UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO TEMPORAIS ENVOLVENDO DADOS QUANTITATIVOS CONTÍNUOS

Rafael Stoffalette João

Orientadora: Profa. Dra. Marcela Xavier Ribeiro

São Carlos – SP

Agosto/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO TEMPORAIS ENVOLVENDO DADOS QUANTITATIVOS CONTÍNUOS

Rafael Stoffalette João

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Bancos de Dados

Orientadora: Profa. Dra. Marcela Xavier Ribeiro

São Carlos – SP

Agosto/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Rafael Stoffalette João, realizada em 07/08/2020.

Comissão Julgadora:

Profa. Dra. Marcela Xavier Ribeiro (UFSCar)

Prof. Dr. Diego Furtado Silva (UFSCar)

Prof. Dr. Gustavo Enrique de Almeida Prado Alves Batista (UNSW)

Profa. Dra. Elaine Parros Machado de Sousa (USP)

Prof. Dr. Agma Juci Machado Traina (USP)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.
O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

A todos os tripulantes da estressante jornada psicocientífica intitulada *"meu doutorado"*. Os quais se perceberam embarcados sem desfrutar do prazer da escolha.

Agradecimentos

A Deus, antes de tudo.

À minha orientadora, professora Dr[a]. Marcela Xavier Ribeiro por todas as contribuições, mas especificamente pela confiança que depositou em mim, conhecimento e experiência na pesquisa que demonstrou nos ensinamentos que a mim fez, pela tranquilidade para lidar com os percalços que surgiram durante estes anos de parceria e pelos vários momentos de diálogos não científicos que nos aproximaram e tornaram este trabalho mais confortável e prazeroso;

Aos professores que compuseram a banca de defesa do doutorado: Dr[a]. Agma Juci Machado Traina, Dr. Diego Furtado Silva, Dr[a]. Elaine Parros Machado de Sousa e Dr. Gustavo Enrique de Almeida Prado Alves Batista, por disponibilizarem parte dos seus preciosos tempos para contribuir com suas experientes visões;

À Fernanda, minha namorada, por acreditar em mim, orgulhar-se e compartilhar dos meus objetivos, ser ouvido dos meus dias de estresse e nas horas vagas, revisora do meu trabalho;

Aos meus pais, Antônio Carlos e Joana pelo incentivo, confiança e orgulho que sempre demonstraram por mim;

Ao meu irmão Renato, amigo de vida e colega de profissão, sempre teremos bons assuntos;

Ao Programa de Pós-graduação em Ciência da Computação (PPGCC) do Departamento de Computação da UFSCar, por me oferecer um caminho para alcançar meus objetivos de vida;

Às verdadeiras amizades que construí durante esse período de trabalho, André Marcatto, Cláudio Paiva, Cleverson, Jaum, Paulão, Flávio, Jasiel e Isaque, bem como aos meus amigos de Osvaldo Cruz Victor, Marco, Juninho, Mario e Eduardo por sempre duvidarem que eu seria capaz e, assim, me motivaram;

Ao professor Dr. Ricardo Rodrigues Ciferri, por um elogio à minha didática que criou em minha cabeça uma questão sobre o que eu realmente sei fazer - hoje eu sei;

À professora Dra. Marilde Terezinha Prado Santos pelo exemplo de sabedoria e carinho sempre presente em nossas conversas; e por repetir tantas vezes a frase: "Confie em você, cara!";

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) por acreditar e confiar no meu trabalho;

Enfim, agradeço a todos que contribuíram de forma direta ou indiretamente para que esse dia chegasse e eu pudesse, enfim, me chamar de Doutor.

"O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001".

To know, is to know that you know nothing. That is the meaning of true knowledge.

(Sócrates)

Lista de Abreviaturas e Siglas

AG – *Algoritmos genéticos*

AIA – *Álgebra intervalar de Allen*

ARMADA – *An algorithm for discovering richer relative temporal association rules from temporal data (um algoritmo para descoberta de regras de associação temporais relativas mais ricas a partir de dados intervalares)*

ART-Q – *Association rules involving temporality and quantitative continuous data*

BDBra – *Base de dados de índices socioeconômicos a respeito do Brasil*

BDMEP – *Banco de dados meteorológicos para ensino e pesquisa*

BDM – *Base de dados meteorológica*

BDRelT – *Base de dados de relações temporais*

BDSocio – *Base de dados de índices socioeconômicos a respeito dos países Brasil, Argentina, Chile, China e Estados Unidos*

BDT – *Bancos de dados temporais*

BDTempo – *Base de dados meteorológica*

BDTempteste – *Banco de dados temporal de teste*

BDTest – *Banco de dados de teste*

BD – *Banco de dados*

Conab – *Companhia nacional de abastecimento*

Conf – *Medida de confiança de uma regra de associação*

Conv – *Medida de convicção de uma regra de associação*

ENE – *Escritório nacional de estatísticas*

IDH – *Índice de desenvolvimento humano*

INMET – *Instituto nacional de meteorologia*

Itemset – *Conjunto de itens*

KDD – *Knowledge discovery in databases (descoberta de conhecimento em bases de dados)*

MDIC – *Ministério do desenvolvimento, indústria e comércio exterior*

MWI – *maximum window for interval (janela máxima para intervalos)*

MWR – *maximum window for relation (janela máxima para relações temporais)*

Min_conf – *Confiança mínima*

Min_sup – *Suporte mínimo*

OFARM – *Optimized fuzzy association rule mining*

PIB – *Produto interno bruto*

POI – *Pontos de interesse*

SGBD – *Sistema gerenciador de bancos de dados*

Resumo

A consideração da temporalidade de forma explícita na mineração de regras de associação que envolvem dados quantitativos contínuos é uma abordagem que tem como intuito principal a contribuição à área da descoberta de informação em bases de dados. A construção de intervalos temporais para os atributos de uma base de dados permite identificar, também, os relacionamentos binários que estes intervalos podem assumir entre si. Este trabalho descreve o desenvolvimento de um novo método, intitulado ART-Q, para a tarefa de mineração de regras de associação temporais que envolvem dados quantitativos contínuos. A temporalidade é assumida, neste trabalho, em sua forma explícita, não somente pelo sequenciamento dos dados. Os padrões que possibilitam a construção das regras são compostos por relações binárias da álgebra intervalar de Allen nos intervalos temporais que descrevem comportamentos de interesse dos atributos quantitativos contínuos. O método provou ser capaz de revelar informações implícitas em bases de dados de diferentes contextos. Os resultados são apresentados por meio de intervalos temporais de interesse dos atributos e suas relações algébricas, padrões e regras de associação temporais. O trabalho demonstra que o método contribui para a evolução da literatura com a definição e busca de um novo tipo de padrão, mais complexo que os existentes. Por meio destes padrões, regras de associação são construídas com semântica que envolve uma vasta quantidade de informações, além da implicação da regra.

Palavras-chave: Mineração de dados. Regras de associação. Temporalidade. Dados quantitativos contínuos.

Abstract

The consideration of temporality in a explicit manner on the task of association rules mining that involves continuous quantitative data is one approach that aims to contribute to the field of study of knowledge discovery in databases. The construction of temporal intervals from attributes of a data set also provides to the method to identify binary relations, which these intervals may have. This work describes the development of a new method, named ART-Q, for the task of mining temporal association rules which involve continuous quantitative data. The temporality is assumed, in the present work, in its explicit form, not only by data sequencing. The patterns that allow the rules construction are made of binary relations from Allen's intervalar algebra in the temporal intervals that describe the continuous quantitative attribute's behavior of interest. The method has proven being able to reveal implicit information in different contexts databases. The results are demonstrated by temporal intervals of interest of the attributes and their algebric relations, patterns and temporal association rules. The work demonstrates that the method ART-Q contributes to the evolution of the literature with the definition and search of a new kind of pattern, more complex than those present in the studies. By the consideration of this kind of patterns association rules are constructed semantically involving a large amount of information among the implication of the rule.

Keywords: Data mining. Association rules. Temporality. Quantitative continuous data.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

Este capítulo introduz o leitor ao tema da mineração de regras de associação temporais que envolvem dados quantitativos contínuos e consideram a temporalidade de forma explícita. Mais especificamente, discorre a respeito das lacunas da área de mineração de dados, cuja solução, este trabalho aborda uma forma de contribuição. Também descreve e justifica a escolha do tema do trabalho, a motivação, hipótese de pesquisa e os objetivos gerais e específicos deste trabalho. Por fim, é apresentada a organização do trabalho, a fim de facilitar a compreensão do documento.

1.1 Considerações iniciais

O volume crescente de dados, ao passo que a tecnologia evolui, fez surgir uma importante questão, assim como interroga Larose (2004): o que fazer com os dados armazenados? As grandes quantidades de dados obtidos geram custos cada vez mais elevados às empresas - equipamentos são caros, a manutenção deve ser constante, há a necessidade de mais pessoas para lidar com os dados, entre outros vários fatores que, mesmo assim, não garantem que a manipulação de tais dados possa ser realizada de forma simples e eficiente. Assim como afirma Han, Pei e Kamber (2012, p.05) "o mundo é rico em dados mas pobre em informação".

O processo de extração de conhecimento de uma base de dados, também conhecido como KDD (*Knowledge Discovery in Databases*), segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), é o processo não trivial de busca e identificação de padrões frequentes em bases de dados. À medida que a evolução da tecnologia se faz presente, além do melhoramento de técnicas e ferramentas já existentes, a inclusão das mesmas em áreas que, até então não se beneficiavam de seu uso, também atrai interesses.

A mineração de dados é uma etapa de grande interesse do KDD. Nesta etapa se lançam mão de algoritmos de inteligência artificial em dados pré-processados para a obtenção de informações implícitas, muitas vezes na forma de padrões. Existem muitos trabalhos que aplicam técnicas de mineração de dados nas mais diversificadas áreas, como o conduzido por Xue et al. (2010) que emprega uma estratégia Fuzzy para gerar regras de associação na análise de qualidade do solo; Romani et al. (2013) realizam uma mineração de séries temporais para a previsão de safras. Trabalhos como os de Nonato e Oliveira (2013) e João et al. (2017), lançam mão de técnicas de mineração de dados para a identificação de plantios de cana-de-açúcar com base em imagens sensoriais, entre outros.

A maioria das técnicas de mineração, ou não consideram a temporalidade dos dados, ou optam por não lidar com atributos quantitativos contínuos. Ou seja, ou perdem a informação temporal ou perdem a informação do valor dos atributos, ou perdem ambas as informações,conforme é melhor justificado no Capítulo 2. No entanto, com a mineração de padrões temporais envolvendo dados quantitativos contínuos é possível enriquecer a análises a respeito da meteorologia, epidemiologia, logística de transportes, entre outros.

1.2 Motivação

A mineração de regras de associação é uma das tarefas mais usadas na mineração de dados. Foi inicialmente proposta por Agrawal R.; Imielinsk (1993) como a busca por relacionamentos entre os itens de transações de compra de um supermercado. Esse problema ficou conhecido como análise de cestas de compra (*market basket*), como evidenciado por Tan, Kumar e Steinbach (2006).

A grande maioria dos trabalhos a respeito das regras de associação, consideram bases de dados compostas por valores discretos ou nominais. Quando os dados são de natureza quantitativa contínua, antes do uso de métodos de mineração de dados, é comum o emprego de uma estratégia de pré-processamento de dados a fim de torná-los discretos, tais como fazem Holte (1993), Liu e Setiono (1995), Ribeiro, Traina e Traina (2008), entre outros. Segundo Olson e Delen (2008) o pré-processamento de dados nos projetos de mineração, corresponde a, no mínimo, 50% de todo o trabalho do projeto. Dependendo do problema esse valor pode chegar a 80%, segundo afirma McCue (2015).

Srikant e Agrawal (1996) propuseram uma das primeiras, e talvez a maior das, referências entre os métodos que realizam o processo de geração de regras de associação em dados quantitativos. Nele, o algoritmo Apriori (apresentado por Agrawal e Srikant (1994))

é estendido a fim de compreender tanto dados quantitativos, quanto dados categóricos. Após sua execução, o método descrito pelos autores gera regras do tipo: *<Idade:30..39>* e *<Casado:Sim>* ⇒ *<NumCarros: 2>*.

Basicamente a estratégia empregada é a de particionar os valores assumidos pelos atributos quantitativos em intervalos (como o intervalo que compreende os valores entre 30 e 39, gerado para o atributo Idade na regra apresentada). Em seguida, conforme o trabalho descreve, uma tabela de mapeamento é construída com valores 1 quando há ocorrência do valor desejado (ou intervalo do atributo) na transação e 0 quando não há. Realiza, desta forma, o processo de discretização dos dados. A Tabela 1.1 exemplifica uma tabela de mapeamento. A partir deste mapeamento, o processo de identificação de padrões e construção de regras de associação segue à semelhança do tradicional algoritmo Apriori.

Tabela 1.1: Mapeamento de dados quantitativos para valores Booleanos.

TID	Idade: 30..39	Idade: 45..60	Casado(a)?	Num. Carros: 2
100	0	1	1	1
200	1	0	1	0
300	0	1	1	0

Fonte: Adaptada de Srikant e Agrawal (1996).

Adhikary e Roy(2015) discutem sobre várias técnicas de mineração de regras de associação quantitativas empregadas em trabalhos disponíveis na literatura. Segundo os autores, as técnicas podem ser classificadas quanto à sua abordagem empregada, podendo ser: de particionamento, como apresentam Srikant e Agrawal (1996), Chan e Au (1997) e Li et al. (2012), agrupamento, como fazem Lian, Cheung e Yiu (2005), Miller e Yang (1997), Yang e Feng (2010) e Fukuda et al. (1996), abordagem estatística, como as de Aumann e Lindell (2003) e Kang et al. (2009), conjuntos Fuzzy, como Zhang (1999) e Zheng et al. (2014) descrevem e as abordagens evolutivas, aqui exemplificadas pelos trabalhos de Kaya e Alhajj (2005) e Martin et al. (2014).

De acordo com Adhikary e Roy (2015), todas as estratégias têm seus pontos positivos e negativos, tudo depende da aplicação a qual elas devem ser utilizadas. Ainda que não haja uma estratégia eleita como a melhor entre todas, a abordagem estatística é empregada em cerca de apenas 9% dos trabalhos na literatura. O que, de fato, é algo que instiga a investigação de sua capacidade para a mineração de regras de associação.

Outro aspecto muito contributivo para o enriquecimento do processo de mineração de regras de associação, além da consideração dos dados quantitativos contínuos, é a

temporalidade associada aos dados. O estudo da temporalidade não é uma área nova de pesquisa, trabalhos como o de Schuster (1906) já relatavam aplicações de séries temporais na análise de fenômenos da natureza. Wiederhold, Fries e Weyl (1975) são os primeiros autores a publicar sobre bases de dados orientadas ao tempo na ciência da computação.

Ainda que não seja uma área de pesquisa recente, a temporalidade explícita dos dados não é amplamente explorada na tarefa de construção de regras de associação. Tão pouco possui um formalismo que é seguido pelos autores, como no caso dos BDT (bancos de dados temporais) que são bem definidos e formalizados por Stam e Snodgrass (1988).

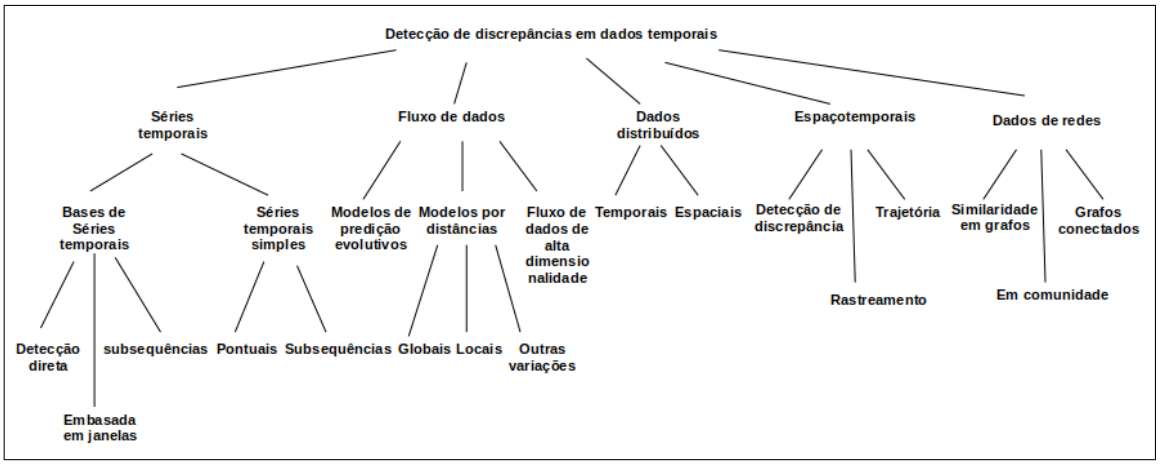
A grande maioria dos trabalhos de mineração de padrões que lidam com a temporalidade, consideram o sequenciamento dos dados como o fator temporal e não manipulam atributos que armazenam informações temporais (a temporalidade explícita do dado), tal como uma data ou hora. Exemplos desta abordagem podem ser: o trabalho de Hirano e Tsumoto (2008) que detecta fatores de risco na área médica, por meio de séries temporais, de Kida, Saito e Arimura (2009), Harms e Deogun (2004) e Koh e Chou (2009), os quais realizam a tarefa de extração de padrões temporais em fluxos de dados.

Não há comprovação da superioridade de uma adoção quanto a outra, a escolha de como a temporalidade será tratada deve ser conduzida pela necessidade do problema. Pode não ser possível considerar o sequenciamento dos dados como fator temporal nos casos em que a base de dados possui valores faltantes. Soma-se à dificuldade de compatibilização de valores temporais de múltiplas fontes de dados, que além de necessitar conversões quando o fuso horário é diferente, tende a ser custosa quanto ao tempo de processamento.

O trabalho de Gupta et al. (2014) é uma contribuição interessante para apoiar a adoção da temporalidade explícita neste trabalho. No trabalho citado, os autores realizam uma inspeção profunda na literatura a respeito de estratégias que detectam discrepâncias em bases de dados temporais, os chamados *Outliers*. Ainda que o foco do trabalho diverge do explorado nesta tese, a varredura na literatura realizada pelos autores mostra-se como uma importante contribuição.

A Figura 1.1, abaixo, apresenta um esquema organizacional das estratégias que lidam com o tema na literatura. No esquema é possível observar que nenhuma ramificação compreende a estratégia que considera a temporalidade de forma explícita. Isso evidencia que existe uma lacuna a ser explorada para contribuir de forma efetiva na tarefa de mineração de dados - ponto este que colabora com a justificativa da escolha deste tema de trabalho.

Figura 1.1: Esquema organizacional de estratégias que lidam com a temporalidade na detecção de discrepâncias (*outliers*).



Fonte: Adaptada de Gupta et al. (2014).

O que se encontra na literatura é uma grande quantidade de trabalhos que optam em lidar com informações temporais implícitas. No entanto, o potencial de padrões que podem ser obtidos ao abordar duas lacunas ainda pouco exploradas na área de mineração de regras de associação: (a) considerar o tempo de forma explícita e (b) a manipular dados quantitativos contínuos, não foi explorado anteriormente na literatura por outros trabalhos.

1.3 Hipótese

A hipótese investigada neste trabalho de pesquisa pode ser descrita pela seguinte afirmação:

A incorporação da temporalidade ao processo de mineração de padrões e regras de associação, especificamente àquelas que consideram dados quantitativos (inclusive contínuos) em sua construção, pode contribuir com a descoberta de informações úteis com composição ainda pouco (ou não) explorada, i.e., que exploram a variabilidade dos valores assumidos pelos atributos de uma base de dados, ao mesmo passo em que a temporalidade implica na mutação e validade das informações.

1.4 Objetivos - geral e específicos

O objetivo principal deste trabalho foi a definição e construção de um método computacional que realiza o processo de mineração de regras de associação temporais quando são

considerados dados quantitativos contínuos. O método desenvolvido teve como princípio a descoberta de novos tipos de padrões temporais, que contemplam relações temporais entre os valores de maior interesse assumidos pelos atributos quantitativos.

Para a contemplação do objetivo geral deste trabalho de pesquisa, os seguintes objetivos específicos foram idealizados:

- Revisão do Estado da Arte sobre a mineração de regras de associação temporais envolvendo dados quantitativos e contínuos;
- Desenvolvimento de um método para a obtenção dos valores de interesse para os atributos contínuos;
- Desenvolvimento de uma estratégia para a construção de intervalos temporais nos quais os atributos assumem os valores de interesse;
- Definição da representação dos intervalos temporais e seus relacionamentos com sua vizinhança. Nesta etapa foi usada a Álgebra Intervalar de Allen (AIA) para tratar as relações temporais;
- Definição e implementação do método ART-Q de mineração de regras de associação envolvendo dados quantitativos contínuos.

1.5 Organização do trabalho

Este documento descreve o trabalho realizado, em nível de doutorado, a respeito do tema mineração de regras de associação temporais que envolvem dados quantitativos contínuos. O documento está organizado da seguinte forma: O Capítulo 2 apresenta a fundamentação teórica que foi considerada para a condução deste trabalho. São apresentados: os conceitos de dados quantitativos contínuos; temporalidade explícita dos dados; a álgebra intervalar de Allen para a representação de relações entre intervalos temporais; e o algoritmo Apriori para a construção das regras de associação. No Capítulo 4 é apresentado o método proposto, intitulado por ART-Q: *Association Rules involving Temporality and Quantitative continuous data* e a forma como realiza a tarefa a qual foi idealizado a cumprir. O Capítulo 5 apresenta os experimentos realizados para validar o método ART-Q. Já o Capítulo 6 apresenta as conclusões, limitações e trabalhos futuros. Por fim, o Capítulo 3 apresenta o conjunto de trabalhos correlatos organizados em: (1) aqueles que lidam com dados quantitativos contínuos, (2) aqueles que lidam com a temporalidade na mineração de regras de associação e (3) os que atacam as duas linhas de pesquisa.

Considerações finais

Este capítulo apresentou a introdução ao tema abordado por este trabalho: a mineração de regras de associação temporais que envolvem dados quantitativos contínuos, seguida pela motivação para a condução deste estudo, a hipótese de pesquisa idealizada, os objetivos geral e específicos que serviram como norte para o êxito na contribuição que o ART-Q provê para a tarefa de mineração de dados quando esta envolve a temporalidade explícita dos dados e a manutenção dos valores quantitativos contínuos, que evita a omissão de nuances nos dados. Por fim, foi apresentada a organização deste documento que descreve o trabalho realizado.

Capítulo 2

Referencial teórico

Neste capítulo são apresentados os principais conceitos usados no desenvolvimento deste trabalho: tipos dos dados; distribuição de probabilidade; representação de dados; o algoritmo Apriori e toda conceituação que o envolve; a temporalidade e suas formas de adoção explícita e implícita; e, a álgebra intervalar de Allen (AIA), como um modelo temporal muito eficiente para a representação das relações temporais entre dois intervalos de tempo.

2.1 Tipos de dados

Segundo fundamentado por Martins (2005), uma característica que assume valores distintos para cada um dos indivíduos é denominada *variável*. Rumsey (2009) define uma variável como a característica que é medida ou avaliada em cada elemento da amostra ou população. Ainda segundo Martins (2005) as variáveis podem ser divididas em dois grupos: (1) as qualitativas e (2) as quantitativas. Valores assumidos por uma variável podem ser armazenados e são denominados dados. Martins comenta que a classificação, quanto ao tipo das variáveis na análise dos dados, segue o sistema proposto por Stevens (1951). Neste trabalho os tipos possíveis de dados são apresentados a seguir e seguem a mesma definição para a classificação das variáveis, apresentada por Stevens (1951).

2.1.1 Dados de natureza qualitativa (ou categóricos)

Os dados qualitativos, também denominados categóricos, são aqueles que representam uma informação referente a alguma qualidade, característica ou categoria. São dados que não são suscetíveis de medidas, i.e., não são mensuráveis. Exemplos reais de dados qualitativos podem ser: a variável que descreve o sentido de um caminho (norte, sul, leste

ou oeste); o estado civil de uma pessoa (solteiro, casado, divorciado ou viúvo); a altura de um determinado indivíduo (alto, médio e baixo); as cores de um carro que pode ser branco, preto, azul, etc.; e outros. Dados do tipo qualitativos podem ser subdivididos entre aqueles que são expressos em (a) escala nominal e aqueles expressos em (b) escala ordinal, detalhadas como segue:

- Os **(a) Dados qualitativos em escala nominal** os quais não representam valores numéricos, mas sim uma categoria dentro das possibilidades existentes. Entretanto não há ordenação entre os valores; uma variável é nominal se cada observação pertence a uma categoria. Por exemplo, a variável estado civil é nominal, uma vez que cada pessoa é solteira, casada, viúva ou divorciada.

  Números podem ser utilizados como dados qualitativos nominais, tal como os números 1 e 2 podem ser considerados para categorizar o gênero de um indivíduo em masculino e feminino, respectivamente. Entretanto não são mensuráveis.

- Como o próprio nome sugere, **(b) dados qualitativos em escala ordinal** respeitam uma ordem entre as categorias. Essa escala de medida pode ser numérica ou não, entretanto assim como na escala nominal, não é possível quantificar diferenças entre os resultados.

  Um exemplo de variáveis qualitativas ordinais são as que categorizam o grau de escolaridade de um indivíduo, que pode ser ensino fundamental, médio e superior.

2.1.2 Dados de natureza quantitativa (numéricos)

Dados quantitativos são aqueles que representam características possíveis de serem medidas ou contadas. Em outras palavras são dados que se referem a números. Exemplos de dados quantitativos são: número de arremessos em um jogo de basquete, o preço de um produto, as notas de alunos em uma disciplina, o peso ou altura de um indivíduo, etc. Dados quantitativos podem ser de natureza (c) contínua (dados contínuos) ou de natureza (d) discreta (dados discretos).

Uma variável é de **(c) natureza contínua** se tem a possibilidade de assumir infinitos valores em um intervalo do domínio da variável. Uma variável quantitativa de natureza contínua é também chamada variável intervalar. Um dado *quantitativo contínuo* é um valor assumido por uma variável quantitativa contínua, ou seja, um valor assumido que pertencente ao conjunto dos números reais $\mathbb{R} = \{-\infty, ..., +\infty\}$. Estes valores podem ser mensurados pela soma, média, desvio padrão, variância, etc. Entretanto não são

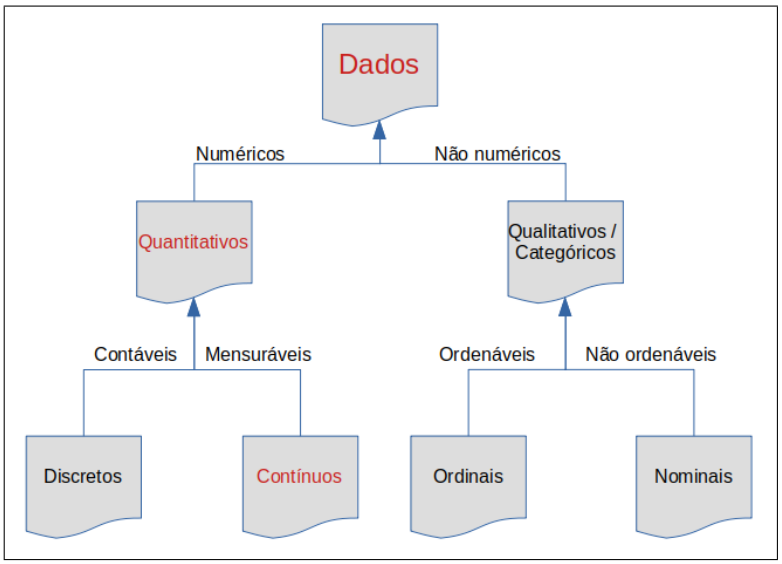
contáveis. Em outras palavras, dados de uma característica que podem ser medidos, mas não são contáveis são *dados quantitativos contínuos*. Exemplos de dados quantitativos de natureza contínua são a altura de um objeto (em centímetros), o tempo de uma viagem (em horas, minutos e segundos), a velocidade do vento (km/hora), a distância entre dois pontos infinitamente próximos, entre outros.

Quando o valor assumido por uma variável é um inteiro, essa variável é de **(d) natureza discreta**. São variáveis que assumem valores numerais inteiros que podem ser contados. Tal como ocorre com o número de filhos de um casal, a quantidade de gols em um jogo de futebol, o número de ligações realizados em uma empresa durante um dia, etc. Desta forma, um dado que representa uma característica e que pode ser contada é um *dado quantitativo discreto*.

Intuitivamente, é fácil concluir que valores assumidos por toda variável qualitativa, tanto as nominais quanto ordinais, são de natureza discreta, i.e., podem ser contados. Mas por se tratar de uma variável qualitativa, não podem ser mensurados pela média.

Como indica o nome, dados contínuos são sempre intermediados por infinitos valores, de forma a construir uma ligação entre eles. Já nas variáveis de natureza discreta a mudança de um valor para outro se dá por um salto. A Figura 2.1 sumariza em uma representação gráfica a hierarquia na classificação dos tipos de dados, assim como detalhado acima.

Figura 2.1: Representação hierárquica da classificação dos dados, quanto aos tipos de valores que podem assumir.



Fonte: Elaborada pelo autor.

Martins (2005) levanta, ainda, uma discussão interessante sobre as classificações das

variáveis: a idade de um indivíduo à primeira vista parece ser um valor de natureza discreta, uma vez que são utilizados números inteiros para representá-la. Entretanto a análise mais profunda da variável mostra que a diferença de idade entre dois indivíduos pode ser tão pequena quanto se queira (um ano, um mês, um dia, horas, minutos, segundos), portanto trata-se, na verdade, de um valor de natureza contínua.

Um dado quantitativo contínuo é armazenado em um banco de dados por meio de um atributo temático quantitativo contínuo. Formalmente definido como segue:

Definição 2.1 (Atributo temático)**.** Um *atributo* que compõe uma relação de uma base de dados é dito *temático* quando os valores que ele assume não representam informações temporais nem espaciais. Pode, portanto, ser um valor numérico ou textual, sem restrições. Um *atributo temático* é notado neste trabalho por *att*.

Definição 2.2 (Atributo temático quantitativo contínuo)**.** Quando um atributo temático (*att*) assume um valor quantitativo contínuo, este é chamado um *atributo quantitativo contínuo* (identificado neste trabalho por $a_{cont} \mid a_{cont}$ é um *att*). Um a_{cont} assume infinitos valores numéricos reais que pertencem a um intervalo $[-v_{a_{cont}}, +v_{a_{cont}}]$, no qual $-v_{a_{cont}}$ é o menor valor que a_{cont} pode assumir e $+v_{a_{cont}}$ o maior valor.

Exemplo 2.1. Os valores de temperatura (aferidos em escala *Kelvin*) de uma determinada cidade *Cid* são armazenados pelo atributo *temperatura(Cid)* em uma relação de uma base de dados da seguinte forma:

$$temperatura(Cid) :< 300,\ 1,\ 299,\ 23,\ 298,\ 7,\ 299,\ 1,\ 299,\ 8,\ 301,\ 3,\ 303,\ 15 >,$$

Nota-se que os valores de temperatura são de natureza quantitativa contínua, portanto *temperatura(Cid)* é dito um atributo quantitativo contínuo (a_{cont}) que assume valores no intervalo [298, 7..303, 15].

No trabalho de Telikani, Gandomi e Shahbahrami (2020) um levantamento na literatura é conduzido a respeito de algoritmos de mineração de regras de associação, mais focado nos que seguem a linha de algoritmos genéticos. No total, o trabalho incorpora 221 algoritmos desenvolvidos entre os anos de 2000 e 2019, divididos em nove grupos de acordo com suas respectivas heurísticas. Conforme ressaltam os autores, a grande maioria dos trabalhos que lidam com dados quantitativos contínuos (referenciado no trabalho como QARs) lança mão de alguma estratégia de discretização dos dados.

O processo de discretização dos dados é a ação de transformar dados quantitativos contínuos em valores discretos. Por exemplo, ao assumir que valores de temperatura são

considerados "elevados"quando ocorrem acima de 35°C, "normais" entre 21°C e 35°C e "baixos" o que ocorre é a construção de três grupos para descrever todos os valores do atributo. A partir deste momento, os valores numéricos de temperatura não precisam mais serem considerados, apenas os grupos construídos.

Ainda segundo Telikani, Gandomi e Shahbahrami (2020), discretizar os valores dos dados nestas condições implica em um resultado mais pobre em informação gerado pelo algoritmo de mineração de regras de associação. Isso porque a eficiência destes métodos depende dos intervalos para a definição dos grupos discretos. Determinar intervalos apropriados é uma difícil tarefa. Entretanto, quando bons conjuntos são definidos, os algoritmos da classe QARs entregam resultados mais expressivos.

Como será discutido mais à frente, neste documento, o método proposto e desenvolvido neste trabalho permite uma flexibilidade na construção de intervalos numéricos que representam informações de interesse dos valores. Não realiza, entretanto um processo de discretização que constrói grupos para descrever todas as possibilidades de valores, que muitas vezes não são interessantes para a análise dos dados que se quer conduzir. Porém, para adentrar com mais profundidade nesta etapa que o ART-Q implementa, conceitos como devem ser apresentados e discutidos, como o de distribuição de probabilidade, a seguir.

2.2 Distribuição de probabilidade

Conforme comentam Stevenson (1986), Martins (2005), Bussab e Morettin (2013), uma variável aleatória é um elemento que assume um valor único para cada um dos resultados possíveis em um experimento que é conduzido. Em outras palavras, uma variável aleatória pode, por exemplo, descrever os resultados de um arremesso de uma moeda ao assumir o valor *cara* ou *coroa*. Ou seja, seu valor é determinado ao acaso. A denominação *aleatória* indica que, em geral, o valor assumido pela variável só pode ser conhecido após a realização do experimento, como afirma Triola (1998).

Uma distribuição de probabilidade associa cada um dos possíveis valores que uma variável aleatória pode assumir (resultado de um arremesso de dados, por exemplo) à probabilidade de que ele ocorra. Assim como as variáveis, uma distribuição de probabilidade pode ser discreta, quando associa valores de probabilidade de uma variável discreta, o que significa que os resultados possíveis são finitos ou contáveis, ou contínua, quando associa probabilidades de uma variável contínua e seus resultados são infinitos e incontá-

veis.

Quando uma distribuição de probabilidade é discreta, ela pode ser representada visualmente por histogramas. Já quando é contínua, o gráfico da distribuição que a representa é uma linha contínua.

Algumas distribuições discretas mais conhecidas são:

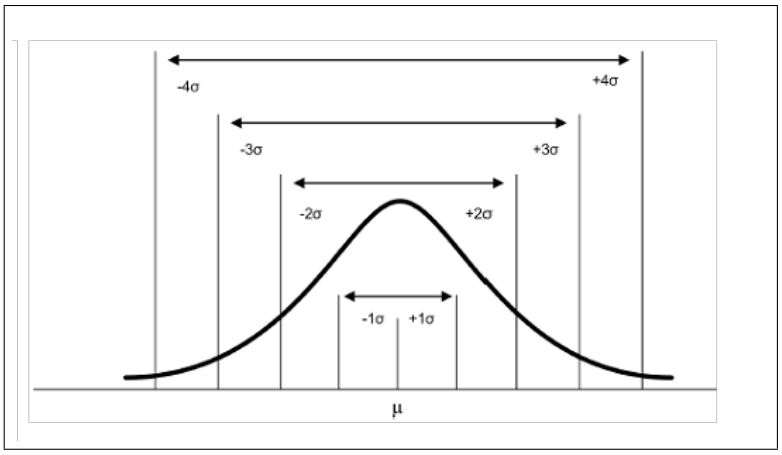
- *distribuição uniforme discreta*: modela um fenômeno aleatório cujos resultados são sempre igualmente prováveis;
- *distribuição de Bernoulli*: corresponde a uma experiência com somente dois resultados (sucesso ou fracasso), que geralmente correspondem aos valores 1 e 0;
- *distribuição binomial*: é a distribuição do número de sucessos, ou fracassos, tal que somente dois resultados são possíveis - assim como na distribuição de Bernoulli, entretanto o número de eventos é definido;
- *distribuição de Poisson*: modela o número de eventos observados por unidade de tempo, ou espaço - o parâmetro λ define a taxa de eventos por unidade; entre outras, tais como a aritmética, geométrica, binomial negativa, etc.

Entre as distribuições contínuas mais conhecidas estão:

- *distribuição exponencial*, na qual a variável aleatória é definida como o tempo entre duas ocorrências;
- *distribuição de Weibull*, que tem sido amplamente utilizada para modelar tempos de eventos consecutivos, ou tempos até que uma falha ocorra e;
- *distribuição normal*: a representação da distribuição normal tem formato de sino. Uma curva na qual a maior parte dos valores se concentra próximos ao centro da distribuição, onde o parâmetro μ especifica a média e o σ, o desvio padrão. A curva é chamada *curva normal* ou *curva de Gauss.* ilustrada na Figura 2.2. A distribuição normal é a mais utilizada para a representação de probabilidades.

Triola (1998) afirma que dado uma variável aleatória x, com distribuição (que pode ser normal, ou não), na medida em que o tamanho da amostra aumenta, a distribuição das médias amostrais de x tende para uma distribuição normal. A aproximação melhora na medida em que aumenta o tamanho da amostra. Em outras palavras, quanto maior

Figura 2.2: Representação gráfica da distribuição normal de probabilidade. Os parâmetros μ e σ descrevem, respectivamente, a média e o desvio padrão da população.



Fonte: Elaborada pelo autor.

a amostragem de dados é, maior é a possibilidade da distribuição normal ser a mais adequada para ser utilizada pelo método ART-Q. Um exemplo pode ser uma repetição de lançamentos de uma moeda. Quando tende ao infinito a quantidade de resultados obtidos iguais (coroas por exemplo), se aproxima de uma distribuição normal.

A distribuição normal é simétrica em relação à média, i.e., a probabilidade dos valores à direita da média é a mesma da dos valores à esquerda, 50% se somados todos os valores de cada um dos lados. A área total abaixo da curva normal, se somada, representa todas as probabilidades, 100%. Por convenção, as probabilidades são referenciadas com relação ao número de desvios padrão de distância da média, $\mu \pm n\sigma$, onde $n \in N$(conjunto de números naturais).

2.3 O algoritmo Apriori e o processo de geração de regras de associação

Inicialmente, a busca por regras de associação em bases de dados de transações foi formalizada por Agrawal R.; Imielinsk (1993) com a aplicação em bases de dados de comércio a fim de identificar relações entre produtos comumente comprados juntos. O algoritmo Apriori, idealizado por Agrawal e Srikant (1994) é o pioneiro e a maior referência dos algoritmos que adotam a estratégia de geração de candidatos no processo de construção de regras de associação. Estratégia pela qual sempre que houver pelo menos dois elementos frequentes de tamanho k na base de dados, com $k-1$ elementos em comum, estes podem

ser combinados de forma a gerar novos elementos, denominados candidatos a padrões frequentes, de tamanho $k+1$.

O algoritmo Apriori é amplamente considerado para a busca de regras de associação, dado que é um algoritmo de fácil entendimento, além de eficiente. Nele, é empregada a antimonotonia da relação (também conhecida como a propriedade Apriori), a qual infere que dados dois conjuntos de itens (chamados *itemsets*) e_1 e e_2 tal que $e_1 \subset e_2$, se e_1 não é considerado um padrão da base de dados então e_2 também não é. Em outras palavras, é impossível que os itens que compõem e_1 (que não é padrão frequente) possam ser combinados a um outro item que os torne frequentes. Desta forma, o algoritmo reduz a quantidade de operações necessárias, pois descarta todas as combinações que seriam originadas por *itemsets* não frequentes.

Antes de detalhar como o Apriori realiza o processo de construção das regras de associação, entretanto, é importante que algumas definições e propriedades sejam apresentadas. Tais como:

Definição 2.3 (Suporte de um item)**.** O suporte de um item i representa a frequência de ocorrência do item i na base de dados BD. É obtido por meio da Equação 2.1, que calcula a razão da quantidade de transações na base de dados que possuem o item i pela quantidade total de transações da base de dados $|BD|$;

$$sup(i) = \frac{\sum_1^{|BD|} T \mid T \subseteq BD, i \subseteq T}{|BD|} \begin{cases} frequente, & sup(i) \geq Min\_sup \\ não\ frequente, & sup(i) < Min\_sup \end{cases} \tag{2.1}$$

Definição 2.4 (Suporte de um conjunto de itens)**.** Da mesma forma, o suporte de um conjunto de itens $X \subseteq I$, i.e., um *itemset*) é a frequência de ocorrências de X em BD, calculado pela Equação 2.1;

Definição 2.5 (Regras de associação)**.** Seja DB uma base de dados de transações. Uma regra de associação é uma implicação da forma $X \Rightarrow Y$, na qual X é denominada parte antecedente e Y consequente. X e Y são compostos por itens diferentes ($X \cap Y = \emptyset$). A regra pode ser interpretada da seguinte forma: se uma transação contém X, então ela tende a ter Y.

Exemplo 2.2. A regra de associação $R_i : arroz,\ ovo,\ sal \Rightarrow farinha$ é uma implicação que dita que a ocorrência dos itens arroz, ovo e sal, em uma mesma transação, implica na ocorrência do item farinha, na mesma transação.

Definição 2.6 (Confiança de uma regra de associação)**.** O valor de confiança de uma regra de associação $R_i : X \Rightarrow Y$, apresentado pela Equação 2.2, é dado pela razão entre

o suporte da regra de associação e o suporte somente de seu antecedente. Esse valor descreve o grau de confiança (força) da regra ao afirmar a ocorrência de Y, dado que é conhecida a ocorrência de X na transação;

$$conf(R_i) = \frac{sup(X \cup Y)}{sup(X)} \begin{cases} forte, & conf(R_i) \geq Min\_conf \\ fraca, & conf(R_i) < Min\_conf \end{cases} \tag{2.2}$$

Definição 2.7 (*Lift* de uma regra de associação)**.** O valor de *lift* (também conhecido como *interest* ou elevação) de uma regra de associação $R_i : X \Rightarrow Y$, estimado por meio da Equação 2.3, é dado pela razão entre o valor de confiança de R_i pelo suporte do consequente da regra (Y). Essa métrica descreve o grau de dependência de uma regra de associação. Por meio do *lift* se pode mensurar o quanto a ocorrência de X incrementa (eleva) a possibilidade de Y também ocorrer. $Lift(\ X \Rightarrow Y) = Lift(\ Y \Rightarrow X)$;

$$Lift(R_i) = \frac{conf(R_i)}{sup(Y)} \tag{2.3}$$

Definição 2.8 (Convicção de uma regra de associação)**.** O valor de convicção de uma regra de associação $R_i : X \Rightarrow Y$, estimado por meio da Equação 2.4, é dado pela razão entre o complemento do valor de suporte do consequente da regra pelo complemento do valor de confiança de R_i. A convicção de uma regra de associação descreve, assim como o valor de *lift*, o grau de dependência de uma regra de associação. Entretanto a convicção de uma regra de associação, respeita o sentido da implicação, i.e., $conv(\ X \Rightarrow Y)! = conv(\ Y \Rightarrow X)$. Quanto maior o valor de convicção de uma regra de associação R_i, maior é a dependência de Y por X;

$$conv(R_i) = \frac{1 - sup(Y)}{1 - conf(R_i)} \tag{2.4}$$

Propriedades das regras de associação

1. O tamanho de uma base de dados, notado por $|BD|$ é estabelecido pela contagem da quantidade de transações que a compõem;

2. Analogamente, o tamanho de uma regra de associação $R_i : X \Rightarrow Y$ é definido pela quantidade de itens que compõem a regra, independente se são pertencentes ao antecedente ou consequente da mesma, i.e., $\mid X \mid + \mid Y \mid$;

3. O *suporte* de uma regra de associação $R_i : X \Rightarrow Y$ é o suporte do conjunto de itens formados pela união dos itens que compõem a parte antecedente e consequente da

regra $(X \cup Y)$;

4. Um item, um conjunto de itens ou uma regra de associação, são ditos *frequentes* se possuírem valor de suporte maior que um valor limite pré-definido inicialmente, denominado Min_sup;

5. Uma regra de associação é dita forte (ou confiante) se possui valor de confiança maior, ou igual, ao limite Min_conf pré-estabelecido.

Exemplo 2.3. Considere a base de dados *BDteste* de transações da Tabela 2.1, de tamanho $|BDteste| = 5$ e a regra de associação $R_i: B, C \Rightarrow R$, de tamanho $|R_i| = 3$. A regra foi obtida por meio de combinações dos *itens frequentes* (B, C, R) - (B, C) parte antecedente e consequente (R) de R_i. O valor de confiança da regra R_i pode ser facilmente obtido por meio da Equação 2.2, da seguinte forma:

De acordo com a Equação 2.2 para calcular o valor de confiança da regra R_i é preciso ter o conhecimento prévio dos valores de suporte de $(X \cup Y)$ e (X):

- $sup(B,C,R) = \frac{\sum_1^5 T \ |(B,C,R) \subseteq T}{|BD|} = \frac{2}{5} = 0,4.$
- $sup(B,C) = \frac{\sum_1^5 T \ |(B,C) \subseteq T}{|BD|} = \frac{3}{5} = 0,6.$
- $sup(R) = \frac{\sum_1^5 T \ |(R) \subseteq T}{|BD|} = \frac{3}{5} = 0,6.$

Portanto:

- $conf(R_i) = \frac{sup(X \cup Y)}{sup(X)} = \frac{(B,C,R)}{(B,C)} = \frac{0,4}{0,6} = 0,667.$
- $Lift(R_i) = \frac{conf(R_i)}{sup(Y)} = \frac{0,667}{(0,6)} = 1.$
- $conv(R_i) = \frac{1 - sup(Y)}{1 - conf(R_i)} = \frac{(1-0,6)}{(1-0,667)} = \frac{0,4}{0,333} = 1,2012.$

Tabela 2.1: BDteste - Base de dados de transações para teste.

Transação	Itens da transação
T1	A, C, H
T2	B, C, J, R
T3	A, B, C, R
T4	B, R
T5	B, C, K

Fonte: Elaborada pelo autor.

Considere, ainda, que anteriormente foi estabelecido um valor de confiança mínimo de 50% para que uma regra seja *confiante*, ($Min\_sup = 0,5$). Dessa forma, é possível afirmar que a regra R_i é uma regra forte, pois $conf(R_i) = 0,667 \geq Min\_sup = 0,5$. Portanto, a seguinte afirmação pode ser feita: *De acordo com a regra* $R_i : B,C \Rightarrow R \mid 0,667$*, a ocorrência de (B,C) em uma transação de BDteste implica na ocorrência, também, de R com confiança de 0,66.* Ou seja, em 66,7% das transações que contém o *itemset* (B,C), também é encontrado o item (R).

Dois parâmetros de entrada são necessários para a execução do algoritmo Apriori: (1) o suporte mínimo (Min_sup) para a busca de padrões; e, (2) o valor de confiança mínima (Min_conf), utilizado na etapa de construção das regras de associação. O algoritmo Apriori é apresentado a seguir, dividido em dois blocos de códigos. O pseudocódigo do processo de identificação de padrões frequentes é apresentado pelo Algoritmo 1, enquanto a etapa de geração das regras de associação por meio dos *itemsets* identificados anteriormente é apresentada no Algoritmo 2. Ambos os algoritmos foram construídos a partir da descrição apresentada por Agrawal e Srikant (1994).

Inicialmente, o Apriori identifica todos aqueles itens (unitários) que são frequentes na base de dados e os armazena no conjunto L_1 (linha 2). Esta tarefa é executada simplesmente pela comparação do parâmetro Min_sup à contagem de quantas ocorrências existem de cada um dos itens na base de dados. Aqueles que ocorrem com maior frequência daquela limitada por Min_sup são considerados frequentes. Os padrões frequentes identificados são chamados *large itemsets* pelos autores.

Algoritmo 1 Procedimento de identificação de padrões frequentes empregado pelo algoritmo Apriori, adaptado de Agrawal e Srikant (1994).

```
1: procedure APRIORI(Min_sup, BD)
2:     L_1 ← todos_itens_unitarios_frequentes;
3:     L ← ∅;
4:     for k ← 2; L_{k-1} ≠ 0; k++ do
5:         C_k ← gera_canidadtos_a_padroes(L_{k-1});          ▷ Novos candidatos
6:         for all transacao t ∈ BD do
7:             C_t ← candidatos_em_t(C_k, t);                 ▷ candidatos contidos em t
8:             for all candidatos c ∈ C_t do
9:                 c.contagem++;
10:            end for
11:        end for
12:        L_k ← {c ∈ C_k | c.contagem ≥ Min_sup};
13:        L ← L ⋃ L_k;
14:    end for
15:    return L;
16: end procedure
```

O algoritmo repete o processo enquanto novos padrões frequentes possam ser identificados na base de dados, i.e., enquanto L_{k-1} não for vazio (linhas de 3 a 13). A tarefa apresentada na linha 4 do Algoritmo 1, chamada *gera_candidatos_a_padroes(L_{k-1})*, é a responsável por gerar os chamados candidatos (de tamanho k) a possíveis *itemsets* frequentes, a partir do conjunto de padrões frequentes de tamanho $k-1$ - padrões frequentes identificados no passo anterior.

A construção dos candidatos por meio da tarefa *gera_candidatos_a_padroes(L_{k-1})* é exemplificada da seguinte forma: dado que o conjunto de padrões frequentes identificados de tamanho $k = 3$ é representado por $L_3 = \{(a,b,c),\ (a,b,d),\ (a,c,d),\ (a,c,e),\ (b,c,d)\}$. Os pares de *itemsets* frequentes de L_3 são combinados seguindo o critério de seleção daqueles que são semelhantes até o seus penúltimos itens (como ocorre com (a,b,c) e (a,b,d)). Como resultado deste exemplo, o conjunto de candidatos a *itemsets* frequentes $C_4 = \{(a,b,c,d),(a,c,d,e)\}$ é gerado. Nota-se que os itens sempre respeitam a ordem lexicográfica. Os candidatos gerados pelo *gera_candidatos_a_padroes(L_{k-1})* são armazenados no conjunto C_k.

Após a construção do conjunto de candidatos C_k, todas as transações t da base de dados BD são visitadas novamente a fim de verificar se cada um dos candidatos pertencentes a C_k está presente, também, em t. Quando o candidato é identificado em t, esse tem sua contagem de frequência incrementada (linhas 7/8). Ao final da contagem de frequência dos candidatos, aqueles que ocorrem com maior frequência que aquela dita

por Min_sup são considerados frequentes e compõem o conjunto de *itemsets* frequentes de tamanho k, o L_k (linha 11). A saída do Apriori é a união de todos os conjuntos L_k para todos os valores possíveis de k (linha 13). Ou seja, todos os *itemsets* de todos os tamanhos que são considerados frequentes.

A partir de todos os padrões frequentes identificados pelo Apriori (etapa descrita pelo Algoritmo 1), o próximo passo é a geração das regras de associação. Para tal, o Apriori faz uso dos padrões frequentes previamente identificados. Os detalhes deste procedimento são apresentados no Algoritmo 2.

Segundo Agrawal e Srikant (1994), o problema da descoberta de todas as regras de associação, pode ser decomposto em dois subproblemas: (1) encontrar *itemsets*, Y, na BD que ocorrem com alta frequência e (2) usar os *itemsets* encontrados em (1) para gerar regras de associação entre os itens que os compõem.

Segundo a definição inicial proposta por Agrawal e Srikant (1994), as regras de associação construídas pelo Apriori são do tipo $X \Rightarrow Y$, em que X é um conjunto de itens (um *itemset*) e Y é um item unitário; tanto X quanto Y são pertencentes a um mesmo padrão identificado; entretanto X e Y não compartilham um mesmo item. Desta forma, cada um dos padrões identificados anteriormente é selecionado e os itens que os compõem são organizados como descreve o padrão de regras do Apriori.

Por exemplo, considere que o *itemset* $it = (a, b, c, d)$ é frequente, i.e., um padrão, e foi identificado pelo Apriori na etapa anterior. Note que it é pertencente a L_4, ou seja, é composto por 4 itens; it pode ser selecionado para a construção das seguintes regras de associação: $R_1: \ a, b, c \Rightarrow d$; $R_2: \ a, b, d \Rightarrow c$; $R_3: \ a, c, d \Rightarrow b$ e $R_4: \ b, c, d \Rightarrow a$.

É exatamente essa ação que a tarefa *possiveisRegras(it)* (linha 4) realiza. A partir de cada um dos *itemsets* do conjunto dos padrões frequentes *PadroesIdentificados*, todas as combinações possíveis de regras são geradas (combinações de antecedentes e consequentes) e armazenadas no conjunto *RegrasPossiveis*.

Cada uma das combinações a possíveis regras construída é, então, verificada quanto a seu valor de confiança (linha 6), i.e., a probabilidade de ocorrência de Y sempre que X é identificado. Todas as regras que possuem valor de confiança maior, ou igual, que o limiar Min_conf previamente estabelecido, são ditas *fortes*.

Apesar da versão inicial do algoritmo Apriori (apresentada por Agrawal e Srikant (1994)) admitir apenas um item como consequente da regra de associação, a grande maioria dos trabalhos (inclusive a própria evolução do algoritmo Apriori, idealizada por Srikant

Algoritmo 2 Procedimento de geração de regras de associação empregado pelo algoritmo Apriori, adaptado de Agrawal e Srikant (1994).

```
1: procedure APRIORI(Min_conf, PadroesIdentificados)
2:     RegrasFortes ← ∅;
3:     for all itemset it ∈ PadroesIdentificados do
4:         RegrasPossiveis ← possiveisRegras(it);
5:         for all regra r ∈ RegrasPossiveis do
6:             if confianca(r) ≥ Min_conf then
7:                 RegrasFortes ← RegrasFortes⋃regra;
8:             end if
9:         end for
10:    end for
11:    return RegrasFortes;
12: end procedure
```

e Agrawal (1996)) constroem regras de associação com consequentes compostos por um, ou vários, itens. Da mesma maneira, a proposta descrita neste trabalho de pesquisa admite que regras compostas por consequentes não unitários possam ser construídas, também.

2.4 A temporalidade na mineração dos dados

A consideração da temporalidade na tarefa de análise de dados, especificamente na mineração de dados, apesar de ser intuitiva, não é simples. Há tempos, trabalhos que realizam mineração de dados têm explorado as mais distintas formas de representação e incorporação do fator temporal. Entretanto a adoção do tempo é feita, comumente, de forma particular, não é formalizada, pois cada autor a trata à sua maneira. Em seu trabalho, ao definir uma álgebra para a representação de relações temporais entre intervalos, Allen (1981) referencia um modelo de representação temporal proposto por Bruce (1972).

A temporalidade da informação é uma grandeza muito abrangente. De fato, não há sistema humano ou computacional que não sofra influência temporal. O tempo é um agente de mudanças, novas informações são criadas à medida que o tempo passa, enquanto outras deixam de existir.

Conforme descrevem Laxman e Sastry (2006), a mineração de dados temporais pode ser agrupada em cinco categorias principais: a predição, classificação, agrupamento, busca e recuperação de informação e descoberta de padrões. Entretanto a mineração de dados, ao que se considera a temporalidade no processo, não apresenta um formalismo comum no que se diz respeito a forma como a temporalidade é tratada. De fato, qualquer abordagem

que considere o fator tempo, seja qual for a forma como o trata, é dita uma abordagem temporal.

Segundo Lin, Orgun e Williams (2002), um algoritmo que busque padrões temporais, ou ajuste um modelo para tal, a partir de dados temporais é dito um algoritmo de mineração de dados temporais. Trata-se, de fato, de uma etapa do processo de *KDD* em bases de dados nas quais estão presentes informações temporais.

Definição 2.9 (Transação temporal)**.** Uma transação é dita temporal quando, em suas informações armazenadas, existe pelo menos um atributo temporal, tal como uma data.

Um exemplo da falta de consenso presente entre os autores dos trabalhos que lidam com a temporalidade pode ser visto quanto a consideração do ponto ou intervalo temporal. Um ponto temporal (ver Definição 2.10) descreve um instante, um momento que está diretamente relacionado a um evento.

Definição 2.10 (Ponto (ou instante) temporal)**.** Um *ponto temporal* descreve um instante de tempo t pontual. Ou seja, um instante (ou momento) que relaciona algum acontecimento instantâneo, sem duração.

O intervalo temporal (ver Definição 2.11) é constituído por um conjunto de um a infinitos pontos temporais e visa descrever um período com a duração da ocorrência de um evento.

Definição 2.11 (Intervalo temporal)**.** Um *intervalo temporal* é definido pelo conjunto infinito de pontos compreendidos entre dois pontos temporais limitantes $[-t, +t]$ que descrevem a duração de algum acontecimento. Necessariamente $-t$ ocorre antes de $+t$, ou seja, $-t \leq +t$.

Definição 2.12 (União de intervalos temporais)**.** Dados dois intervalos temporais $t_1 = [-t_1, +t_1]$ e $t_2 = [-t_2, +t_2]$, nos casos em que t_1 e t_2 compartilham pelo menos um ponto temporal, os intervalos temporais podem ser unificados a fim de compor um novo intervalo temporal. O novo intervalo temporal t_3 é construído com o menor $-t$ e o maior $+t$, i.e., $t_3 = t_1 \cup t_2 = [min(-t_1,\ -t_2),\ max(+t_1,\ +t_2)]$.

A adoção do ponto temporal ou intervalo temporal, assim como discutem Allen e Hayes (1985) é uma escolha que, de fato, pode mudar o resultado do processo conduzido. Winarko e Roddick (2007) afirmam que a maioria dos trabalhos relacionados à mineração temporal defende o uso de pontos temporais e não intervalos. Muitos trabalhos que

buscam regras de associação temporais estão disponíveis na literatura, como os conduzidos por Srikant e Agrawal (1996), Lu, Feng e Han (2000), Ozden, Ramaswamy e Silberschatz (1998), Li et al. (2001), que focam na busca de regras de associação que consideram pontos temporais. Como comentam Winarko e Roddick (2007) e Bohlen, Busatto e Jensen (1998), em muitos casos, certos acontecimentos ocorrem em um intervalo de duração e não são instantâneos, portanto são melhores de serem tratados como intervalos.

Definição 2.13 (Atributo temporal)**.** Um *atributo temporal* (notado por a_{temp}) descreve um valor de tempo que está relacionado à informação, tal como o instante em que um determinado evento ocorreu, quando foi armazenado, qual a sua duração, etc.

Exemplo 2.4. A evolução salarial de um funcionário da empresa *Fabrica A* pode ser armazenada em transações de uma base de dados da seguinte forma:

cod_funcionario; data_evolucao; novo_salario;...

3012; 2015-02-02; R$1.690,00;

3012; 2016-03-02; R$2.080,00;

3012; 2017-01-22; R$2.387,00;

Na qual cada transação é composta por três atributos, a saber: o código do funcionário, a data em que o reajuste foi realizado e o novo valor de salário. O atributo cod_funcionario é um atributo temático quantitativo não contínuo (ou seja, discreto), data_evolucao é um atributo temporal (a_{temp}) que indica o instante de tempo em que a evolução começa a vigorar, enquanto novo_salario é um atributo quantitativo contínuo (a_{cont}) que descreve o valor do salário que o funcionário (cod_funcionario = 3012) possui em cada instante de tempo (data_evolucao). É possível observar que o salário do funcionário apresenta valores no intervalo $[-v_{a_{cont}} = 1.690,00,\ +v_{a_{cont}} = 2.387,00]$.

2.4.1 Representação de informações temporais

Um outro aspecto que divide opiniões no processo de mineração de dados é a escolha pela adoção da temporalidade em sua forma explícita ou implícita. A temporalidade explícita é a forma que expressa a informação temporal com um atributo (ou campo) especificamente descritor de um valor temporal, tal como uma data, hora, intervalo de tempo, duração, etc. Já a temporalidade implícita considera que o simples encadeamento dos dados indica a ordem cronológica de ocorrência dos fatos armazenados. Nesta, não é necessário armazenar um valor extra para indicar o tempo.

Segundo afirmam Kirchgässner, Wolters e Hassler (2007), Chatfield (2003), séries temporais podem ser definidas como um conjunto de observações quantitativas ordenadas cronologicamente. Mitsa (2010) reforça que, de forma mais detalhada, uma série temporal pode ser entendida como uma sequência de observações cronologicamente ordenadas, com intervalos de tempo regulares entre cada par de observações.

Giusti (2017) defende que uma série temporal pode ser definida como a realização de um processo estocástico - um processo aleatório - que se desenvolve ao longo do tempo. Praticamente qualquer atividade ou fenômeno que pode ser imaginado incorpora informações temporais. Muitas vezes esse aspecto não é levado em consideração em uma análise ou processamento de dados pois não contribui, de forma efetiva, com a tarefa realizada. Entretanto em alguns casos a consideração do fator temporal associado ao dado é um facilitador da compreensão do problema.

Uma série temporal pode ser classificada em dois grupos: (a) séries unidimensionais (monovariadas) e (b) séries multidimensionais (multivariadas). A primeira delas é constituída quando uma série é composta por apenas uma variável; é multivariada quando é composta por um conjunto de variáveis. A temperatura aferida durante um período de tempo é um exemplo de série monovariada, enquanto a probabilidade (em valores entre 0 e 1) de um indivíduo estar resfriado leva em conta, além da temperatura corporal, níveis de coriza, quantidade de espirros e tosses durante um dia e contagem de glóbulos brancos. Portanto, trata-se de uma série multivariada, ou multidimensional. A Figura 2.3, abaixo, apresenta duas representações gráficas que identificam, mais à esquerda (imagem (a)) uma série unidimensional (monovariada) e mais à direita (imagem (b)) uma série multidimensional (multivariada).

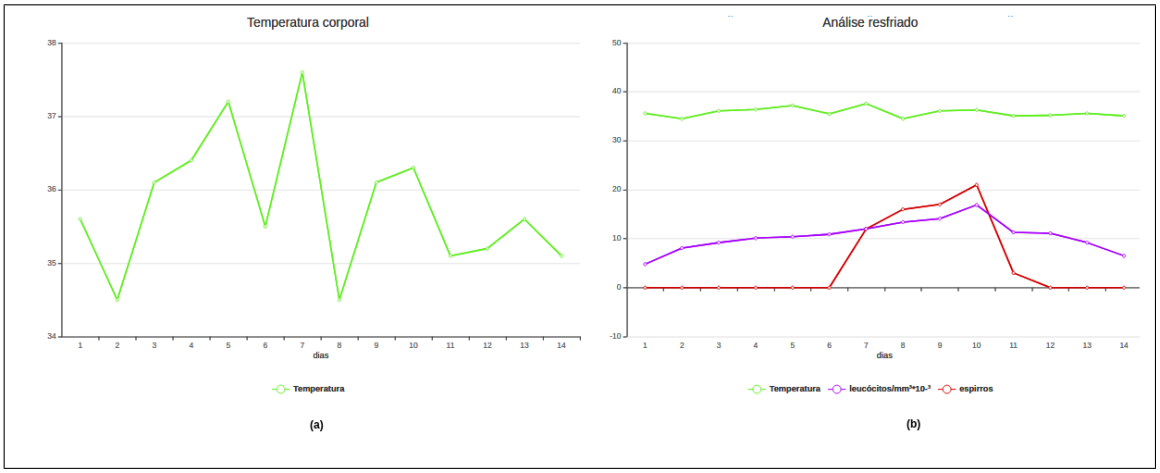
Bases de dados constituídas por séries temporais são utilizadas em diversas áreas de conhecimento, como em medicina, na realização de eletrocardiogramas; em economia, no acompanhamento das taxas de juros e ações na bolsa de valores; em agricultura, por meio do acompanhamento do vigor vegetativo das safras ao longo do ano e em meteorologia, para analisar a evolução de variáveis climáticas ao decorrer das décadas.

Adaptada da definição descrita por Amaral (2020), uma série temporal pode ser formalmente definida da seguinte forma:

Definição 2.14 (Série temporal)**.** Uma *série temporal*, indicada por $S = \{s_1, s_2, s_{...}, s_n\}|i \in [1..n]$, onde um fato f é descrito por um valor s_i quantitativo que ocorre em um instante t_i da função S sem duração, ocorre em um ponto.

Exemplo 2.5. A tarefa de *backup* automático, realizada pontualmente às 01h50, todos os

Figura 2.3: Representação gráfica de séries temporais. À esquerda, uma série unidimensional composta por valores de temperatura corporal e à direita, multidimensional composta por três séries unidimensionais que representam a informação resfriado.



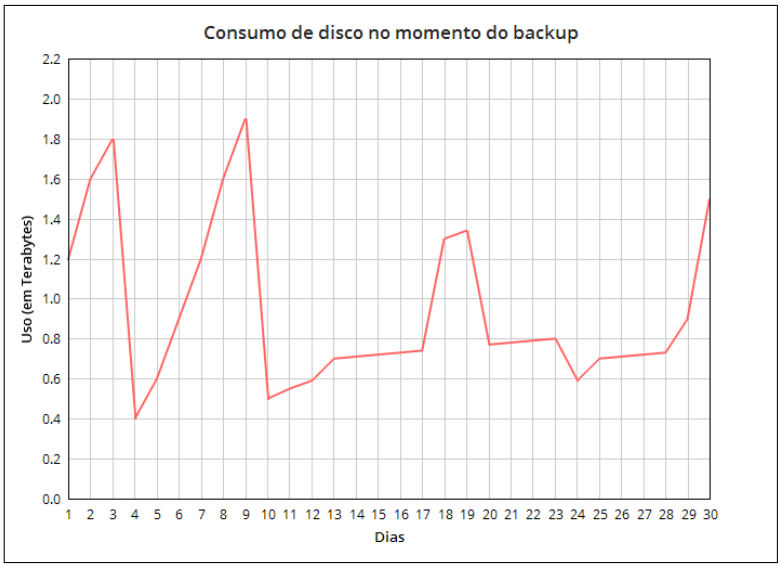
Fonte: Elaborada pelo autor.

dias, no servidor de uma empresa fictícia (imaginada para esse exemplo) gera um arquivo de controle (*log*) com várias informações sobre o estado do servidor. Uma das informações coletadas é a quantidade de arquivos armazenadas no disco rígido de *backup*, em *Terabytes*. Trata-se, então, de uma série temporal unidimensional que descreve o uso de disco do servidor, uma vez que os valores são gerados a cada intervalo de tempo, uniformemente espaçado. A série temporal pode ser formalmente escrita como $S = \{s_1, s_2, s_{...}, s_n\} = \{1,2,1,6,1,8,0,4,...,1,5\}$, graficamente ilustrada na Figura 2.4, apresentada logo abaixo.

Séries temporais são um caso particular dos dados temporais segundo Fu (2011). A temporalidade é dividida em temporalidade implícita e explícita. A Figura 2.5 ilustra a diferença entre a temporalidade explícita e a implícita. Na figura é possível observar que a organização dos dados no armazenamento, quando a temporalidade é expressa explicitamente, não requer um planejamento prévio para garantia de semântica, uma vez que para a recuperação de alguma informação uma consulta deverá ser realizada nos registros até que se encontre a informação correspondente; ao exato instante que se deseja. Já na adoção implícita, pode-se entender a organização dos registros como se respeitassem uma pilha, onde registros mais recentes são armazenados no topo da estrutura. É importante evidenciar que, ainda que as séries temporais sejam organizadas pelo encadeamento dos dados armazenados, a Definição 2.14 mostra que os dados armazenados respeitam intervalos idênticos entre si.

A escolha pela adoção da temporalidade explícita, ou implícita, deve ser guiada pela

Figura 2.4: Representação gráfica da série temporal utilizada no exemplo 2.5 - série temporal que descreve o uso diário de disco de um servidor.



Fonte: Elaborada pelo autor.

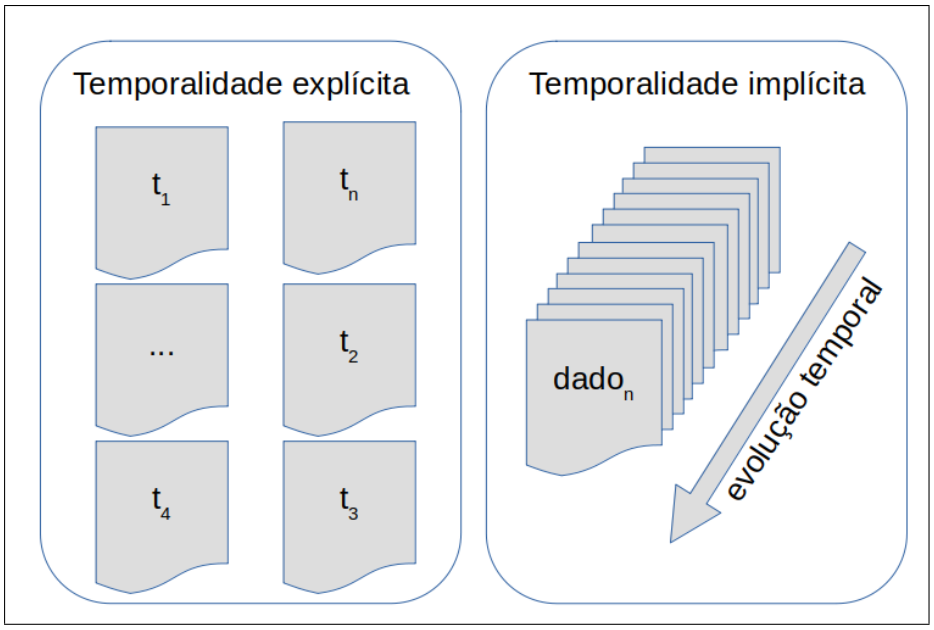
necessidade do problema. Não é mérito deste trabalho discutir sobre a superioridade de uma adoção sobre a outra. Entretanto nota-se na literatura, que a grande maioria dos trabalhos opta por lidar com dados encadeados em série. Devido a este simples fato, a dúvida sobre a possibilidade de contribuir com a evolução desta linha de ciência, por meio da consideração de uma abordagem menos utilizada, surgiu e inspirou grande parte deste trabalho.

Ao considerar informações temporais explícitas na base de dados, alguns benefícios podem ser obtidos. Tais como maior agilidade na recuperação de informações, bem como a garantia de que lacunas entre os dados não prejudiquem os resultados gerados. Isso ocorre pois é impossível identificar uma lacuna entre dados de uma série temporal, mas é fácil selecionar porções de dados em uma base de dados que contempla informações temporais explícitas.

2.5 A Álgebra intervalar de Allen (AIA)

Conforme afirmam Nicoletti, Lisboa e Hruschka-Junior (2013, p.134), "O aspecto temporal de situações e eventos é um aspecto decisivo para ser considerado na implementação de sistemas inteligentes". Muitos trabalhos focados em formas de representação temporal estão presentes na literatura, o que reforça essa afirmação. Tal como pode ser observado nos trabalhos de Ladkin (1986), Knight e Ma (1993), Bellini, Mattolini e Nesi (2000),

Figura 2.5: Representação gráfica da diferença entre as duas formas de adoção da temporalidade no processo de mineração de dados temporais, a saber: a temporalidade explícita (à esquerda) e a temporalidade implícita (à direita).



Fonte: Elaborada pelo autor.

Chittaro e Montanari (2000), Furia et al. (2010), entre outros.

A Álgebra Intervalar de Allen (AIA), proposta por Allen (1981) é um modelo matemático de representação temporal intuitivo que, segundo Allen e Ferguson (1994), tem como premissa básica um objeto primitivo denominado *intervalo de tempo*. Um intervalo de tempo é composto por infinitos pontos temporais ordenados, compreendidos entre dois pontos limitantes, $[-t, +t] \mid -t < +t$ cuja finalidade é descrever um tempo associado a algo ou um evento. Ainda segundo Allen e Ferguson, a AIA tem como principal objetivo o de indicar relacionamentos binários, por meio de uma relação denominada *meets*, entre dois intervalos de tempo.

Inicialmente foram descritas cinco possibilidades para relações entre os intervalos temporais, a saber: *meets, before, overlaps, during* e *equal*, como introduzido por Allen (1983). Entretanto, de acordo com Allen, a relação *during* na verdade é constituída por três possibilidades que ainda satisfazem a relação: (1) quando dois intervalos se iniciam no mesmo ponto temporal, entretanto um tem seu fim anteriormente ao outro; (2) quando um intervalo se inicia e termina enquanto outro ocorre; e (3) quando dois intervalos compartilham do mesmo ponto temporal para indicar seus términos, entretanto se iniciaram em pontos temporais distintos. Portanto, a relação *during* foi subdividida em três outras, a *starts*, a *finishes* e a *during* de fato.

A Tabela 2.2, adaptada daquela encontrada no trabalho de João, Nicoletti e Monteiro (2016) apresenta as sete relações básicas propostas pela AIA. Dois intervalos temporais genéricos são considerados a fim de expressar as relações da AIA, a saber, o intervalo $A = [a-, a+]$, $(a- < a+)$ e o intervalo $B = [b-, b+]$, $(b- < b+)$, tal que $\{a-, a+, b-, b+\} \in \Re$.

Tabela 2.2: Sete relações básicas da AIA, considerando os dois intervalos temporais $A = [a-, a+]$ $(a- < a+)$ e $B = [b-, b+]$ $(b- < b+)$ | $\{a-, a+, b-, b+\} \in \Re$.

Relação temporal entre A e B	Descrição temporal
$meets(A, B)$	$a- < b-$, $a- < b+$, $a+ = b-$, $a+ < b+$
$before(A, B)$	$a- < b-$, $a- < b+$, $a+ < b-$, $a+ < b+$
$overlaps(A, B)$	$a- < b-$, $a- < b+$, $a+ > b-$, $a+ < b+$
$during(A, B)$	$a- > b-$, $a- < b+$, $a+ > b-$, $a+ < b+$
$starts(A, B)$	$a- = b-$, $a- < b+$, $a+ > b-$, $a+ < b+$
$finishes(A, B)$	$a- > b-$, $a- < b+$, $a+ > b-$, $a+ = b+$
$equal(A, B)$	$a- = b-$, $a- < b+$, $a+ > b-$, $a+ = b+$

Fonte: Elaborada pelo autor.

Conforme ressaltam João, Nicoletti e Monteiro (2016), a segunda coluna da tabela que descreve as condições de existência das relações binárias entre dois intervalos temporais (definidos nesta tese por $[-t, +t]$) assume que sempre é verdade que $-t < +t$ e a vírgula representa o operador lógico E (ou AND).

Allen e Hayes (1985) descrevem, ainda, que as relações binárias propostas podem ser, no total, 13 quando são consideradas as suas relações inversas. Em outras palavras, toda relação xRy, tem uma inversa correspondente $xRy^{-1} = yRx$; são elas: $meets(A, B)^{-1} = met\_by(B, A)$, $before(A, B)^{-1} = after(B, A)$, $overlaps(A, B)^{-1} =$ *overlapped_by(B,A)*, $starts(A, B)^{-1} =$ *started_by(B,A)*, $finishes(A, B)^{-1} =$ *finished_by(B,A)* e, por fim, $during(A, B)^{-1} =$ *contains(B,A)*. Note que a relação $equal(A, B)$ possui como inversa ela mesma, ou seja, a inversa de uma relação de igualdade é a igualdade.

A AIA é um modelo linear de tempo que considera o intervalo de tempo como contínuo, o qual é constituído por infinitos pontos temporais. Allen (1983) reforça que a adoção de pontos temporais em oposição aos intervalos deve ser pensada cautelosamente. Assim como Mani, Pustejovsky e Sundheim (2004, p.06) afirmam, "dependendo da representação e da escolha da primitiva (intervalos, pontos ou ambos), uma variedade de diferentes relações entre tempos pode ser definida".

Tabela 2.3: Sete relações básicas da AIA e suas correspondentes relações inversas. No total 13 relações binárias são propostas pela AIA.

Relação binária	Relação inversa	Descrição temporal
$meets(A,B)$	$met\_by(B,A)$	$a- < b-$, $a- < b+$, $a+ = b-$, $a+ < b+$
$before(A,B)$	$after(B,A)$	$a- < b-$, $a- < b+$, $a+ < b-$, $a+ < b+$
$overlaps(A,B)$	$overlapped\_by(B,A)$	$a- < b-$, $a- < b+$, $a+ > b-$, $a+ < b+$
$during(A,B)$	$contains(B,A)$	$a- > b-$, $a- < b+$, $a+ > b-$, $a+ < b+$
$starts(A,B)$	$started\_by(B,A)$	$a- = b-$, $a- < b+$, $a+ > b-$, $a+ < b+$
$finishes(A,B)$	$finished\_by(B,A)$	$a- > b-$, $a- < b+$, $a+ > b-$, $a+ = b+$
$equal(A,B)$	$equal(B,A)$	$a- = b-$,$a- < b+$, $a+ > b-$,$a+ = b+$

Fonte: Elaborada pelo autor.

Considerações finais

Este capítulo apresentou alguns conceitos e o embasamento teórico relacionado ao tema deste trabalho. Foram apresentados os tipos possíveis de dados evidenciando a hierarquia da divisão dos tipos, uma forma de evidenciar e justificar a escolha da consideração dos dados quantitativos contínuos neste trabalho, uma vez que a manutenção dos dados em sua forma mais bruta permite a revelação de detalhes que podem ser omitidos quando tarefas de discretização de dados são conduzidas. Também foi discutido sobre a distribuição de probabilidade e sua importância para representar fenômenos aleatórios, tais como os que descrevem os dados a serem considerados pelo método desenvolvido ART-Q. Justificou, ainda, a escolha pela representação da forma explícita na temporalidade dos dados que permite ao método lidar com bases de dados não ordenadas cronologicamente, ao mesmo passo que possibilita a seleção de registros da BD sem o risco de dados faltantes prejudicarem o processo de mineração de dados. Por fim, a álgebra intervalar de Allen (AIA) foi apresentada como a forma usada pelo ART-Q para a representação das relações temporais que viabilizam a construção do tipo de padrão idealizado neste trabalho, mais complexo que os tipos existentes na literatura.

Capítulo 3

Revisão da literatura e trabalhos correlatos

Este capítulo apresenta uma discussão a cerca do conjunto de trabalhos que descrevem o atual estado da arte da mineração de regras de associação temporais em dados quantitativos contínuos, de forma a delinear as semelhanças e diferenças entre as ferramentas e estratégias do estado da arte com o método descrito por este trabalho. Particularmente, divididos em três grupos de estratégias, estão descritos nesta seção, os trabalhos que lidam com dados quantitativos contínuos, que envolvem a temporalidade dos dados (implícita ou explicitamente) ou atacam ambos os problemas de forma simultânea. Esta etapa foi apoiada pelo uso da ferramenta StArt (apresentada por Zamboni et al. (2010)), que possibilitou, além da busca em diferentes fontes de informação, uma maneira de definir critérios para a seleção, ou exclusão, dos trabalhos mais relacionados ao tema escolhido.

3.1 Considerações iniciais

Os trabalhos discutidos a seguir foram identificados em diferentes fontes de dados, por meio de uma sistemática assistira por uma ferramenta computacional que auxilia no processo de revisão sistemática. Os detalhes deste processo de levantamento bibliográfico estão descritos em maior nível de detalhes no Apêndice C.

3.2 Trabalhos que lidam com dados quantitativos contínuos

A grande maioria dos trabalhos na literatura que buscam regras de associação em dados quantitativos contínuos realiza um processo de categorização dos dados quando

necessário. Entretanto o método ART-Q foi idealizado para evitar a necessidade da realização de um processo de pré-processamento dos dados. De forma a ser robusto o suficiente para lidar com o dado em sua forma mais próxima à realidade possível. A seguir, são descritos os trabalhos selecionados que, segundo afirmam os seus respectivos autores, lidam com dados quantitativos contínuos sem que um processo de discretização (categorização) seja realizado previamente.

O trabalho de Aumann e Lindell (2003) é uma das referências mais exploradas quanto à busca de regras de associação estatísticas. Nele, os autores introduzem uma nova definição para regras quantitativas, com vistas à teoria de inferência estatística. Mais precisamente, na distribuição dos valores de atributos em um registo da base de dados. Um exemplo, apresentado no trabalho, da regra buscada pela proposta é:

$$sexo = feminino \Rightarrow salário\text{-}médio=\$7{,}90/hora\ (média\ geral\ salarial = \$9{,}02),$$

A partir da regra apresentada é possível afirmar que, na média, o salário para trabalhadoras do sexo feminino é de $7,90 por hora, enquanto a média geral de salários, sem distinção de sexo, é de $9,02 por hora. Essa regra é interessante pois revela a existência de um grupo que recebe um salário inferior à média geral. É importante notar que essa regra não é gerada com auxílio de nenhum método de discretização, mas lida com os atributos contínuos.

Segundo afirmam os autores, uma regra de associação tem vistas a identificar comportamentos de interesse em uma base de dados. Em atributos categóricos, o comportamento pode ser compreendido como uma lista de itens e seus respectivos valores de probabilidade de ocorrência. Nas regras de associação categóricas, um comportamento de interesse é identificado mais vezes do que certos atributos. Segundo afirmam os autores, essa é a definição estatística de distribuição de probabilidade de um conjunto de itens para uma dada população. Entretanto para valores numéricos contínuos, as melhores distribuições que representam seus comportamentos são a média e a variância de seus valores. Um subconjunto de uma população que apresenta uma distribuição significativamente diferente do restante, em termos de média e variância de seus valores, é dito interessante.

Ainda, segundo os autores, no trabalho não foi utilizado nenhum método de discretização dos dados e, mesmo assim, não sofreram com a explosão exponencial na quantidade de regras identificadas, como comumente ocorre. Os primeiros testes foram realizados com uma base de dados real de média salarial dos trabalhadores americanos (*Determinants of Wages from the 1985 Current Population Survey*), contendo 534 transações com

7 atributos categóricos e 4 quantitativos. Como resultados, foram obtidas regras que descrevem características sociais do período, tais como a que descreve que pessoas com menor grau de escolaridade são, geralmente, as mais velhas. Um sinal positivo de progresso na sociedade.

No trabalho, os autores realizaram o teste Z para o valor de média, a fim de estabelecer a significância da média. Uma regra é dita significante se a hipótese nula é rejeitada com confiança acima do limiar pré-definido. São considerados dois tipos de regras de associação como resultados do trabalho: (1) aquelas que são compostas por ambos os lados (antecedente e consequente) com um atributo quantitativo, tal como a regra *R: Educação* $\in [14, 18]$ *anos* $\Rightarrow$ *salário-médio=$18,64/hora*; e (2) aquelas que são compostas por um atributo categórico em seu antecedente e um quantitativo em seu consequente (como a regra apresentada anteriormente).

Para o teste em uma base de dados real, foram analisados dados de linguística, obtidas em um estudo de hábitos de escrita em inglês por não nativos. A base é composta por 643 transações contendo 15 atributos categóricos e 27 quantitativos. Os resultados obtidos (regras) foram apresentados a um especialista do contexto que categorizou-os como de grande importância. Apesar de comprovadamente eficaz, para k atributos quantitativos e n transações da base de dados, a estratégia tem complexidade computacional $O(k * n * long_n + k^2 * n)$.

Ainda que tenha sido uma grande influência para a idealização e implementação do método ART-Q, a estratégia descrita por Aumann e Lindell (2003) não permite a definição de comportamentos de interesse para cada um dos atributos. Fica, entretanto, restrita a buscar informações sempre dentro do comportamento descrito como normal aqui neste trabalho.

Haldulakar e Agrawal (2011) comentam que algoritmos genéticos são utilizados para buscas em geral. Eles buscam uma solução ótima para um determinado problema, tendo como base um conjunto de soluções - uma população que é evoluída a cada ciclo de execução. No trabalho, os autores afirmam que em casos nos quais atributos contínuos são introduzidos na tarefa de mineração de regras de associação, estes precisam ser discretizados, selecionando pontos de corte para dividir os valores dos atributos contínuos em intervalos. Desta forma, a estratégia de algoritmos genéticos é utilizada para lidar com os dados quantitativos contínuos, mais precisamente para identificar os pontos de corte pelos quais serão criados intervalos para os valores quantitativos contínuos.

O algoritmo apresentado por Haldulakar e Agrawal (2011) lança mão de técnicas

embasadas em algoritmos genéticos logo após a execução do conhecido algoritmo Apriori, com vistas a incorporar, também, aquelas regras que foram descartadas pelo Apriori mas que podem contribuir com a evidenciação de mais informações.

Inicialmente, no trabalho, os valores quantitativos contínuos são codificados convertendo-os em sequências de bits (zeros e uns) e, então, técnicas de algoritmos genéticos são empregadas, repetidamente, a fim de selecionar os melhores indivíduos para representar os valores dos atributos. Após esta etapa, o algoritmo Apriori é executado para a obtenção do conjunto de regras de associação. A partir deste ponto, aquelas regras ditas fortes são armazenadas e, posteriormente combinadas às ditas fracas após passarem por uma etapa de refinamento. Etapa esta de refinamento é responsável por submeter as regras fracas (descartadas pelo Apriori) a um algoritmo genético que emprega uma função de aptidão modificada.

Esta técnica é empregada a fim de selecionar as regras que seriam descartadas mas são potencialmente interessantes. Os autores destacam, ainda, que a parte mais importante do algoritmo genético é dada pela função de aptidão, definida pela razão entre o suporte da regra gerada e o suporte mínimo. A variação na função de aptidão é dada pelo fato de que as regras que não superam o suporte mínimo, mas estão próximas do valor de suporte mínimo estabelecido, não são excluídas do conjunto resultante. Elas são verificadas novamente quanto às que superam o valor de suporte mínimo em 0,5. Nesta etapa, valores de bits são mudados para seus opostos, randomicamente. Como condição de parada do algoritmo é definido que deva ser executado enquanto não atinja uma quantidade pré-definida de regras ditas ótimas globais. Com a modificação proposta no trabalho, uma quantidade menor de regras foi obtida. Os autores defendem que a mineração de regras considerando a função de aptidão modificada no trabalho garante um resultado melhor ao final.

O fato de limitar a busca por uma quantidade pré-definida de regras a serem buscadas diverge do princípio do ART-Q, de buscar a maior quantidade possível de informações implícitas na base de dados.

No trabalho de Alvarez e Vazquez (2012) é apresentada uma ferramenta não supervisionada para identificar regras de associação que lidam com atributos quantitativos contínuos e discretos sem a necessidade de uma discretização prévia dos dados. Os autores defendem a não discretização dos atributos devido ao fato que os intervalos são obtidos pelo processo evolutivo por si só, durante a etapa de aprendizado. A ferramenta, intitulada *GARplus* é embasada na teoria de algoritmos evolutivos, de Goldberg (1989) e faz uso do conceito de intervalos, assim como definido por Corcoran e Sen (1994). A

ferramenta se propõe a identificar as regras de associação discretas de maior importância, geradas a partir de uma base de dados com atributos discretos ou numéricos.

No trabalho, cada indivíduo corresponde a uma regra de associação, no qual os genes são os valores máximo e mínimo dos intervalos de cada atributo numérico que compõem a regra, ou o conjunto de valores que cada atributo discreto pode assumir. As regras de associação geradas pelo *Garplus* foram inspirações para a construção do ART-Q, uma vez que podem ser compostas por vários atributos em suas partes antecedentes e consequentes.

A função de aptidão que a ferramenta emprega, considera além dos valores de suporte e confiança, medidas como número de atributos, amplitude média dos valores numéricos dos atributos que compõem a regra, amplitude médias dos valores discretos que compõem a regra e contagem de regras as quais cada atributo faz parte. Todos os valores são, então normalizados para terem a mesma representatividade no cálculo da aptidão. Os atributos são diferenciados entre os que compõem a parte antecedente e os que compõem a parte consequente das regras - processo realizado somente ao final, de forma randômica. Os atributos que não compõem nenhuma regra de associação são descartados e não armazenados na memória.

As regras geradas pelo *Gar-plus* apresentam-se da seguinte forma: *"Se* $c_6 \in$ *[1,5] e* $c_7 \in (v_1, v_2, v_3)$ *então* $c_8 \in$ *[8,10]"*, onde pode ser observado que os atributos numéricos são associados a seus intervalos enquanto os discretos às possibilidades finitas de valores que podem assumir. Ao final dos experimentos, os autores concluem que os resultados foram satisfatórios nos vários testes realizados. Ainda, que o algoritmo mostrou-se escalável para bases de dados muito grandes, uma vez que nem todos atributos precisam ser armazenados na memória.

Yang (2010) é outro exemplo de estratégia que faz uso de algoritmos genéticos (AG), mais precisamente um mecanismo imunológico aplicado ao AG para lidar com atributos contínuos. Segundo comenta o autor, métodos tradicionais de mineração de regras de associação utilizando-se de algoritmos genéticos são suscetíveis a perda de informação, uma vez que realizam o processo de discretização dos dados de forma independente e anteriormente ao processo de mineração de regras. O algoritmo descrito no trabalho, intitulado MAR_IGA, integra o processo de discretização dos valores contínuos, a redução de atributos e a extração de regras ao mesmo tempo.

O sistema imunológico faz uso das medidas de aptidão e concentrações dos indivíduos. Assim, os melhores podem ser selecionados para criar uma nova geração. Quanto maior a aptidão calculada de um indivíduo, maior a probabilidade dele ser selecionado para

compor uma nova geração. Quanto maior a concentração de um indivíduo é, menor é a chance dele ser selecionado. Isto garante que haja diversidade na população.

As regras de associação são consideradas úteis de acordo com a função de aptidão e os operadores genéticos de mutação e cruzamento. Uma regra é considerada forte sempre que superar um limiar (confiança) mínimo definido pelo usuário. Para a realização dos experimentos, o autor faz uso da base de dados Abalone, obtida do repositório de aprendizado de máquina UCI, com 4177 instâncias - 1 atributo discreto e 8 contínuos. Os resultados obtidos foram comparados aos obtidos pelo algoritmo genético clássico (SGA) e o Apriori, sob as mesmas condições.

Zheng et al. (2014) descrevem a proposta e implementação do método OFARM (*Optimized Fuzzy Association Rule Mining*) para mineração de regras de associação em dados quantitativos contínuos, por meio da lógica Fuzzy. Segundo descrevem os autores, as vantagens do OFARM são: (1) propor um novo método para adicionar nebulosidade e flexibilidade nas funções de pertinência em conjuntos fuzzy, (2) otimizar conjuntos Fuzzy e seus pontos de partição com funções objetivas múltiplas depois de categorizar os dados quantitativos e (3) modelar uma iteração de dois níveis para filtrar conjuntos de itens frequentes e regras de associação fuzzy.

No trabalho, os autores afirmam que quando conjuntos fuzzy são criados, existem pontos de partição que nem sempre representam a realidade. Assim como ocorre na grande maioria dos trabalhos, tais pontos de partição precisam ser otimizados por métodos de otimização e o conhecimento especialista, para aumentar a precisão dos conjuntos fuzzy. O OFARM refina o processo de criação de regras de associação considerando, além da medida de confiança, outras medidas tais como a convicção, interesse e fator de certeza.

O OFARM difere da estratégia tradicional por adicionar uma iteração de dois níveis para otimizar o conjunto mínimo inicial de itens frequentes e regras de associação. O algoritmo otimiza as regras de associação por meio do refino das partições dos conjuntos fuzzy e busca de padrões frequentes, repetitivamente.

O trabalho apresenta uma comparação dos resultados obtidos pelo OFARM com o algoritmo GFARM (*General Fuzzy Association Rules Mining*) Delgado e others. (2003), seu inspirador. O experimento considera três bases de dados contendo diagnósticos e prognósticos de câncer e diabetes, extraídas do repositório UCI e configurações diferentes para os valores de suporte mínimo, confiança e fator de certeza.

O trabalho é concluído com a afirmação que o método proposto OFARM supera seu

inspirador GFARM em ambos os aspectos, quantitativamente e qualitativamente, isso graças a otimização dos conjuntos de pertinência fuzzy, promovida pela função objetiva múltipla e a iteração em dois níveis. Os experimentos, com variações de suporte e confiança mínimos demonstraram que o OFARM é robusto e estável. Entretanto, como afirmam os autores, o trabalho não mantém os dados em sua forma quantitativa contínua, mas sim aplicam uma estratégia para categorizar os dados quantitativos.

3.3 Trabalhos que consideram o aspecto temporal

Nesta seção são detalhados os trabalhos identificados e selecionados na literatura, que atuam na tarefa de mineração de regras de associação temporais. É possível notar que a relação de trabalhos contempla, também, alguns estudos que levam em consideração a temporalidade não explícita dos dados, i.e., trabalhos que assumem que a informação temporal dá-se pelo fato do sequenciamento dos dados. Essa consideração foi permitida para que uma melhor comparação entre os trabalhos pudesse ser realizada, uma vez que os trabalhos que lidam com a temporalidade explícita dos dados representa uma parcela bem menor dos trabalhos disponíveis na literatura.

O algoritmo ARMADA (An algorithm for discovering Richer relative teMporal Association rules from temporal DAta) proposto por Winarko e Roddick (2007) é um algoritmo de mineração de padrões sequenciais que estende o algoritmo MEMISP, de Lin e Lee (2002). O ARMADA tem dois objetivos específicos: (1) encontrar padrões temporais e (2) gerar regras de associação que correlacionam os eventos contidos na base de dados levando em consideração a temporalidade associada ao dado, i.e., temporalidade implícita, o que o difere do método ART-Q.

O algoritmo ARMADA busca padrões sequenciais temporais e, para tal, considera que *itemsets* são eventos. Cada evento ocorre em um intervalo de tempo e, portanto, possui dois pontos temporais a ele associados, que determina a duração do evento. Para tanto, adota a definição de sequência de estado, proposta por Höppner (2001).

As relações entre os intervalos de tempo, consideradas pelo algoritmo, são baseadas na lógica temporal intervalar de Allen (1981, 1983). Assim como o seu sucessor (o MEMISP), o ARMADA não gera candidatos e tampouco faz projeção de base de dados. Faz, entretanto, a indexação de memória para organizar os padrões e suas respectivas posições na base de dados. Nota-se, portanto, que o ARMADA considera a temporalidade implícita nos dados, diferente do que faz o ART-Q, mas ainda assim o ARMADA foi uma das

maiores inspirações para a elaboração do método ART-Q, uma vez que ambos consideram as relações da AIA.

Inicialmente, o algoritmo executa uma única varredura inicial na base, pela qual a copia para a memória interna, calculando a frequência de todos os itens unitários frequentes. Na segunda etapa de execução do ARMADA, um conjunto de índices é gerado com o objetivo de associar cada uma das sequências frequentes à sua posição na base de dados. Na terceira e última etapa do algoritmo, novas buscas são feitas e os conjuntos de índices são atualizados de forma recursiva enquanto padrões frequentes são buscados pelo algoritmo. Quando a base de dados é muito volumosa para ser armazenada inteiramente na memória, o ARMADA a divide em partes, por meio do procedimento *partition-and-validation* e realiza a mineração em cada uma delas.

No trabalho, um novo conceito, o *maximum_gap*, é introduzido, a fim de refinar a identificação de padrões temporais reduzindo o volume de eventos encontrados. O maximum_gap define a janela máxima permitida entre a ocorrência de dois eventos identificados, para que possam ser considerados frequentes. Outro conceito que motivou a condução deste trabalho, especialmente à definição dos parâmetros MWI e MWR.

Segundo afirmam os autores, embora seja verdade que há a necessidade de espaço suficientemente grande para armazenar a base de dados e o índice na memória, no entanto, o tamanho do conjunto de índice fica menor a medida em que os prefixos para criar o conjunto de índice ficam maiores (envolvem mais eventos).

João, Nicoletti e Monteiro (2016) apresentam um sistema computacional capaz de identificar padrões frequentes em uma base de dados composta por transações de venda de varejo e, por meio dos padrões identificados, inferir regras de associação que incorporam o aspecto temporal dos dados. O algoritmo proposto, denominado S_MEMISP+AR (*System MEMISP + Association Rules*), assim como o ARMADA de Winarko e Roddick (2007) estende o algoritmo MEMISP (*Memory Indexing for Sequential Pattern Mining*) proposto por Lin e Lee (2002), entretanto o S_MEMISP+AR incorpora o tratamento temporal em uma base de dados real.

Assim como o MEMISP, o S_MEMISP+AR realiza apenas uma varredura na base de dados, a copiando completamente para a memória, então busca os padrões frequentes de forma recursiva com ajuda de uma tabela de índices que associa cada padrão identificado. Ao final da tarefa de identificação dos padrões, o S_MEMISP+AR faz uso de uma variação do algoritmo Apriori Agrawal e Srikant (1994) para a tarefa de construção das regras de associação temporais.

Após uma etapa de pré-processamento da base de dados, as transações de venda contidas nela foram convertidas em conjuntos de itens ordenados alfabeticamente e dispostos em séries temporais de conjuntos de itens, a fim de representar todas as vendas realizadas no período de forma realista. Duas abordagens foram consideradas para os experimentos: (1) a primeira para buscar padrões associados aos dias da semana, a fim de identificar padrões semanais e (2) a segunda que considera intervalos de estados, seguindo a definição de Höppner (2001) com a finalidade de representar as relações temporais detectadas por meio da álgebra intervalar de Allen (AIA) Allen (1981, 1983). Além de não considerar dados quantitativos contínuos em sua execução, o S_MEMISP+AR diverge do método ART-Q ao considerar a temporalidade de forma implícita nos dados.

No trabalho descrito por Akhlagh, Tan e Khak (2012) um algoritmo de classificação baseado em árvores de decisão temporais é apresentado. Para a classificação de dados, um atributo alvo é chamado atributo de decisão ou de classificação e os demais atributos são chamados atributos de condição. O atributo de decisão é predito por meio de processos aplicados aos atributos de condição. Regras são criadas com os atributos de condição, que determinam certas relações existentes entre os valores dos atributos de condição com o valor do atributo de decisão.

Na grande maioria dos trabalhos dessa abordagem, as relações dos atributos de condição são consideradas em um mesma transação da base de dados. Entretanto podem existir relações inter transações quando considerado o sequenciamento das transações. Desta forma, um atributo classe pode ser estimado levando em consideração, também, os atributos condicionais de transações anteriores. Provê, assim, um resultado mais preciso na classificação. Os autores ressaltam que, para essa abordagem ser contributiva, é necessário que as transações respeitem uma ordem no seu sequenciamento.

No trabalho é apresentado um procedimento denominado temporalização, pelo qual transações consecutivas em uma base de dados são unificadas, respeitando uma janela temporal deslizante w pré-definida. Este procedimento constrói-se novas transações, nas quais a temporalidade é assumida (implicitamente) no sequenciamento, ou seja, a ordem das transações unificadas é a instanciação da temporalidade. Não é assumida, portanto, a temporalidade explícita dos dados, não é preciso que a base de dados possua atributos com informações temporais, tais como data, hora, etc.

Uma vez unificadas, em uma transação temporalizada, existem valores para os atributos em w tempos diferentes. Neste procedimento, os atributos de decisão das transações unificadas são descartadas, exceto pela última transação, esta definirá o valor de decisão,

i.e., a classe. Quando um nó folha é multi-valorado, ou seja, a quantidade de atributos condicionais não é suficientemente grande para definir apenas uma classe no nó folha, a estratégia para a tomada de decisão é feita aplicando a probabilidade condicional de Bayes.

Segundo concluem os autores, nos experimentos realizados o procedimento da temporalização para unir transações consecutivas provou ser uma estratégia melhor que o uso das árvores de decisão convencionais. Por este motivo, a estratégia de temporalização foi considerada para a definição da estratégia utilizada pelo método ART-Q de considerar relações entre intervalos temporais de diferentes atributos da BD.

A Temporal Association Rule Mining Algorithm Based on Attribute Reduction (ARTAR) como é denominado o algoritmo proposto por Ni et al. (2016) que explora uma forma de minerar regras de associação temporais combinando tecnologias de computação paralela e a teoria dos conjuntos aproximados. O ARTAR foi idealizado para lidar com dados de alta dimensionalidade, empregando a redução de atributos para diminuir a quantidade de dados em cada transação da base de dados.

A mineração de regras empregada pelo ARTAR é dividida em três principais fases, a saber: (1) Pré-processamento dos dados: pela qual tarefas como limpeza de dados, redução de ruídos e a redução de atributos são implementadas; (2) Mineração de padrões frequentes: na qual os algoritmos Apriori e TFP-growth são empregados para identificar padrões frequentes; e (3) Mineração de regras temporais: onde as constantes temporais e os padrões identificados são utilizados para minerar as regras que respeitam o valor de confiança mínima no intervalo.

Ou seja, a partir de um conjunto mínimo de atributos que represente a transação, dada alguma característica de interesse, o ARTAR emprega o algoritmo TFP-growth, juntamente com o Apriori para construir uma árvore de decisão, na qual os nós são compostos por padrões frequentes, respeitando um valor pré-definido de suporte mínimo e um intervalo temporal. Posteriormente, os caminhos da árvore são consultados e seus valores de confiança calculados, caso superem aquele pré-definido, este é considerado como uma regra de associação temporal. Pelo simples fato de realizar uma grande quantidade de tarefas de pré-processamento dos dados para reduzir a dimensionalidade, o ARTAR se diverge do método ART-Q. Entretanto foi um grande inspirador quanto à geração de regras de associação temporais.

Como (LEE; CHEN; LIN, 2003) discorrem em seu trabalho, no qual o algoritmo minerador de partições progressivo "*Progressive Partition Miner*"(PPM), a tarefa de mi-

neração de regras de associação temporais é custosa. Para tal, os autores propõem uma estratégia de particionamento da base de dados e identificação de padrões temporais em cada partição. Numa primeira varredura à base de dados, o PPM particiona a base de dados e identifica os padrões em cada partição, que posteriormente são combinados para compor padrões globais. Entretanto os dados que o PPM é apto a lidar são discretos, o que impede uma comparação direta com o método descrito nesta tese.

3.4 Trabalhos que consideram o aspecto temporal e lidam com dados quantitativos contínuos

A seguir, são apresentados os dois trabalhos selecionados mais recentes que lidam com os dois desafios estudados nesta pesquisa de doutorado, a temporalidade e a manipulação de dados quantitativos contínuos. Embora os trabalhos selecionados empreguem estratégias não tão próximas às deste trabalho, são os trabalhos mais recentes identificados e, portanto, representam o estado da arte.

No trabalho realizado por Chen et al. (2016) é proposto um algoritmo para mineração de regras de associação temporais fuzzy (FTARM), ou seja, regras de associação que lidam com dados quantitativos contínuo, considerando o tempo de duração dos itens da base de dados, denominado *vida útil*.

Segundo comentam os autores, as duas contribuições do trabalho são: (1) o FTARM é o primeiro algoritmo de mineração de regras de associação fuzzy que considera a *vida útil* dos itens e (2) regras de associação fuzzy mais úteis podem ser obtidas em termos da média de suporte e confiança, devido aos experimentos realizados em bases de dados reais e sintéticas. No trabalho, conceitos como itens, transações, suporte e confiança são redefinidos para incorporarem a lógica fuzzy e o aspecto temporal, tornando-se: item temporal, base de dados de transações temporais, item quantitativo temporal, suporte temporal fuzzy e confiança temporal fuzzy.

Como entrada do algoritmo é necessário que cada item nas transações da base de dados já possua sua *vida útil* definida. As transações da base de dados são transformadas para a representação fuzzy e a *vida útil* de cada um dos itens das transações é coletado. Os valores de suporte dos itens unitários (e seus graus de pertinência) são calculados e os com suporte superior ao limiar pré-definido são selecionados para gerar os candidatos de tamanho 2, de forma similar àquela executada pelo algoritmo Apriori original. Ao final destes passos, os padrões frequentes obtidos são utilizados para construir regras de associação

temporais fuzzy, combinando os itens que compõem o padrão nos papéis de antecedente e consequente das regras. Considerando o padrão-2 frequente *(A.Low, D.High)*, as possíveis regras temporais fuzzy são da seguinte forma: *if D.High, Then A.Low* e *if A.Low, Then D.High*. Cada uma das possibilidades é, então, verificada quanto aos seus valores de confiança e as que excedem o limiar pré-definido são consideradas para o resultado final da execução do FTARM.

Os autores concluem que a estratégia baseada no Apriori, proposta por eles, é capaz de gerar mais regras fuzzy que seu antecessor, o FAR. A contribuição mais importante, segundo destacam os autores, é a definição da *vida útil* dos itens que refina o processo de mineração de regras de associação fuzzy. Entretanto alguns problemas ainda precisam ser solucionados, tais como escolher uma função de pertinência apropriada para cada um dos itens.

O conceito de vida útil, idealizado pelos autores foi essencial para a construção e utilização do conceito de tempo de vida de uma regra de associação, utilizado pelo ART-Q.

Wang, Yang e Muntz (2001) descrevem um modelo para regras de associação temporais que envolvem atributos numéricos, especificamente para a mineração de regras de associação que capturam correlações entre as evoluções temporais de atributos numéricos, assim como afirmam os autores. Uma evolução temporal descreve as mudanças temporais dos valores de atributos de um dado objeto, como por exemplo funcionários que possuem aumento salarial anual. Evoluções temporais de atributos podem ser mapeadas para pontos de alta dimensão. A distancia entre dois pontos de um mesmo atributo representa a disparidade temporal (evolução) do atributo. Quando muitos atributos de um mesmo objeto têm distancias curtas, assume-se que há um agrupamento, portanto deve haver um padrão de comportamento nos objetos similares.

No trabalho, é assumido que uma base de dados consiste em um conjunto de objetos, cada um deles é descrito por um identificador único e um conjunto de atributos numéricos variando ao longo do tempo. Por exemplo, cada funcionário armazenado na base de dados pode ser um objeto com um conjunto de atributos associados ao identificador, como idade, salário, altura, etc. Os instantes de armazenamento dos atributos de cada objeto são obtidos respeitando uma determinada frequência e cada atributo é registrado em um mesmo instante de tempo. Esse tipo de consideração dos elementos que compõem a base de dados foi uma das inspirações para a construção da $BDRelT$, utilizada pelo ART-Q e que contempla relações temporais construídas a partir dos intervalos de interesse dos

atributos da base de dados.

Uma regra de associação temporal é definida por $R : X \Longleftrightarrow Y$, onde X é a conjunções de evoluções de atributos, i.e., $E(A_1) \cap E(A_2) \cap ... \cap E(A_k)$ e Y só diz respeito a evolução de um único atributo. Um exemplo de regra de associação temporal de tamanho igual a 2, conforme exemplificam Wang, Yang e Muntz (2001) tem a seguinte forma:

$$R_i : (salario \in [40000, 55000]) \rightarrow (salário \in [40000, 50000]) \Longleftrightarrow$$

$$(despesas \in [10000, 15000]) \rightarrow (despesas \in [10000, 12000]))$$

Além das medidas tradicionais de suporte e confiança da regra de associação, o modelo faz uso da densidade, que estabelece uma conexão natural entre uma regra de associação temporal e um agrupamento baseado em densidade. A densidade atua, também, como um mecanismo eficiente de redução do espaço de busca. Segundo os autores, é uma garantia de que a regra é válida durante todo o intervalo de tempo selecionado.

A estratégia empregada pelo TAR para identificação de regras de associação temporais, basicamente é: Após particionar cada domínio de atributo em intervalos de base, encontrar todos os grupos em relação ao limiar de densidade pré-definido e de acordo com os grupos identificados, buscar todos os conjuntos de regras válidas com uso do algoritmo Apriori de Agrawal e Srikant (1994).

Mais recentemente, trabalhos como o de Chamazi e Motameni (2019), Moslehi, Haeri e Martínez-Álvarez (2019) e Telikani, Gandomi e Shahbahrami (2020) descrevem o atual estado da arte da mineração de regras de associação. Telikani, Gandomi e Shahbahrami (2020) disponibilizam um levantamento bibliográfico de estratégias que consideram dados quantitativos contínuos ou a temporalidade em meio as descritas pelo trabalho. Entretanto o estudo demonstra que a grande maioria dos trabalhos recentes seguem a linha dos algoritmos genéticos como aposta de melhorias às estratégias. Na grande maioria dos trabalhos relacionados, técnicas de discretização dos dados quantitativos são empregadas e nos que incorporam a temporalidade, séries temporais são consideradas.

No trabalho de Moslehi, Haeri e Martínez-Álvarez (2019) é proposta uma mineração de regras de associação quantitativas por meio da integração dos resultados de uma estratégia híbrida que considera duas heurísticas, a habilidade de exploração de algoritmos genéticos (AG) com a alta velocidade para convergência da otimização de partículas de enxames (PSO (*Particle Swarm Optimization*)).

O objetivo do trabalho concentra-se em identificar a melhor combinação de suporte e

confiança para gerar a melhor quantidade regras de associação. Para os testes os autores consideram o uso de cinco bases de dados quantitativas, entretanto normalizadas. Os resultados gerados são, então comparados a outros trabalhos que lançam mão de AG para busca de regras de associação e, segundo os autores, demonstram a superioridade aos trabalhos anteriores da mesma linha.

Chamazi e Motameni (2019) descrevem o que pode ser considerado entre os trabalhos recentes, o mais próximo do ART-Q. Os autores discorrem no trabalho que o foco do estudo concentra-se na mineração de regras de associação em dados quantitativos contínuos que envolvem a temporalidade, por meio da consideração de conjuntos fuzzy.

Os autores defendem que algoritmos genéticos (AG) e o conhecido enxame de abelhas podem ajudar na seleção dos melhores conjuntos fuzzy para os atributos quantitativos contínuos. Em outras palavras, conduzem um processo automatizado para identificar os grupos da discretização dos dados. No trabalho o conceito de *lifespan* também é explorado, assim como no ART-Q. Para os autores, o *lifespan* descreve o tempo esperado que um item temporal deve ocorrer, a partir de um dado ponto temporal. Entretanto no trabalho, a base de dados é particionada em períodos que contemplam conjuntos de transações (registros de BD) e o *lifespan* dos itens é estimado pela soma da quantidade de transações dos períodos nos quais o item ocorre.

Embora a terminologia se assemelhe à do ART-Q, os resultados dos trabalhos não servem como parâmetro de comparação, uma vez que o ART-Q manipula os itens da base de dados para construir intervalos de interesse que serão analisados quanto à sua frequência. Ou seja, considera a busca por um padrão com estrutura totalmente diferente de todos os presentes na literatura.

Considerações finais

Neste capítulo foram apresentados os trabalhos identificados na literatura que têm o maior grau de relação com o tema de pesquisa discutido neste trabalho. O levantamento bibliográfico foi apoiado pelo uso da ferramenta StArt que possibilitou a busca em diferentes fontes de informação. Especificamente são detalhados os trabalhos que lidam com dados quantitativos contínuos, envolvem a temporalidade dos dados (implícita ou explicitamente) ou atacam ambos os problemas de forma simultânea. Dois dos trabalhos selecionados, mais precisamente os apresentados na Seção 3.4 representam o estado da arte da área de mineração de regras de associação temporais que lidam com a temporalidade dos dados. Ainda que os trabalhos tenham sido grandes

inspiradores na construção do método ART-Q, descrito por este trabalho, nenhum deles se aproxima a ponto de estabelecer uma comparação direta dos resultados obtidos, uma vez que não consideram a identificação de intervalos de interesses dos atributos e não lidam com padrões complexos como os identificados pelo ART-Q.

Capítulo 4

O Método ART-Q

*Neste capítulo são apresentados os detalhes do desenvolvimento do método **ART-Q**: **A**ssociation **R**ules involving **T**emporality and **Q**uantitative continuous data, para a mineração de regras de associação que envolvem dados quantitativos contínuos e a temporalidade do dado. O método ART-Q foi idealizado para assumir a temporalidade explícita na construção das regras de associação, ao mesmo passo em que lida com bases de dados compostas por dados quantitativos contínuos, sem a necessidade de discretizá-los. O capítulo apresenta um conjunto de definições necessárias para o entendimento do método ART-Q na seção de materiais e métodos, seguida pela seção que descreve o detalhamento do método desenvolvido e como o ART-Q implementa suas etapas de execução para atingir o objetivo de estabelecer uma nova forma de mineração de dados com padrões mais ricos em informação e regras de associação mais contributivas. Além de fornecer uma estratégia inovadora na definição de comportamentos de interesse e dos atributos e identificação dos intervalos temporais os quais os atributos assumem seus comportamentos de interesse.*

4.1 Materiais e métodos

Embora a lida com valores quantitativos contínuos no processo de mineração de dados, unida à incorporação do tratamento temporal de forma explícita, seja uma tarefa difícil, a estratégia que é empregada pelo ART-Q é intuitiva. A partir de uma base de dados composta por atributos contínuos e que contém pelo menos um atributo temporal, o ART-Q identifica pontos temporais em que os atributos contínuos apresentam um comportamento de interesse. A união desses pontos temporais imediatamente vizinhos permite a construção de intervalos temporais que os contemplam. Por sua vez, os intervalos podem ser melhores descritos por meio da AIA, com suas relações temporais.

Entretanto para a melhor compreensão das etapas de execução do método desenvolvido, algumas definições são necessárias:

Definição 4.1 (Comportamento de interesse)**.** Dado um atributo quantitativo contínuo (a_{cont}), um *comportamento de interesse* de a_{cont} é identificado quando o atributo assume valores dentro (ou fora) de um intervalo numérico $[-v, +v]$ que são mais representativos para uma determinada análise.

Definição 4.2 (Representação do comportamento de interesse)**.** O comportamento de interesse de cada um dos atributos pertencentes à BD pode ser representado por uma distribuição de probabilidade. A média (μ) e o desvio padrão (σ) dos valores são considerados para estabelecer os limites que os comportamentos de interesse possuem. Por padrão e quando o comportamento de um atributo quantitativo contínuo é desconhecido, é possível assumir que ele siga a distribuição normal de probabilidade.

Definição 4.3 (Intervalo de interesse de atributos quantitativos contínuos)**.** Um intervalo de interesse de um atributo é o espaço temporal no qual o atributo assume valores de seu comportamento de interesse. Um intervalo de interesse é identificado pela tripla ($-t_x$, x, $+t_x$), na qual $-t_x$ é um ponto temporal que identifica o início de uma ocorrência de um intervalo de interesse, x refere-se ao atributo assumindo valores no comportamento de interesse e $+t_x$ é o ponto temporal que identifica o fim do intervalo. Um intervalo temporal pode ser constituído simplesmente por um único ponto. Nestas situações, $-t_x = +t_x$.

Exemplo 4.1. Ao realizar leituras de valores de temperaturas durante um ano, nota-se que nos meses de junho, julho e agosto os valores são, em média iguais a $18,7°$. Entretanto observou-se que em 13 dos primeiros 15 dias do mês de julho, as temperaturas foram acima da média, em torno de $21,4°$. Se considerado que o comportamento de interesse do atributo temperatura ocorre quando seus valores fogem do padrão esperado, um intervalo temporal [01-07-2015,15-07-2015] pode ser construído para representar o período em que o atributo *temperatura* assumiu valores dentro do seu comportamento de interesse. Portanto, a tripla $(01-07-2015, temp_{elevada}, 15-07-2015)$ pode ser construída para representar o intervalo de interesse do atributo temperatura.

Definição 4.4 (Regra de associação temporal)**.** Uma *regra de associação temporal* descreve uma associação entre itens que contém uma informação temporal ligada a ela, seja essa informação um ponto temporal t ou um intervalo. Nos casos em que a regra de associação temporal é associada a um ponto temporal, a regra apresenta-se da seguinte forma $R_i[t] : X \Rightarrow Y \mid Sup, Conf, Lift, Conv$ em que t pode indicar o exato momento em que a regra foi obtida, é válida, deixa de ter validade, etc. De forma semelhante, quando

a regra de associação temporal está associada a um intervalo temporal, esta apresenta-se da seguinte forma: $R_i[-t,\ +t] : X \Rightarrow Y \mid Sup, Conf, Lift, Conv$.

Definição 4.5 (Tempo de vida de uma regra de associação temporal)**.** O *tempo de vida de uma regra de associação temporal* R_i, denominado *lifespan*(R_i) é o intervalo temporal no qual a regra de associação foi identificada. Ou seja, a duração de uma regra de associação temporal com garantia de valor de confiança. O *lifespan* de uma regra de associação temporal é construído pela união dos intervalos temporais nos quais a regra comparece. A união de dois intervalos temporais $[-t_1,\ +t_1]$ e $[-t_2,\ +t_2]$ é realizada pela construção de um novo intervalo temporal $[-t_3,\ +t_3]$, no qual $-t_3$ é o menor ponto temporal dos intervalos temporais e $+t_3$ é o maior deles.

Definição 4.6 (Frequência de ocorrência de uma regra de associação temporal)**.** A *frequência de ocorrência de uma regra de associação temporal* R_i, é um valor inteiro que descreve a periodicidade na qual a regra ocorre durante o tempo de vida de $R_i = [-t_R, +t_R]$, onde $-t_R$ e $+t_R$ são pontos temporais de início e fim, respectivamente, da janela w em relação à regra R.

Exemplo 4.2. Considere a base de dados (BDTempteste), apresentada na Tabela 4.1, na qual estão presentes quatro atributos quantitativos contínuos precedidos de um identificador da transação e a data em que os dados foram coletados. Assume-se, ainda, que os comportamentos de interesse de cada um dos atributos são identificados quando: o *atributo_A* assume valores negativos; o *atributo_B* assume valores entre $0,4$ e $0,6$; o *atributo_C* entre $1,0$ e $1,2$; e o *atributo_D* entre os valores $0,29$ e $0,31$. As ocorrências de interesse dos atributos são destacadas na Tabela 4.1.

Tabela 4.1: BDTempteste - Base de dados de transações composta por 4 atributos quantitativos contínuos após um temporal explícito.

		Itens da transação			
Transação	Data	atributo_A	atributo_B	atributo_C	atributo_D
T_1	2017-01-01	0,5	**0,6**	**1,1**	**0,295**
T_2	2017-03-02	**-0,5**	**0,5**	**1,0**	**0,30**
T_3	2017-05-04	0,3	0,9	-1,1	**0,30**
T_4	2017-06-01	1,8	**0,5**	-1,2	**0,295**
T_5	2017-06-03	0,6	**0,6**	**1,0**	0,287

Fonte: Elaborada pelo autor.

A regra de associação temporal $R_k[2017-01-01,\ 2017-06-03]$: atributo_B$[0,4..0,6]$, atributo_C$[1,0..1,2] \Rightarrow$ atributo_D$[0,29..0,31] | Conf : 0,667$ que pode ser gerada a partir dos padrões contidos na *BDTempteste* identifica que, durante o intervalo temporal compreendido entre [2017-01-01, 2017-06-03], é confiante, com 66,7% probabilidade de que o

atributo_R também assuma valores de seu comportamento de interesse ($[0,29..0,31]$) nas transações as quais sejam identificadas que ambos, o atributo_B e o atributo_C assumam valores de seus respectivos comportamentos de interesse. A regra R_k, neste exemplo onde a janela temporal é definida por $[2017-01-01,\ 2017-06-03]$, tem frequência igual a 2, ou seja, duas ocorrências a cada cinco meses.

Definição 4.7 (MWI - *Maximum Window for Interval*)**.** Dado um ponto temporal t, a *janela máxima para compor um intervalo temporal* $I = [-t, +t]$ é um valor inteiro que descreve a quantidade máxima de unidades de tempo que pode existir entre I e t para que t possa pertencer ao intervalo temporal. Quando $(-t\ -\ MWI) \leq (+t\ +\ MWI)$, t é incorporado a I por meio do processo de unificação de intervalos temporais, apresentado pela Definição 2.12

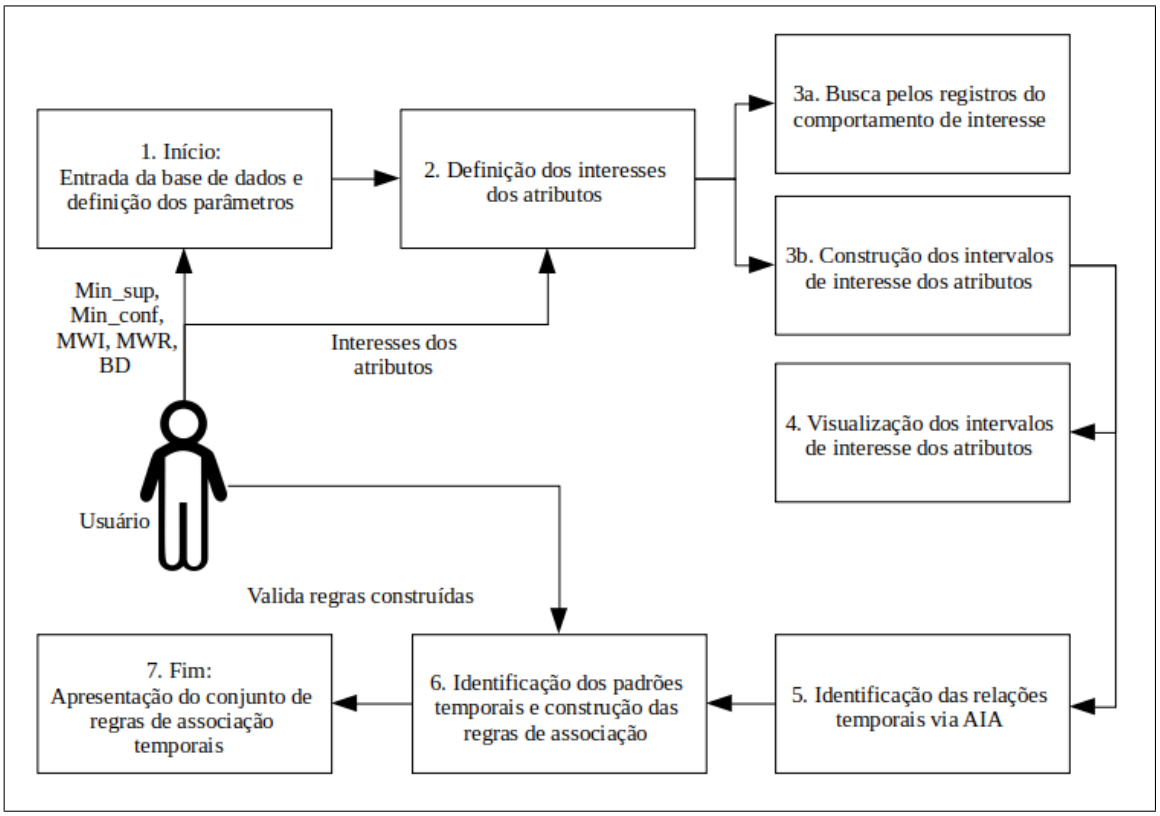
Definição 4.8 (MWR - *Maximum Window for Relations*)**.** A *janela máxima para relacionamentos intervalares* é um valor inteiro que descreve a quantidade máxima de unidades de tempo que podem existir entre dois intervalos $I_1 = [-t_1, +t_1]$ e $I_2 = [-t_2, +t_2]$ para que uma relação intervalar da AIA possa ser considerada entre ambos.

Exemplo 4.3. Seja o ponto temporal $t_k = 10-09-2019$, o qual representa um dia de ocorrência de chuva na cidade de *Exemplolândia*. Ainda, sejam os intervalos $I_1 = [-t_1, +t_1] = [12-09-2019, 17-09-2019]$ e $I_2 = [-t_2, +t_2] = [01-10-2019, 15-10-2019]$ períodos que descrevem dias com ocorrência de chuva e temperaturas baixas, respectivamente, na mesma cidade. Ao admitir $MWI = 3$ e $MWR = 15$ (unidades de tempo iguais a dias), nota-se que $t_k + MWI \geq -t_1$ e $t_k + MWI \leq +t_1$, o que, de acordo com a Definições 4.7 e 2.12, pode ser unificado em um único intervalo temporal. O novo intervalo, então, é definido por $I_1 = [-t_1, +t_1] = [10-09-2019, 17-09-2019]$. Nota-se, ainda, que $+t_1 + MWR \geq -t_2$, o que indica que pode existir um relacionamento temporal entre I_1 e I_2. Segundo a AIA, o predicado que pode ser utilizado para representar a relação temporal entre I_1 e I_2 é o $OVERLAPS$, portanto $OVERLAPS(I_1, I_2)$. O que, semanticamente, pode ser interpretado como: "um período de chuvas na cidade de *Exemplolândia* se sobrepõe a um período de frio na mesma cidade".

4.2 Detalhamento do método ART-Q

O método ART-Q tem vistas a identificar regras de associação em relações temporais descritas pela álgebra intervalar de Allen, a partir de bases de dados que comportam dados quantitativos contínuos. O esquema apresentado na Figura 4.1 ilustra em alto nível cada uma das etapas do método ART-Q com as interações que o método tem com o usuário.

Figura 4.1: Diagrama de fluxo de execução do método ART-Q: Association Rules involving Temporality and Quantitative continuous data.



Fonte: Elaborada pelo autor.

O método parte de (1) uma base de dados que contém atributos quantitativos contínuos e pelo menos uma informação temporal para cada registro (transação), (2) carrega a base de dados na memória RAM e recebe as definições dos comportamentos de interesse de cada um dos atributos que compõem a base de dados fornecida.

Em seguida (3a) analisa o comportamento dos atributos quantitativos contínuos e identifica as transações em que os atributos assumem valores do comportamento de interesse. Ao mesmo passo em que esta etapa ocorre, (3b) os pontos temporais associados às transações identificadas em (3a) são usados para construir intervalos temporais, nos quais o atributo assume valores de seu comportamento de interesse.

Com os intervalos de interesse definidos para os atributos, o método ART-Q fornece em (4) uma maneira de visualização para os intervalos, com a finalidade de facilitar a compreensão do comportamento de interesse dos atributos. Posteriormente, (5) a álgebra intervalar de Allen é utilizada para descrever as relações temporais que podem ser identificadas entre os intervalos construídos na etapa anterior e em seguida (6) o método

lança mão de uma estratégia similar à empregada pelo algoritmo Apriori para identificar padrões temporais e gerar regras de associação entre os padrões identificados.

Por fim, (7) um conjunto de regras de associação temporais é exportado para um arquivo de saída como resultado da execução do método ART-Q desenvolvido. É importante ressaltar, entretanto, que o ART-Q foi projetado para atuar em conformidade com o usuário/especialista do domínio. O usuário tem a tarefa indispensável de definir os parâmetros de entrada, tais como *Min_sup*, *Min_conf*, *MWI*, *MWR*, comportamentos de interesse dos atributos, a base de dados (BD) e, também, a tarefa de validar e filtrar as regras que são interessantes ao contexto.

4.2.1 Etapa 1 - Início: Entrada da base de dados e definição dos parâmetros

A primeira etapa de execução do ART-Q é a responsável pela verificação da existência de atributos temporais e quantitativos contínuos na base de dados (BD) fornecida pelo usuário. Para tal, é necessário que a base de dados seja composta por, pelo menos, um atributo temporal e um atributo quantitativo contínuo. O ART-Q não é projetado para lidar com atributos qualitativos (ordinais ou nominais), mas sim somente com atributos numéricos (quantitativos discretos e/ou contínuos). Os atributos que não são quantitativos contínuos, nem temporais, são ignorados.

Uma transação (ou registro) da BD esperada como entrada do ART-Q deve ser do tipo $t_i : Ca_{temp} \cup Ca_{cont}$, onde Ca_{temp} é um conjunto de atributos temporais e Ca_{cont} um conjunto de atributos que podem assumir valores quantitativos contínuos. Um exemplo possível seria um conjunto de informações de uma determinada área de plantio de uma cultura qualquer, $t_i : 2017-03-29, 11:48:00, 0, 33, 28, 5, 9, 15, 0, 43$, composto respectivamente por: data e hora, atributos temporais (a_{temp}) e valores de precipitação, temperatura, intensidade do vento e índice de vegetação, ambos atributos quantitativos contínuos (a_{cont}). Apesar do ART-Q ser projetado para lidar com valores quantitativos contínuos, o comparecimento de atributos discretos, ou qualitativos, não impede sua execução. Inicialmente no ART-Q atributos qualitativos são ignorados e não comparecerão nas regras geradas pelo método ART-Q.

Após a identificação dos atributos temporais e quantitativos contínuos da BD, a próxima etapa de execução do ART-Q é a definição dos comportamentos de interesse dos atributos quantitativos contínuos. Etapa esta, apresentada a seguir.

4.2.2 Etapa 2 - Definição dos interesse dos atributos

Atributos quantitativos contínuos podem assumir infinitos valores. Ora podem ser considerados importantes, ora não. Isso depende do tipo de análise que está sendo conduzida. Por exemplo, a síndrome do amarelecimento foliar (Amarelinho) pode ocasionar de 30% a 50% de perdas em plantios em diversas culturas, conforme descrevem Santiago e Rossetto (2009b). O amarelinho é causado por um vírus do grupo *Polerovirus*, transmitido por insetos e se manifesta em condições de estresse hídrico, por excesso ou falta. Ao buscar regras de associação temporais em um conjunto de dados que representam somente áreas contaminadas pelo amarelinho, é esperado que os atributos que expressam informações sobre umidade ou precipitação de chuva assumam valores elevados ou muito baixos.

Regras de associação podem ser geradas a partir de dados somente de áreas contaminadas pelo amarelinho. Tais regras são úteis, por exemplo, para identificar a doença. Para a construção de regras em dados de plantios não saudáveis, os valores dos atributos que descrevem a umidade, ou precipitação de chuva são valores do comportamento de interesse de uma área doente.

Entretanto ao considerar um conjunto de dados sem o prévio conhecimento da disseminação da doença, pode ocorrer que o comportamento normal para os atributos que descrevem a umidade e precipitação de chuva, seja o de assumir valores moderados e médios. Nesta busca, os valores interessantes para tais atributos são os que se apresentam fora do comportamento normal do atributo. Pois podem indicar o surgimento de uma doença, como o amarelinho.

Para a condução da próxima etapa de execução do ART-Q é necessário, entretanto, que sejam definidos os comportamentos de maior interesse para cada um dos atributos que compõem a base de dados fornecida como entrada ao método.

Como descrito na Seção 2.2, o método ART-Q assume o pressuposto que toda variável aleatória pode ser melhor representada por uma distribuição normal. Desta forma, o método permite que o usuário indique qual o comportamento de interesse de cada um dos atributos a partir de quatro possibilidades, a saber:

1. **acima do normal**: quando os valores mais interessantes são aqueles que superam a média somada a m (fornecido pelo usuário) desvios padrões ($> \mu + m\sigma$);
2. **normal**: quando os valores interessantes são aqueles que estão entre a média μ e m

desvios padrões ($[\mu - m\sigma, \mu + m\sigma]$);

3. **abaixo do normal**: quando os valores mais interessantes são aqueles que são inferiores a média decrescida de m (fornecido pelo usuário) desvios padrões ($< \mu - m\sigma$);

4. **fora do normal**: quando os valores mais interessantes são aqueles ou abaixo ou acima do normal ($< \mu - m\sigma$ ou $> \mu + m\sigma$).

Quando o comportamento de interesse do atributo é desconhecido pelo usuário, ou é optado por não informá-los, o ART-Q assume que o comportamento de interesse é aquele que se encontra no normal.

Esta etapa de execução do ART-Q permite, ainda, que o usuário defina qual a faixa de valores da distribuição que deve ser considerada pela variável m, descrita acima. Trata-se de um número inteiro que considera a quantidade de deslocamentos (quantos desvios padrões de distanciamento) devem ser considerados quando as opções selecionadas são *acima* ou *abaixo* do normal.

Por exemplo, em uma análise sobre doenças em um plantio de soja, quando é sabido que a alta umidade propicia o surgimento de fungos. Pode ser que assumir que os valores acima do normal são o comportamento de interesse para o atributo umidade do solo ainda implique na consideração de muitos registros nos quais o valor não destoa tanto da normalidade. Em casos como este, é interessante selecionar registros nos quais o valor do atributo se encontra muito acima da normalidade. Portanto m pode ser definido com um valor maior que 1.

A próxima etapa de execução do método ART-Q é a busca dos registros (transações) nos quais os atributos quantitativos contínuos podem assumir valores de seu comportamento de interesse, a partir das definições feitas pelo usuário nesta etapa de execução. Os detalhes da busca pelos registros do comportamento de interesse são descritos a seguir.

4.2.3 Etapa 3a - Busca pelos registros do comportamento de interesse dos atributos

Nesta etapa de execução, o ART-Q emprega uma estratégia que percorre toda a base de dados a fim de identificar quais são as transações da base de dados nas quais cada um dos atributos quantitativos assume valores dentro de seu comportamento de interesse. Essa estratégia é inspirada no trabalho de Aumann e Lindell (2003) que apresentam um método estatístico para a construção das regras de associação. Entretanto a definição dos comportamentos de interesse dos atributos é uma tarefa inovadora que não é considerada pelos trabalhos na literatura. Como descrito pela Definição 4.2, uma distribuição normal de probabilidade pode ser assumida para melhor representar o comportamento de cada um dos atributos quantitativos.

A estratégia do método ART-Q na condução desta etapa de execução é a de, a partir do primeiro registro na base de dados, verificar todos os atributos quantitativos quanto ao valor que assumem, se estão ou não compreendidos pelo intervalo de valores que descreve seu respectivo comportamento de interesse.

O Algoritmo 3, apresentado abaixo, descreve os passos seguidos pelo ART-Q para a realização desta etapa. Quando um valor quantitativo é identificado como pertencente ao comportamento de interesse de um atributo, o ponto temporal do registro a ele associado é armazenado em uma lista de pontos de interesse (POI) do atributo (ver linhas 6 e 7 do Algoritmo 3). Ao final desta etapa de execução, é gerado um conjunto de pontos de interesse para cada um dos atributos quantitativos que compõem a base de dados. Estes conjuntos são armazenados em arquivos de texto.

A busca de valores de interesse empregada pelo ART-Q, permite a evidenciação de mais nuances nos atributos que as estratégias presentes na literatura, como a implementada por Haldulakar e Agrawal (2011) que converte valores contínuos em binários, Yang e Feng (2010) que faz uso de algoritmos genéticos e discretiza os dados e Zheng et al. (2014) que constrói conjuntos nebulosos onde há incerteza. A estratégia empregada pelo método ART-Q também estende a seleção de características idealizada por Ribeiro, Traina e Traina (2008), pelo fato de permitir ao usuário a definição do comportamento de interesse do atributo, quando este é conhecido.

Ao mesmo passo em que ocorre a identificação das transações em que os atributos quantitativos contínuos assumem um comportamento de interesse, a próxima etapa de execução do método ART-Q também ocorre: a construção dos intervalos de interesse dos

Algoritmo 3 Procedimento para a identificação de pontos de interesse (POI) implementado pelo ART-Q.

```
1: procedure IDENTIFICAPOI(BD, interesses_set : def_1, ..., def_f, ..., def_n)
2:     POI ← ∅;
3:     for all registro r ∈ BD do
4:         for all atributo f ∈ r do
5:             if !temporal(f) then
6:                 if f ∈ def_f then
7:                     POI_f ← POI_f ∪ r;
8:                 end if
9:                 POI ← POI ∪ POI_f;
10:            end if
11:        end for
12:    end for
13:    return POI;
14: end procedure
```

atributos. Os detalhes dessa etapa são descritos a seguir.

4.2.4 Etapa 3b - Construção dos intervalos de interesse dos atributos

A assunção da temporalidade explícita dos dados na BD implica que, para cada uma das transações na BD, exista uma informação temporal a ela relacionada. Ou seja, ao identificar todas as ocorrências de interesse dos atributos quantitativos contínuos (Etapa 3a de execução do ART-Q) também se identifica cada um dos instantes (momentos) em que o atributo assume tais valores do comportamento de interesse.

Quando a temporalidade é considerada de forma implícita, pode ocorrer que existam dados faltantes entre duas transações da base de dados, ou pode haver desordem entre as transações. Ao considerar a temporalidade explícita nos dados esse problema pode ser evitado pois os valores temporais de duas transações subsequentes podem ser observados e validados quanto à real proximidade. Desta forma, uma possível situação de desordem na base de dados que contempla dados temporais de forma explícita também pode ser corrigida.

As informações tornam-se mais completas se relacionadas aos respectivos instantes de tempo em que os atributos assumem valores no comportamento de interesse. Ou seja, a informação temporal pode ser útil para expressar em quais instantes de tempo, ou por quanto tempo, um atributo assume valores interessantes, ou não. E é esta filosofia que o ART-Q segue nesta etapa de execução, o questionamento de *por quanto tempo um*

comportamento de interesse se mantém em um atributo?

À medida que registros da BD nos quais as ocorrências de comportamentos de interesse são identificadas, os instantes de tempos associados a eles são, também. Após esta identificação, intervalos temporais podem ser construídos pela união dos instantes de tempo (pontos temporais) que são próximos o suficiente para respeitar a janela máxima de consideração para construção de um intervalo temporal (MWI). Desta forma, os instantes em que os atributos tem comportamentos de interesse podem ser unidos para constituir intervalos de interesse do atributo.

O conceito de um *intervalo de interesse* de um atributo, é embasado nos intervalos de estados de Höppner (2001) que os define pela tripla (b, s, f) que descreve um estado s que ocorre em um intervalo definido pelos pontos temporais de início (b: *begin*) e fim (f: *finish*). Analogamente à definição proposta por Höppner (2001), um intervalo de interesse é identificado, neste trabalho, pela tripla $(-t_x,x,+t_x)$, onde $-t_x$ é um ponto temporal que identifica o início de uma ocorrência de um comportamento de interesse; x é a variável que representa a ocorrência do atributo quando assume valores do seu respectivo comportamento de interesse e $+t_x$ é o ponto temporal que identifica o fim do intervalo de interesse.

A estratégia de construir intervalos temporais para as ocorrências de comportamentos de interesse dos atributos é elaborada a partir do raciocínio de Chen et al. (2016) que considera o tempo de duração dos itens da base de dados para o processo de busca de regras de associação. Entretanto a proposta é mais antiga, o conceito de *lifespan* proposto por Ale e Rossi (2000) identifica o tempo de duração (em quantidade de transações da base de dados) de um padrão em uma base de dados.

À semelhança do processo de temporalização apresentado por Akhlagh, Tan e Khak (2012), o ART-Q constrói intervalos temporais, entretanto o faz de uma forma mais detalhista. Os intervalos temporais construídos pelo ART-Q levam em conta nuances das transações e não simplesmente unificam transações adjacentes na base de dados. A estratégia apresentada por Ale e Rossi (2000) é, de fato, muito interessante pela originalidade e eficiência. Por isso foi selecionada a ser a inspiradora do ART-Q nesta etapa.

Um exemplo que pode ser considerado para simplificar estes conceitos é a presença da ferrugem da cana-de-açúcar, uma doença de disseminação das mais rápidas, causada pelo fungo *Puccinia melanocephala*, cujos esporos são facilmente dispersos pelos ventos, como descrito por Santiago e Rossetto (2009a). Em uma análise, pode ser mais interessante identificar as transações em que o atributo relacionado à intensidade do vento assume

valores mais elevados, ou seja, fora de seu comportamento normal. Ao mesmo passo em que os registros que contemplam valores do comportamento de interesse do atributo são identificados, os instantes de tempo a eles relacionados também são. Desta forma, intervalos temporais podem ser construídos a fim de representar quando as ocorrências de comportamento de interesse (neste caso, intensidade do vento elevada) foram identificadas e por quanto tempo elas se sustentaram, i.e., seu intervalo de interesse.

Para a construção dos intervalos de interesse dos atributos, o método ART-Q implementa um procedimento que lança mão da teoria descrita pela Definição 2.12 que formaliza a construção de um intervalo temporal. O Algoritmo 4, apresentado abaixo, detalha como é realizado o procedimento de construção dos intervalos temporais de interesse dos atributos de BD.

Algoritmo 4 Procedimento para a construção de intervalos temporais de interesse, a partir dos pontos de interesse (POI) identificados pelo ARTQ.

```
1:  procedure CONSTROIINTERVALOS(POI, BD, MWI)
2:      conjIntervalos ← ∅;
3:      for all atributo f ∈ BD do
4:          for all pontotemporal t ∈ POI_f do
5:              if conjIntervalos_f ≠ ∅ then
6:                  for all intervalo int ∈ conjIntervalos_f do
7:                      t_i, t_f ← int;                          ▷ t_i,t_f = −t,+t
8:                      if t ∈ [t_i − MWI, t_f + MWI] then
9:                          atualiza(int, t);
10:                     else
11:                         novoIntervalo ← [t,t];
12:                         conjIntervalos_f ← conjIntervalos_f ⋃ novoIntervalo;
13:                     end if
14:                 end for
15:             else
16:                 novoIntervalo ← [t,t];
17:                 conjIntervalos_f ← conjIntervalos_f ⋃ novoIntervalo;
18:             end if
19:         end for
20:         conjIntervalos ← conjIntervalos ⋃ conjIntervalos_f;
21:     end for
22:     return conjIntervalos;
23: end procedure
```

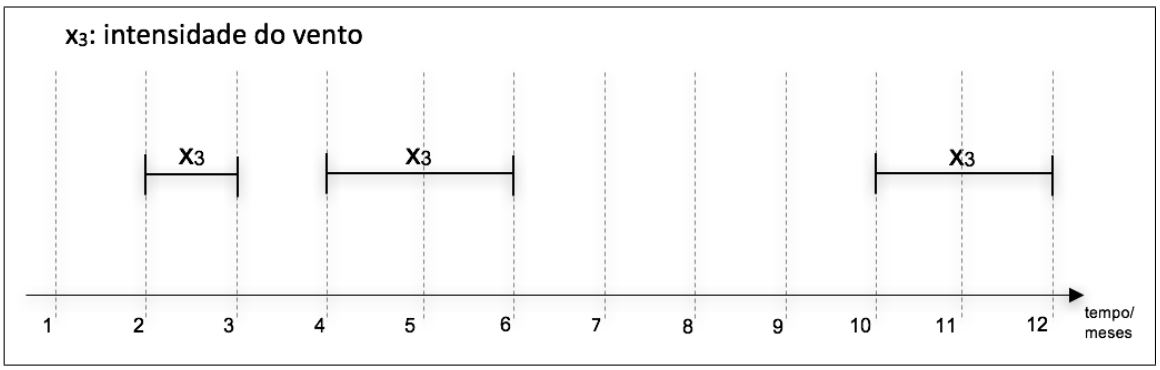
Para cada atributo f pertencente à BD (ver linha 3), o procedimento seleciona cada um dos pontos temporais t pertencentes ao conjunto POI de f. t é então verificado quanto à proximidade de cada um dos intervalos de interesse que f já possui (linhas 6 a 9). Quando t não se aproxima de nenhum intervalo de f, um novo intervalo é construído

com pontos temporais de início e fim descritos por t. Para a verificação da proximidade de t com os intervalos de f, o parâmetro MWI é levado em consideração. A condição da linha 5 do algoritmo verifica se existem intervalos já construídos para f. Quando nenhum intervalo foi construído ainda para f os comandos das linhas 16 e 17 indicam a construção de um intervalo definido por $[t, t]$.

4.2.5 Etapa 4 - Visualização dos intervalos de interesse dos atributos

Após a definição dos intervalos nos quais um atributo assume valores de interesse, o método ART-Q provê uma forma de representação gráfica que contribui com a interpretação e identificação de informações implícitas. Nesta etapa de execução o usuário pode visualizar os intervalos de interesse para todos os atributos quantitativos da BD. A Figura 4.2 ilustra um exemplo de visualização provida pelo ART-Q. Considere que x_3 indica um atributo quantitativo contínuo que descreve valores de intensidade do vento. Define-se que o comportamento de interesse de x_3 é atingido quando assume valores acima do normal. Na figura, o eixo horizontal (eixo x) representa o tempo, em meses. Observa-se na figura que houve ventos intensos nos períodos dos meses 2 a 3 (fevereiro a março), 4 a 6 (abril a junho) e 10 a 12 (outubro a dezembro), de maneira formal, (2,x_3,3), (4,x_3,6) e (10,x_3,12), respectivamente. Nos demais meses o atributo assume valores dentro do comportamento normal.

Figura 4.2: Intervalos de interesse do atributo x_3, que representa a intensidade do vento acima do comportamento normal (ventos fortes). Observam-se ocorrências de x_3 nos intervalos (2,x_3,3), (4,x_3,6) e (10,x_3,12).



Fonte: Elaborada pelo autor.

É importante ressaltar que somente um atributo foi considerado no exemplo da representação da Figura 4.2, entretanto para cada um dos atributos quantitativos, uma representação da mesma forma pode ser construída. A próxima etapa de execução do

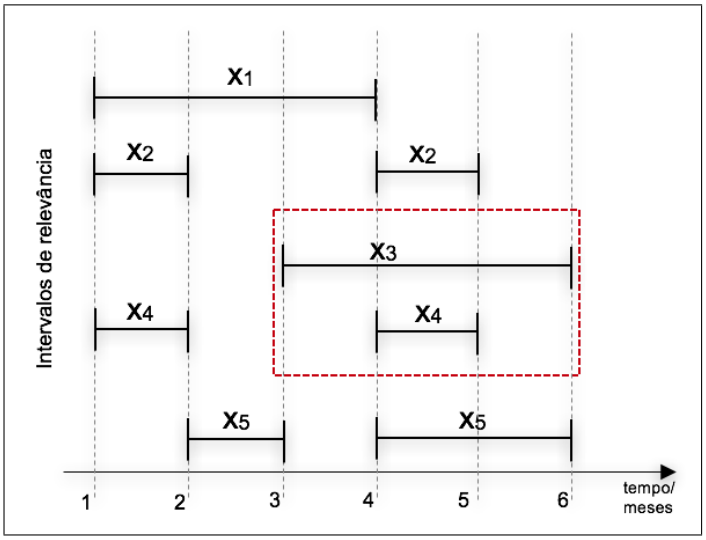
ART-Q é a responsável por identificar as relações temporais existentes entre os intervalos de interesse dos atributos, por meio da consideração da álgebra intervalar de Allen.

4.2.6 Etapa 5 - Identificação das relações temporais via AIA (Álgebra Intervalar de Allen)

Na representação da Figura 4.2 é possível observar que o atributo x_3 é recorrente, ou seja, se repete. Um intervalo de interesse não necessariamente é único, mas pode repetir ao longo do tempo em diferentes durações. Assim como ocorre com o atributo x_3, é possível que também ocorra com os demais atributos quantitativos contínuos de uma BD. Não somente envolvendo apenas um atributo. Cada relação intervalar possível descrita pela AIA é binária e, portanto, pode envolver intervalos de dois atributos distintos ou referentes a um mesmo atributo. Em outras palavras, a AIA é utilizada para facilitar a identificação de relacionamentos entre intervalos temporais de interesse dos atributos que não são de entendimentos triviais ao usuário.

Para cada um dos atributos temáticos quantitativos contínuos (a_{cont}) da BD, um conjunto de intervalos de interesse pode ser construído e um conjunto de relações de seus intervalos com outros intervalos de qualquer atributo da BD, também. A Figura 4.3 ilustra um exemplo com oito intervalos temporais de interesse que descrevem os comportamentos de interesse de três atributos, a saber x_1, x_2 e x_3.

Figura 4.3: Intervalos de interesse de três variáveis (x_1, x_2 e x_3). As variáveis representam comportamentos fora do padrão, assumidos em atributos quantitativos contínuos da BD. A seleção pontilhada representa uma relação $CONTAINS(x_3, x_4)$.



Fonte: Elaborada pelo autor.

Como pode ser observado na Figura 4.3 o quadro pontilhado identifica a ocorrência de uma relação do tipo $CONTAINS$ entre os atributos x_3 e x_4, i.e., existe uma ocorrência de um intervalo de interesse de x_4 durante um intervalo de interesse de x_3. Outras, tais como $EQUAL(x_2, x_4)$, $MEETS(x_4, x_5)$, $CONTAINS(x_3, x_2)$, entre outras, também podem ser observadas na figura.

Das possibilidades que a AIA provê para a representação das relações entre intervalos temporais, o método ART-Q considera o uso de sete. São elas: *BEFORE, CONTAINS, OVERLAPS, MEETS, STARTS, FINISHES* e *EQUAL*. Todas elas são consideradas a partir do parâmetro MWR que descreve a janela temporal máxima entre dois intervalos para que se possa ser considerada uma relação entre ambos. Em outras palavras, para se configurar a relação $BEFORE(A, B)$, por exemplo, entre os intervalos dos atributos A e B, deve-se ter um distanciamento temporal de no máximo MWR entre o fim de A e o início de B.

À medida em que as relações temporais são identificadas entre os intervalos de interesse, uma nova base de dados é construída, a $BDRelT$ (Base de Dados de Relações Temporais) que armazena todas as relações identificadas entre os intervalos de interesse. Cada transação da BDRelT é construída a partir de uma unidade de tempo (um ponto temporal) e é composta por todas as relações possíveis dos intervalos de interesse, que se iniciam neste exato ponto, com os demais intervalos de interesse que se distanciam no máximo a MWR unidades de tempo. O processo se inicia a partir do instante t_0 e após todas as relações dos intervalos que se iniciam em t_0 serem registradas, o valor de t é incrementado e o processo é repetido sucessivamente. Cada instante t_i de tempo analisado é denominado *passo*.

Uma transação da BDRelT que descreve as relações temporais do passo $-t$ é um conjunto de relações do tipo: *relação*(x_A, x_B). Por exemplo, a relação $OVERLAPS(x_A, x_B)$ que envolve os intervalos $[-t_A, +t_A, -t_B, +t_B] = [2, 8, 5, 10]$, indica uma relação de sobreposição do intervalo de interesse x_B, i.e., x_B se inicia enquanto x_A já ocorria, entretanto termina depois de x_A terminar.

A Tabela 4.2 exemplifica a BDRelT, construída a partir dos intervalos de interesse apresentados na Figura 4.3. Nela é possível notar que cada uma de suas linhas refere-se a um passo (unidade de tempo). Enquanto a primeira coluna lista os passos (unidades de tempo), a segunda apresenta as relações temporais dos intervalos de interesse de cada variável (x_i). As variáveis (x_i) descrevem, cada uma delas, um atributo quantitativo contínuo da BD, tal como: intensidade do vento assumindo valores acima do padrão,

precipitação de chuva elevada, ou escassa, umidade do ar normal, dentre outras.

Na primeira linha e coluna da tabela, observa-se a indicação do primeiro passo ($t = 1$) seguido de todas as relações possíveis entre os intervalos de interesse que se iniciam no passo $t = 1$, no caso x_1, x_2 e x_4. Ao analisar x_1 em $t = 1$ na Figura 4.3 vê-se que as relações possíveis, segundo a AIA (ver Tabela 2.2) são *MEETS(x_1,x_2), OVERLAPS(x_1,x_3), MEETS(x_1,x_4)* e *MEETS(x_1,x_5)*. De forma semelhante, *STARTS(x_2,x_1), BEFORE(x_2,x_3), EQUAL(x_2,x_4)* e *MEETS(x_2, x_5)* para x_2 em $t = 1$ e para x_4 em $t = 1$, *STARTS(x_4,x_1), BEFORE(x_4,x_3)* e *MEETS(x_4,x_5)*. Desta forma, a transação da BDRelT referente a $t = 1$ é composta pela união das relações dos intervalos de interesse. O passo t é, então, incrementado e o processo se repete até o último t possível.

Tabela 4.2: BDRelT, construída a partir dos intervalos de interesse apresentados na Figura 4.3.

passo (t)	relações temporais identificadas
$t = 1$	MEETS(x_1,x_2), OVERLAPS(x_1,x_3), MEETS(x_1,x_4), MEETS(x_1,x_5), STARTS(x_2,x_1), BEFORE(x_2,x_3), EQUAL(x_2,x_4), MEETS(x_2,x_5), STARTS(x_4,x_1), BEFORE(x_4,x_3), MEETS(x_4,x_5)
$t = 2$	CONTAINS(x_1,x_5), BEFORE(x_5,x_2), MEETS(x_5,x_3), BEFORE(x_5,x_4), BEFORE(x_5,x_5)
$t = 3$	-
$t = 4$	CONTAINS(x_3,x_2), EQUAL(x_2,x_4), STARTS(x_2,x_5), CONTAINS(x_3,x_4), STARTS(x_4,x_5), FINISHES(x_5,x_3)
$t = 5$	-
$t = 6$	-

Fonte: Elaborada pelo autor.

Nota-se que nos passos $t = 5$ e $t = 6$ nenhum intervalo de interesse se inicia, portanto nenhuma relação é possível de ser identificada. Já no passo $t = 3$ existe um intervalo de interesse do atributo x_3 que se inicia. As possíveis relações seriam *AFTER(x_3, x_2), AFTER(x_3, x_4), IS_OVERLAPED_BY(x_3, x_1), IS_FINISHED_BY(x_3, x_5)*, entre outras. Embora tais relações sejam, na verdade, inversas àquelas já registradas em outros passos. O método ART-Q não considera todas as 13 relações propostas pela AIA, mas somente as 7 básicas. Essa decisão é tomada pois caso as relações inversas também fossem consideradas, haveria muita redundância no conjunto final de relações entre os intervalos de interesse. Por exemplo, após a identificação de *AFTER(x_3, x_4)* no passo $t = 3$, caso a relação *BEFORE(x_4, x_3)* fosse identificada no passo $t = 4$, nenhuma informação nova seria provida, uma vez que ambas compartilham da mesma semântica.

A escolha pela adoção da álgebra intervalar de Allen (AIA) para representação das relações temporais deve-se ao fato que a AIA é suficientemente capaz de representar, de forma objetiva e clara, todas as relações possíveis entre intervalos temporais. Dessa forma, o método ART-Q é habilitado a prover bons resultados. A eficiência do uso da AIA no processo de mineração de regras de associação pode ser observada no trabalho que descreve o algoritmo ARMADA, de Winarko e Roddick (2007).

O procedimento descrito pelo Algoritmo 5 demonstra como o ART-Q identifica as relações entre os intervalos de interesse dos atributos de BD e constrói a $BDRelT$. Para cada atributo de BD (linha 3), identificado por f, o procedimento seleciona todos os intervalos de interesse que estão relacionados à f. Na linha 6 do algoritmo todos os intervalos existentes de todos os atributos são analisados quanto à proximidade do intervalo de f selecionado, sempre com respeito a MWR. Estes são chamados de candidatos a relacionamentos com o intervalo de f selecionado.

Cada um dos candidatos, então, é verificado quanto à relação que possuem com o intervalo corrente. Essa relação identificada é, então, inserida ao conjunto de relações de f (linha 8) e, posteriormente, passam a pertencer à $BDRelT$. Por fim, o procedimento tem como saída o conjunto de relacionamentos entre intervalos de interesse dos atributos, que descrevem a nova base de dados $BDRelT$.

Algoritmo 5 Procedimento para a identificação das relações temporais segundo a AIA, implementado pelo ARTQ.

```
1:  procedure IDENTIFICARELACOESAIA(MWR, conjIntervalos, BD)
2:      BDRelT ← ∅;
3:      for all atributo f ∈ BD do
4:          relacoesDef ← ∅;
5:          for all intervalo int ∈ conjIntervalos | int ∈ f do
6:              candRelacionamentos ← selIntervalos(int_ti, int_tf + MWR);
7:              for all intervalorelacionado intRel ∈ candRelacionamentos do
8:                  relacoesDef ← relacoesDef ⋃ nomeiaRelacao(int, intRel);
9:              end for
10:             BDRelT ← BDRelT ⋃ relacoesDef;
11:         end for
12:     end for
13:     return BDRelT;
14: end procedure
```

O próximo passo do ART-Q é o responsável por construir o conjunto das regras de associação, por meio da $BDRelT$.

4.2.7 Etapa 6 - Identificação dos padrões temporais e construção das regras de associação

Ainda que a construção de um conjunto de relações temporais entre os intervalos de interesse de atributos quantitativos contínuos da base de dados seja uma tarefa inovadora, pois não há relatos na literatura de trabalhos que a realizam, foi pretendido ir mais além. O método ART-Q gera um conjunto de regras de associação que consideram intervalos de interesse e seus relacionamentos, segundo a AIA. Em um exemplo simplista, uma regra de associação gerada pelo ART-Q pode ser interpretada da seguinte forma: *a ocorrência de períodos de chuva intensa enquanto os índices de vento são elevados, pode ser predecessora de um período de temperatura baixa, além do normal.* Esta informação apresentada pela regra pode prevenir um produtor de uma possível geada.

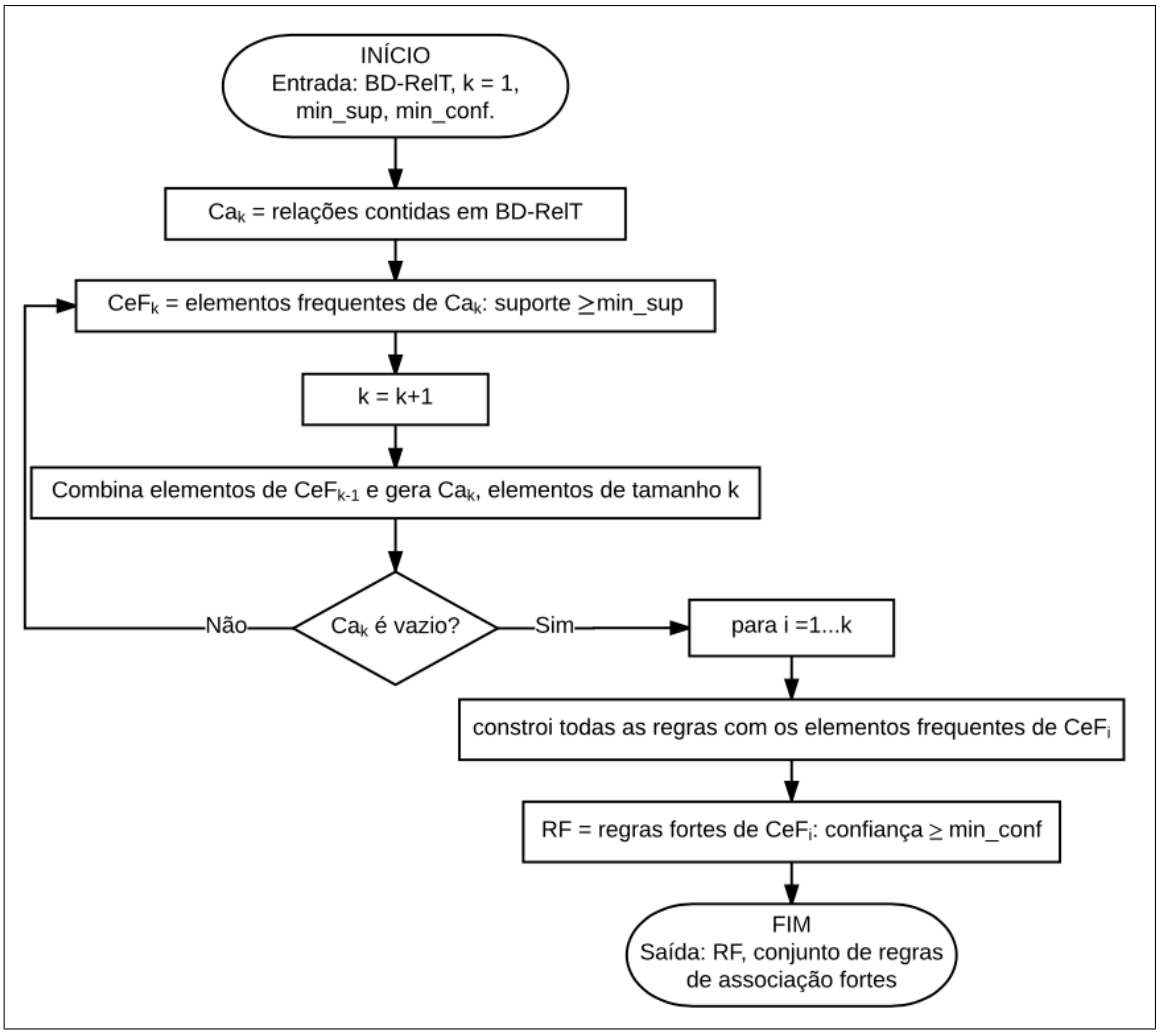
Para esta etapa, o ART-Q atua de forma semelhante àquela empregada pelo algoritmo Apriori, descrito por Agrawal e Srikant (1994) a fim de gerar regras de associação com as transações que compõem a BDRelT (4.2). Cada uma das relações entre dois intervalos de interesse, do tipo $MEETS(x_2,\ x_1)$, pode ser vista como um item.

A Figura 4.4 apresenta, em alto nível, um fluxograma que descreve o processo de geração das regras de associação empregado pelo ART-Q, a partir da BDRelT. É possível observar que, assim como o Apriori, o ART-Q inicialmente busca os elementos da base de dados que são ditos padrões frequentes e, posteriormente, os considera para construir as regras de associação. Para executar este procedimento, é necessário que, além da base de dados, os valores de Min_sup, Min_conf sejam informados como parâmetros de entrada.

Mais precisamente, o ART-Q realiza a tarefa de geração de regras de associação em 10 passos, descritos a seguir. Os passos são ilustrados na Figura 4.4:

1. Lista todas as relações em BDRelT, independente da transação a qual se encontram;
2. Calcula o valor de suporte de cada uma das relações, por meio da Equação 2.1. As que apresentam valor de suporte maior que Min_sup são chamadas frequentes (itens frequentes);
3. Armazena todos os itens frequentes no conjunto de elementos frequentes CeF_k, em que inicialmente $k = 1$;
4. Combina todos os elementos frequentes (em CeF_k) e incrementa o contador k para formar elementos de tamanho $k = k + 1$, i.e., estende os elementos frequentes com

Figura 4.4: Fluxograma que descreve o processo de geração das regras de associação empregado pelo ART-Q, a partir da BDRelT (que contém as relações temporais dos intervalos de interesses), à semelhança do que faz o algoritmo Apriori.



Fonte: Elaborada pelo autor.

mais uma relação, também frequente - as combinações são chamadas de candidatos a elementos frequentes e compõem o conjunto Ca_k;

5. Se o conjunto Ca_k é vazio, ou seja, nenhum elemento novo (de tamanho k) foi gerado a partir da combinação dos itens frequentes em CeF_{k-1} a busca de padrões interrompe e o processo continua a partir do passo (8);

6. Caso contrário, calcula o valor de suporte de todos os elementos de Ca_k, por meio da Equação 2.1. Os frequentes são armazenados em CeF_k;

7. O processo é repetido, a partir do passo (4) enquanto novos elementos puderem ser gerados a partir das combinações de elementos frequentes;

8. Para cada um dos elementos nos conjuntos de elementos frequentes CeF_i, i de 1 até k, constrói todas as regras de associação possíveis. Para cada CeF_i, um conjunto de regras de associação, compostas por i itens, é gerado;

9. Calcula o valor de confiança de todas as regras de tamanho i, por meio da Equação 2.2. As que possuem valor de confiança maior que Min_conf são chamadas fortes, as demais são descartadas;

10. O conjunto contendo todas as regras de associação geradas pelo método ART-Q é entregue como saída do procedimento.

Como saída do procedimento, um conjunto de regras de associação é entregue ao usuário. O conjunto é composto por regras de tamanhos variados, tanto quanto dos padrões que puderem ser identificados, uma vez que as regras de associação são geradas a partir dos padrões frequentes.

A fim de otimizar a quantidade e a força (valor de confiança) das regras de associação geradas, as transações que não possuem relações, tais como t_3, t_5 e t_6 da Tabela 4.2, são descartadas. De fato, ao considerá-las nenhum padrão seria incluído, além dos identificados nas demais transações. O valor de suporte dos padrões seria, também, reduzido e, consequentemente o valor de confiança das regras geradas, também. Parte desta etapa do ART-Q foi previamente implementada no algoritmo RAMiner, de João et al. (2017).

A escolha pelo algoritmo Apriori para servir como inspirador do processo de geração de regras de associação no ART-Q, é justificada pelo fato de ser um algoritmo simples e eficiente. A princípio, a abordagem iterativa do Apriori não limita a capacidade de construção de regras de associação do ART-Q. Portanto o algoritmo Apriori é, de fato, um grande colaborador para o método proposto.

4.2.8 Etapa 7 - Fim: apresentação do conjunto de regras de associação temporais

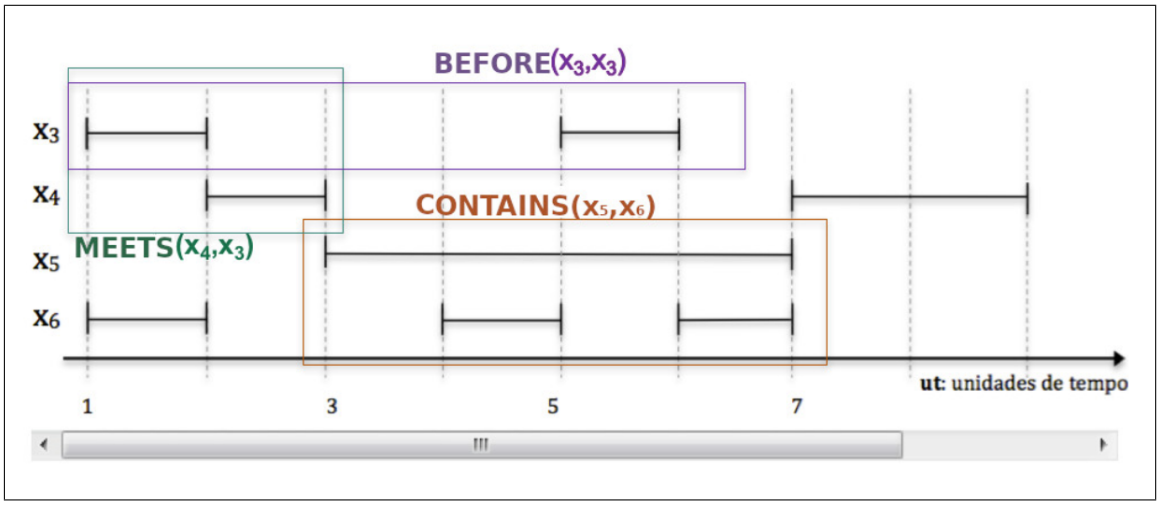
Por fim, a última etapa de execução do ART-Q é a responsável por gerar a saída com os resultados obtidos por ele. Todas as regras de associação geradas, bem como os intervalos temporais considerados e os padrões identificados são armazenados, cada um, em um arquivo. Antes da exportação, entretanto, entre as etapas 6 e 7 o ART-Q provê uma forma de filtrar os resultados gerados por meio da possibilidade de selecionar quais regras de associação são as mais contributivas para a análise conduzida. Desta forma,

o arquivo que contém as regras de associação geradas pelo método contempla apenas as regras selecionadas pelo usuário. Naturalmente uma execução do ART-Q em uma base de dados que configure muitas relações temporais pode tornar este trabalho custoso e cansativo. Por isso o ART-Q considera todas as regras geradas para compor o arquivo de saída caso o usuário não selecione nenhuma entre as opções.

Como comentado anteriormente, a regra gerada pelo método ART-Q considera um padrão, temporal, mais complexo que aqueles identificados pelos trabalhos na literatura. Consequentemente a regra de associação contempla uma grande quantidade de informação em sua semântica. Algumas regras de associação geradas pelo ART-Q podem ser de difícil compreensão, especialmente se o usuário não é um especialista do domínio da base de dados.

A Figura 4.5, apresentada abaixo, é incorporada neste trecho do documento para simplificar a compreensão da regra de associação resultante. Nela, um espaço temporal com 9 unidades de tempo é considerado e todos os intervalos de interesse identificados nesse espaço temporal são apresentados. As relações que compõem a regra $R_i[1,7]$: $BEFORE(x_3,x_3)$, $MEETS(x_4,x_3) \Rightarrow CONTAINS(x_5,x_6)$ | $Sup = 0,333$, $Conf = 0,83$, $Lift = 2,54$, $Conv = 5,67$ são destacadas e identificadas na Figura 4.5 a fim de simular como será realizada a visualização quando o usuário selecionar uma regra de associação.

Figura 4.5: Visualização dos resultados obtidos pelo ART-Q. As relações que compõem a regra $R_i[1,7] : BEFORE(x_3,x_3), MEETS(x_4,x_3) \Rightarrow CONTAINS(x_5,x_6)$ | $Sup = 0,333, Conf = 0,83, Lift = 2,54, Conv = 5,67$ são destacadas e identificadas.



Fonte: Elaborada pelo autor.

O intervalo de duração da regra de associação introduzido neste trabalho é inspirado pelo conceito de *lifespan* criado por Ale e Rossi (2000) e descrito pela Definição 4.5. Na regra $R_i[1,7]$ o intervalo temporal relacionado à regra informa que durante o intervalo

compreendido entre a unidade de tempo 1 e 7 a regra R_i é confiante.

Considerações finais

*Neste capítulo a descrição dos materiais e métodos empregados para a condução deste trabalho é apresentada, onde foram definidos conceitos importantes para o entendimento deste projeto. Também detalha as etapas de execução do método desenvolvido, denominado ART-Q (**A**ssociation **R**ules involving **T**emporality and **Q**uantitative continuous data). Mais precisamente, este capítulo teve o intuito de descrever as estratégias consideradas que permitem ao ART-Q identificar os intervalos nos quais os atributos da BD assumem valores mais interessantes à análise, bem como introduziu qual o formato inovador de padrão que o método trabalha, construído a partir da álgebra intervalar de Allen, que permite identificar todas as possíveis relações entre dois intervalos temporais. Ressaltou ainda a quantidade de informações que a regra de associação gerada pode revelar, uma vez que considera dados em sua forma bruta, sem omissão de detalhes e a temporalidade explícita nos dados que permite ao ART-Q ser mais robusto quanto a registros faltantes. O capítulo demonstrou, ainda, que o método é capaz de construir uma visualização para os intervalos de interesse dos atributos da BD.*

Capítulo 5

Experimentos e resultados

Neste capítulo, o método ART-Q é colocado à prova na condução de três experimentos que foram idealizados para comprovar o potencial que o método tem para contribuir com a área de mineração de dados. (1) O primeiro deles leva em consideração uma base de dados sintéticos, constituída por atributos quantitativos que podem ser representados por diferentes distribuições de probabilidade. Objetiva-se com esta configuração de base de dados, provar que o ART-Q lida com dados quantitativos, ainda que não sejam regidos pela distribuição normal de probabilidade. Uma maneira de validação dos resultados nesta base de dados foi idealizada para provar que o método tem reconhecida garantia e que pode atuar em bases de dados reais. A validação consiste em inserir informações manualmente à base de dados sintéticos, que devem ser identificadas pelo ART-Q. A partir da validação dos resultados, dois outros experimentos foram conduzidos e são apresentados neste capítulo: (2) quando é considerada uma base de dados de índices socioeconômicos; e (3) quando uma base de dados a respeito da ocorrência do fenômeno El Niño é utilizada. Ambas as bases de dados dos experimentos (2) e (3) contemplam dados reais. Nos três experimentos conduzidos, mais de uma forma de análise dos dados foi realizada. i.e., a flexibilidade do método para conduzir várias formas de extração de informação da base de dados foi explorada. Toda essa estratégia e a seleção das bases de dados consideradas nestes experimentos leva em consideração o questionamento se a manutenção da continuidade nos dados quantitativos, juntamente com a consideração da temporalidade explícita da BD pode contribuir com o processo de descoberta de informação. Mais detalhes de cada um dos experimentos realizados podem ser vistos no decorrer deste capítulo.

5.1 Experimento 1 - Base de dados sintéticos

A fim de comprovar a hipótese levantada (apresentada na Seção 1.3) e atingir os objetivos definidos, alguns experimentos foram conduzidos tanto em bases de dados sintéticos quanto em bases de dados reais.

A seguir, o primeiro dos experimentos realizados com o ART-Q é apresentado. Trata-se de um experimento conduzido sob uma base de dados sintéticos, idealizada para validar os resultados obtidos pelo método e, assim, provar a garantia que o método tem para lidar com bases de dados reais.

5.1.1 Descrição da base de dados *BDTest*

O primeiro dos experimentos foi realizado, considerando uma base de dados sintéticos, gerada com auxílio da ferramenta Orange Canvas, descrita por Demšar et al. (2013). A base de dados $BDTest$ compreende 1000 registros, sem importância de ordem cronológica, compostos, cada um deles, por 11 atributos. O primeiro dos atributos, como um requisito do ART-Q, descreve a informação temporal de forma explícita. Trata-se de uma data no formato $dd/mm/aaaa$. Para compor os 10 atributos ($Vari$, com i de 01 a 10) quantitativos restantes, foram usadas as seguintes distribuições:

- (a) $Var01, Var02$: Normal, o primeiro deles com variância de 10 e média igual a 0 e o segundo, com média 10 e variância igual a 1;
- (b) $Var03, Var04$: Bernoulli (um deles com probabilidade de sucesso em 85% e outro 50%);
- (c) $Var05, Var06$: Binomial (com 100 tentativas e probabilidade de sucesso em 85% e 50%, respectivamente);
- (d) $Var07, Var08$: Poisson (com números esperados de ocorrências num dado intervalo de tempo, definidos em 15 e 5, respectivamente) ;
- (e) $Var09, Var10$: Exponencial.

A escolha de usar várias distribuições de probabilidade ao construir a base de dados ($BDTest$) foi para aplicar o método em dados com vários comportamentos. As distribuições usadas construção da base de dados são as mais conhecidas e comuns. A Figura 5.1, abaixo, sumariza a $BDTest$.

Figura 5.1: Recorte da base de dados sintéticos *BDTest*, utilizada no experimento 1, para a validação dos resultados do ART-Q.

	A	B	C	D	E	F	G	H	I	J	K
1	Date	Var01	Var02	Var03	Var04	Var05	Var06	Var07	Var08	Var09	Var10
2	01/01/1980	10.6227	9.06929	1	1	88	53	9	5	1.74859	0.163121
3	02/01/1980	-7.00372	10.8796	1	0	85	49	17	5	3.32016	0.731049
4	03/01/1980	-3.51465	10284	1	1	89	51	17	6	0.372142	1.33845
5	04/01/1980	-8.18579	8.28783	1	0	84	44	17	7	0.0582338	1.12579
6	05/01/1980	7.4842	9.94861	1	1	82	51	11	6	1.87627	0.88134
7	06/01/1980	-8.85533	8.75116	1	0	82	51	15	4	0.226743	0.210843
8	07/01/1980	-13.3869	12.0707	1	0	79	55	22	6	1.50893	0.372132
9	08/01/1980	-17.6674	10.3821	1	1	85	63	14	6	1.79447	0.239429
999	...										
1000	25/09/1982	10.2796	10.5892	1	0	82	54	19	4	1.07498	0.644343

Fonte: Elaborada pelo autor.

Note que o atributo que contribui com a informação temporal, de forma explícita, apresenta-se mais à esquerda no conjunto e contempla valores de dias a partir de 01/01/1980.

5.1.2 Condução do experimento

Inicialmente, para a execução do ART-Q sob a $BDTest$, os parâmetros do método foram ajustados para os seguintes valores: ***Min_sup = 0,1***, ***Min_conf = 0,5***, ***Min_window_for_interval (MWI) = 5*** e ***Min_window_for_relation (MWR)= 30***. Em outras palavras, para ser identificado como frequente, um elemento tem que se repetir a uma taxa igual ou acima de 10% em $BDRelT$ (que será construída na etapa 5 de execução do ART-Q, descrita pela Seção 4.2.6); para ser considerada forte, uma regra de associação deve possuir confiança igual ou superior a 50%; para que dois pontos temporais possam ser considerados pertencentes a um mesmo intervalo, não podem se distanciar em mais de 5 unidades de tempo (dias neste caso) e; para que uma relação entre dois intervalos temporais $i_1 =< -t_1, i_1, +t_1 >$ e $i_2 =< -t_2, i_1, +t_2 >$ possa ser considerada existente, não deve ultrapassar 30 unidades de tempo (dias) entre si $(+t_1 + MWR \geq -t_2)$, considerando que $i_1 \leq i_2$, temporalmente.

A escolha de tais valores para os parâmetros dá-se simplesmente pelo fato que, após algumas execuções, a quantidade de padrões e regras de associação resultantes foi suficiente para identificar as informações inseridas e não tornar o trabalho de análise dos resultados cansativo.

Ao submeter a base de dados ao ART-Q, o mesmo identifica quais valores são numéricos (compatíveis com o método) e quais devem ser ignorados - neste exemplo nenhum valor não numérico (além da data) foi considerado. O atributo que informa o valor tem-

poral também é identificado, mas neste momento não é utilizado. Cada um dos atributos numéricos então é processado para que seja encontrado seu valor de média (μ) e desvio padrão (σ).

A Tabela 5.1, apresentada abaixo, relaciona cada um dos atributos (nomeados por $Vari \mid i \in [01, 10]$) seguidos pela informação que descreve a qual distribuição de probabilidade pertencem, a média dos valores que o atributo assume e o desvio padrão.

Tabela 5.1: Descrição dos atributos que compõem a base de dados *BDTest*, composta por dados sintéticos.

Atributos	Descrição dos valores		
	Distribuição estatística	Média dos valores (μ)	Desvio padrão (σ)
Var01	Normal	29,27	2721,503
Var02	Normal	593,908	247,157
Var03	Bernoulli	0,843	0,364
Var04	Bernoulli	0,507	0,5
Var05	Binomial	85,216	3,441
Var06	Binomial	49,781	5,08
Var07	Poisson	15,195	3,868
Var08	Poisson	4,988	2,202
Var09	Exponencial	7,67	135,419
Var10	Exponencial	5,887	92,429

Fonte: Elaborada pelo autor.

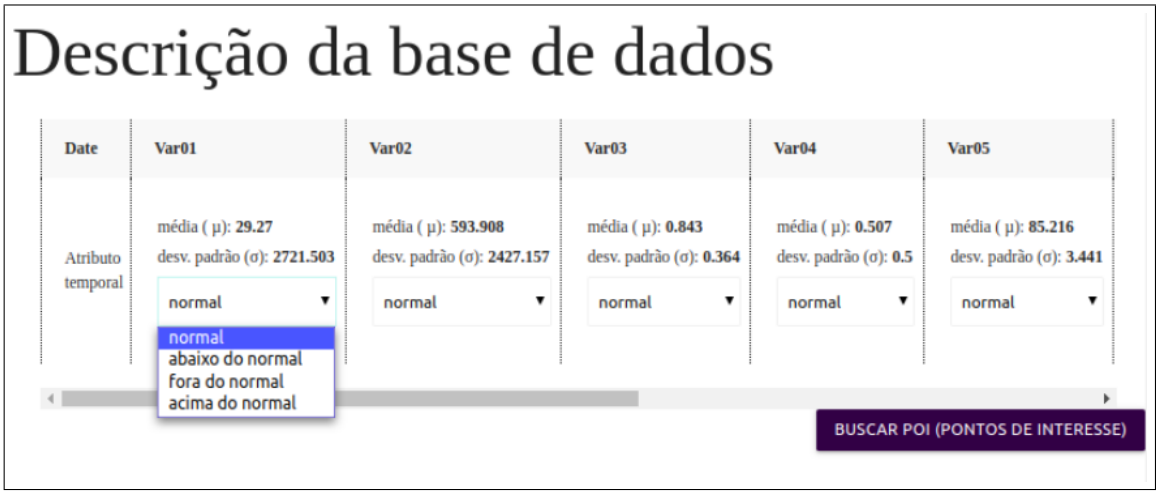
A próxima etapa do ART-Q depende da interação com o usuário que o opera. Este é responsável de informar ao método, para cada atributo, qual é o comportamento de maior interesse. Ou seja, o atributo em questão tem valores mais interessantes para a análise que está sendo conduzida, quando estão em uma determinada faixa de valores. Por exemplo, imagine que uma análise que busca informações implícitas em doenças em um plantio de soja e é sabido que a alta umidade propicia o surgimento de fungos. Ainda, sobre o exemplo, imagine que a base de dados fornecida contempla valores de precipitação de chuva sob a região do plantio. É intuitivo afirmar que os períodos (ou registros da base de dados) nos quais há maior concentração de chuva são mais propícios para o surgimento de fungos. Desta forma, pode-se afirmar que valores de precipitação de chuva mais elevados são mais interessantes para esta análise - comportamento de interesse do atributo precipitação de chuva.

Toma-se como base, a afirmação apresentada por Triola (1998) a respeito da distribuição das médias amostrais de uma variável aleatória x que tende para uma distribuição normal ao passo em que o tamanho da amostra cresce (ver Seção 2.2). Portanto, como descreve a Etapa 2 de execução do ART-Q (ver Seção 4.2.2), o método permite que o

usuário indique quatro possibilidades (comportamentos de interesse) para cada um dos atributos, a saber: **acima do normal**: quando os valores mais interessantes são os maiores que $\mu + m\sigma$; **normal**; **abaixo do normal**; e **fora do normal**. A Figura 5.2 ilustra as quatro possibilidades que o ART-Q provê para o usuário indicar qual é o comportamento de maior interesse para cada atributo quantitativo.

Para conduzir este experimento, foi definido que o comportamento de interesse é aquele onde os valores assumidos são *fora do normal* - para os atributos identificados por Var01, Var02, Var09 e Var10. Para os demais, o comportamento de interesse foi definido como *normal*.

Figura 5.2: Possibilidades que o ART-Q provê para o usuário indicar qual o comportamento de interesse dos atributos numéricos que compõem a base de dados no experimento 1.



Fonte: Elaborada pelo autor.

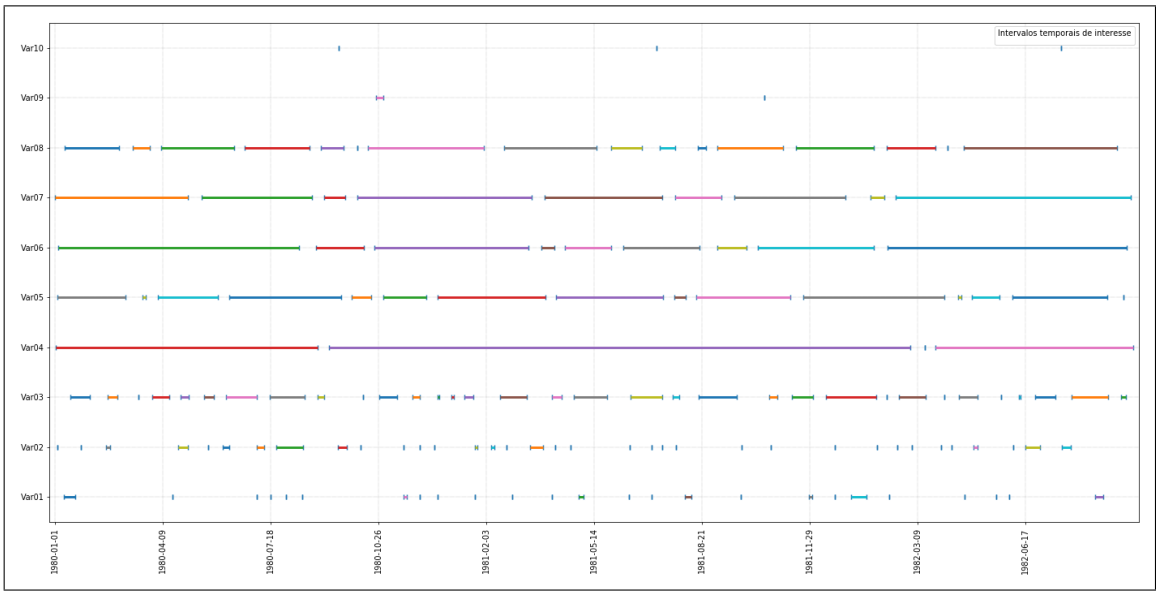
Por levar em consideração que a distribuição normal de probabilidade concentra cerca de 68% dos valores assumidos por uma variável aleatória entre o intervalo compreendido por $\mu \pm m * \sigma$, nos experimentos descritos neste trabalho, é adotado o valor $m = 1$.

Após a definição dos comportamentos de interesse de cada atributo, a tarefa a ser realizada pelo ART-Q é a detalhada na Seção 4.2.3, a busca pelo comportamento de interesse dos atributos. Nesta etapa de execução, o método varre a base de dados e analisa, para cada atributo, se o valor que assume no dado instante é de interesse ou não. Ao mesmo passo que identifica os pontos temporais nos quais o atributo assume valor em seu comportamento de interesse, o ART-Q também lança mão da tarefa de construção dos intervalos de interesse (ver Seção 4.2.4). Ao final da realização destas etapas, dois conjuntos de dados são produzidos: um com os pontos de interesse para cada atributo (POI); e, por meio da união dos pontos de interesse que satisfaz a condição imposta pela

janela (MWI), o outro conjunto com todos os intervalos de interesse para cada um dos atributos de $BDTest$.

Um intervalo de interesse descreve um período no qual um atributo assume valores de interesse que não superam a janela MWI de distanciamento entre si. Os intervalos de interesse para cada atributo de $BDTest$ identificados pelo ART-Q, neste experimento, são apresentados pela Figura 5.3. Note que a figura, por si só, já é uma forma de extrair informações da base de dados. Ela serve como um facilitador da compreensão do domínio, mais especificamente dos comportamentos de interesse definidos para os atributos. A figura permite a observação que os atributos $Var01$ e $Var02$ apresentam uma considerável quantidade de intervalos de interesse de curta duração, que podem revelar pontos de interesse que são espaçados o suficiente para não ocorrer dentro da janela MWI.

Figura 5.3: Intervalos de interesse dos atributos que compõem a *BDTest*, identificados pela etapa 3 de execução do ART-Q no experimento 1.



Fonte: Elaborada pelo autor.

A figura revela ainda que tais atributos possuem comportamentos mais atípicos que os de $Var04$, $Var05$, $Var06$, $Var07$ e $Var08$. Estes possuem intervalos de interesse (normal) mais longos; com destaque para o atributo $Var04$ que apresenta poucos períodos nos quais não se comporta dentro do seu comportamento de interesse.

Ainda que os atributos com seus respectivos comportamentos de interesse definidos como *normal* apresentem *buracos* que indicam comportamentos fora do normal, não é possível distinguir, nesta configuração, a qual situação esses *buracos* correspondem (se são acima, abaixo, ou fora do normal).

A Figura 5.3 permite observar, também, que os atributos identificados por *Var09 e Var10*, com seus comportamentos de interesse definidos como fora do normal, quase não possuem intervalos de interesse. Fato que pode ser interpretado como atributos que raramente oscilam para fora da normalidade. Quando intervalos de interesse para estes atributos são identificados, podem indicar ocorrências importantes.

Após a identificação dos intervalos de interesse, o ART-Q segue para a próxima etapa de execução, i.e., a identificação das relações temporais segundo a AIA (ver Seção 4.2.6). Etapa na qual o método seleciona todos os pontos temporais t_i nos quais intervalos de interesse se iniciam, independentemente do atributo o qual pertencem. A partir de cada um dos t_i selecionados, identifica todos intervalos temporais que se iniciam em t_i e buscam as possíveis relações temporais com outros intervalos.

Para configurar uma relação entre intervalos de interesse (por exemplo, entre os intervalos i_1 e i_2), i_2 não deve se distanciar além da janela (MWR) de i_1. De forma mais clara, o ponto temporal que determina o término de i_1 não deve se distanciar mais que MWR unidades de tempo do ponto temporal que determina o início de i_2. Neste experimento, o parâmetro MWR foi definido com o valor igual a 30 unidades de tempo (dias).

Uma observação que deve ser feita é que a unidade de tempo não depende do ART-Q, mas sim da base de dados fornecida como entrada. Por exemplo, se uma base de dados composta por registros de leituras mensais de um sensor qualquer é utilizada como entrada do ART-Q, a unidade de tempo considerada pelo método será quantificada em meses. Neste experimento, a base de dados contém valores diários, portanto a unidade de tempo é um dia.

Após a etapa de identificação das relações temporais, um conjunto composto por 461 relações entre os intervalos temporais de interesse de $BDTest$ foi gerado e utilizado pelo ART-Q para a construção de uma nova base de dados, a $BDRelT$. No total, a nova base de dados compreende 120 registros de tamanhos independentes quanto à quantidade de relações que compreendem. Isso porque cada registro r_i de $BDRelT$ é composto por todos os relacionamentos temporais dos intervalos int de interesse que se inicial no ponto temporal t_i com todos os intervalos que não se distanciam mais que MWR unidades de tempo de int. É possível, entretanto, deduzir que a quantidade de intervalos de interesse (independentemente do atributo associado) que o ART-Q pôde construir tem, no mínimo, o mesmo valor da quantidade de registros de $BDRelT$, uma vez que cada registro é construído a partir de, pelo menos, um intervalo de interesse que se inicia neste ponto temporal.

A partir da $BDRelT$ construída, o ART-Q lança mão de uma estratégia muito próxima àquela utilizada pelo algoritmo Apriori para identificar os padrões que pertencem à base de dados. Ao final desta etapa, o ART-Q pôde identificar a existência de 16 padrões temporais. A Tabela 5.2 apresenta cada um dos 16 padrões identificados nesta etapa. Cada padrão é identificado, na tabela, por um número (coluna mais à esquerda) seguido pela apresentação do padrão e, por fim, seu respectivo valor de suporte. Entre os padrões identificados pelo ART-Q, apenas um é constituído por mais de duas relações entre intervalos de interesse, o padrão identificado por 16 na tabela. A semântica do padrão pode ser descrita como:

> *10% dos relacionamentos entre intervalos de interesse dos atributos de BDTest descrevem a informação que Var02 assume valores de interesse (fora do normal) por um período que antecede (não imediatamente) um período de comportamento também fora do normal de Var01. O comportamento fora do normal de Var02 se repete periodicamente e, também, antecede (não imediatamente) um período de normalidade de Var03.*

Tabela 5.2: Padrões identificados pelo ART-Q na base de dados de relações entre intervalos temporais de interesse (*BDRelT*), no experimento 1.

#	Padrão temporal identificado	Frequência (suporte)
1	BEFORE(Var01;Var02)	0,15
2	BEFORE(Var01;Var03)	0,15
3	BEFORE(Var02;Var01)	0,25
4	BEFORE(Var02;Var02)	0,1916
5	BEFORE(Var02;Var03)	0,25
6	BEFORE(Var02;Var05)	0,1083
7	BEFORE(Var02;Var08)	0,116
8	BEFORE(Var03;Var02)	0,125
9	BEFORE(Var03;Var03)	0,133
10	BEFORE(Var01;Var02), BEFORE(Var01;Var03)	0,1083
11	BEFORE(Var02;Var01), BEFORE(Var02;Var02)	0,1083
12	BEFORE(Var02;Var01), BEFORE(Var02;Var03)	0,133
13	BEFORE(Var02;Var02), BEFORE(Var02;Var03)	0,1083
14	BEFORE(Var02;Var03), BEFORE(Var02;Var05)	0,1083
15	BEFORE(Var02;Var03), BEFORE(Var02;Var08)	0,1083
16	BEFORE(Var02;Var01), BEFORE(Var02;Var02), BEFORE(Var02;Var03)	0,1

Fonte: Elaborada pelo autor.

É importante, pois, lembrar que inicialmente o parâmetro $Min\_sup = 0,1$ limitou a identificação de padrões àqueles que ocorrem em, pelo menos, 10% de $BDRelT$. Neste

caso, para ser considerado um padrão, uma relação temporal entre intervalos de interesse deve ocorrer em, no mínimo 12 registros da base de dados, uma vez que $BDRelT$ é composta por 120 registros.

Por fim, após a identificação dos padrões temporais, o ART-Q realiza a etapa que constrói as regras de associação a partir dos padrões temporais identificados que contemplam mais de 1 item, i.e., que possuam pelo menos 2 relações temporais entre intervalos de interesses. Relembre que o valor de Min_conf foi inicialmente definido em $0,5$. Ao final desta etapa de execução, o ART-Q foi capaz de gerar um conjunto com 6 regras de associação consideradas fortes. As regras são apresentadas abaixo, na Tabela 5.3 com seus respectivos valores de suporte, confiança, elevação (lift) e convicção.

Tabela 5.3: Regras de associação temporais geradas pelo ART-Q a partir dos padrões temporais identificados em ($BDRelT$), no experimento 1.

#	Regra	Suporte	Confiança	Lift	Convicção
1	BEFORE(Var01;Var02) → BEFORE(Var01;Var03)	0,108	0,722	4,815	3,058
2	BEFORE(Var02;Var01) → BEFORE(Var02;Var02)	0,108	0,722	3,768	2,908
3	BEFORE(Var02;Var01) → BEFORE(Var02;Var03)	0,133	0,889	3,556	6,757
4	BEFORE(Var02;Var02) → BEFORE(Var02;Var03)	0,183	0,957	3,826	17,442
5	BEFORE(Var02;Var01) → BEFORE(Var02;Var02), BEFORE(Var02;Var03)	0,1	0,667	3,636	2,452
6	BEFORE(Var02;Var01), BEFORE(Var02;Var02) → BEFORE(Var02;Var03)	0,1	0,923	3,692	9,74

Fonte: Elaborada pelo autor.

Dentre as regras de associação construídas e mensuradas pelo ART-Q neste experimento, destacam-se as regras identificadas pelos números 5 e 6 na tabela. Ambas são compostas por três padrões temporais. Note que na regra (5) o antecedente é composto por apenas um padrão, o que indica que o padrão $BEFORE(Var02, Var01)$ tem força suficiente para implicar na ocorrência de outros dois padrões. O valor positivo de *Lift* nesta regra indica a chamada dependência positiva. Em outras palavras, isso indica que o padrão que compõe o antecedente da regra eleva a probabilidade da ocorrência do consequente.

Já na regra (6) é possível notar que a incorporação do padrão temporal $BEFORE$ ($Var02$,$Var02$) ao antecedente (originalmente composto por $BEFORE(Var02, Var01)$),

além de elevar a confiança da regra, aumenta consideravelmente o valor de convicção (*Conv*). O que indica que há uma forte dependência do consequente pelo antecedente. Fato que pode ser novamente observado na regra (4) que compartilha dos mesmos padrões, exceto pelo padrão $BEFORE(Var02, Var01)$. O valor de convicção (17,442) indica, também, a força da dependência declarada na implicação.

5.1.3 Validação dos resultados

Além de servir como um teste inicial para a verificação da corretude do ART-Q, a condução deste experimento possibilitou a validação dos resultados obtidos pelo método desenvolvido. Para tal, logo após a construção de $BDTest$, algumas informações foram manualmente inseridas na base de dados. A estratégia por trás desta ação consiste em verificar se o método ART-Q é capaz de revelar informações implícitas na base de dados.

Como inicialmente era planejado considerar que o comportamento de interesse das variáveis (atributos) $Var01$ e $Var02$ fosse ao assumir valores acima das suas respectivas médias somadas a um desvio padrão, a base $BDTest$ foi manipulada propositalmente. Vários registros, nos quais o valor assumido por $Var02$ satisfaz a condição para ser um valor de interesse (acima do desvio padrão), foram identificados e alguns registros cronologicamente após cada um destes identificados foram também alterados. Entretanto essa alteração deu-se nos valores do atributo $Var01$. Essa manobra visou incluir uma informação à base de dados de que pontos de interesse de $Var02$ antecedessem pontos de interesse de $Var01$, com frequência. As informações inseridas respeitaram o distanciamento máximo permitido por MWR, para que uma relação pudesse ser identificada futuramente e, ao mesmo tempo, não foram imediatas, mas sim com o distanciamento de 1 unidade de tempo.

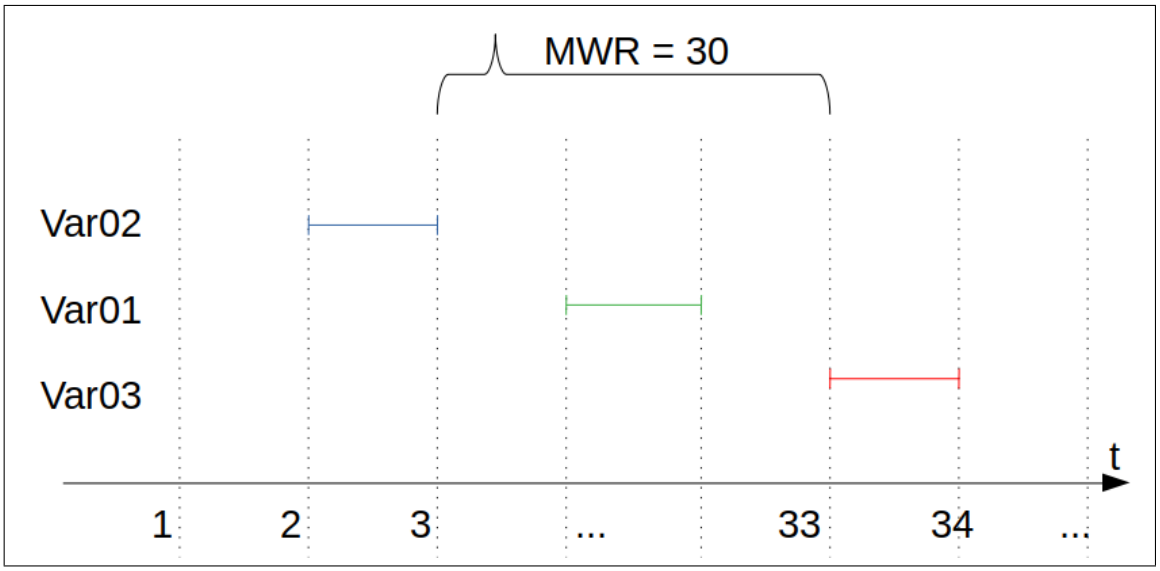
Com o resultado obtido após a execução do método ART-Q foi, realmente, possível identificar $BEFORE(Var02; Var01)$ como um padrão da base de dados. O que indica que o método ART-Q foi capaz de encontrar a relação manualmente inserida na BD. Ainda, além da identificação do padrão contemplando a informação manualmente inserida, o método ART-Q foi refletida no conjunto de regras de associação construídas $R_2, R_3, R_5 e R_6$. Todas essas regras foram geradas considerando a relação inserida como termo a ser encontrado. A regra $R_3 : BEFORE(Var02; Var01) \rightarrow BEFORE(Var02; Var03)$, por exemplo, pode ser selecionada entre todas construídas pelo ART-Q para ser interpretada. A semântica da regra pode ser entendida como:

É esperado, com confiança de 88,9% e convicção superior a 6,7 vezes, que a ocorrência de um período de interesse de $Var01$ que sucede (não imediatamente) um período de interesse de $Var02$ implique, também, na ocorrência de um período fora da normalidade de $Var02$ seguido de um período de normalidade $Var03$.

O *lifespan* de R_3 é definido pelo intervalo $i = [-t, +t] = [01-01-1980, 25-09-1982]$, intervalo no qual a regra R_3 foi identificada. Em outras palavras, ao notar que $Var02$ assume períodos de interesse em sequência de períodos de interesse de $Var01$, é interessante observar $Var03$ pois existe uma grande probabilidade de se comportar de forma a despertar interesse por um período de tempo, uma vez que a ocorrência já identificada eleva a chance da previsão ocorrer.

Uma possível visualização da regra R_3 pode ser vista na Figura 5.4, a fim de facilitar a sua compreensão. Nela é possível observar com clareza a janela MWR delimitando o espaço de busca por intervalos que podem se relacionar.

Figura 5.4: Visualização da regra de associação R_3, com espaço de busca por relações temporais definido por MWR, no experimento 1.



Fonte: Elaborada pelo autor.

O parâmetro $MWR = 30$ determina que a partir do ponto temporal que identifica o final do intervalo de interesse de Var02, representado simbolicamente na figura pelo número 3, somente os intervalos que se iniciem até o ponto temporal 33 possam ser considerados para a identificação de uma relação temporal.

5.2 Experimento 2 - Dados socioeconômicos

Em seguida à validação do método ART-Q, realizada por meio do experimento 1 (ver Seção 5.1) um novo experimento foi conduzido. Este, entretanto considera uma base de dados real. O objetivo deste experimento é contribuir com a premissa que o método ART-Q pode ser usado em bases de dados de diversos contextos, não se especifica a um determinado assunto. Desta forma, o ART-Q pode contribuir efetivamente com diversas análises ao proporcionar uma visão temporal que as estratégias anteriores de mineração de regras de associação não contemplam. A seguir, o experimento 2 é detalhado, a começar pela descrição da base de dados utilizada, seguida pela descrição da condução do experimento e, por fim, uma discussão sobre os resultados obtidos é apresentada.

5.2.1 Bases de dados de índices socioeconômicos (*BDBRA* e *BDSocio*)

O Banco Mundial[1] é uma das maiores fontes de financiamento e conhecimento para países em desenvolvimento; uma parceria de instituições globais que trabalham com o apoio a soluções sustentáveis para reduzir a quantidade da população mundial que vive em extrema pobreza para 3% até o ano de 2030 - proporção que chegou a 10% em 2015. A organização assume que uma pessoa é condicionada a viver em extrema pobreza se vive com menos de U$1,90 por dia. Mais de 1,9 bilhão de pessoas (26,2%) da população mundial vive com menos de U$3,2 por dia; 46% com menos de U$5,5. No total, o Banco Mundial reúne 189 países envolvidos em mais de 12 mil projetos, espalhados em mais de 130 localidades, com financiamento superior a 45 bilhões de dólares.

Uma das estratégias adotadas pelo Banco Mundial é manter a transparência e disponibilizar conhecimento e dados sobre seus projetos. Em sua plataforma na Internet, existe um grande conjunto de dados socioeconômicos, ricos em detalhes, que são disponibilizados com a finalidade de fomentar a pesquisa sobre o assunto. A título de conhecimento, no total, o banco de dados do The World Bank (2020) disponibiliza um conjunto de 1431 índices socioeconômicos referentes a 264 países.

Para a condução deste experimento, duas bases de dados foram obtidas da plataforma do banco mundial. A primeira delas, intitulada *BDBra*, é composta por 10 atributos que descrevem um conjunto de índices socioeconômicos durante um período de 20 anos (20 registros). Todos os atributos da base de dados relatam indicadores socioeconômicos do

[1] World Bank Group: https://www.worldbank.org/

Brasil durante o período de 20 anos, a contar do ano de 2000. A segunda, intitulada *BDSocio* considera os mesmos indicadores, entretanto envolve mais países. Mais detalhes sobre a constituição da *BDSocio* podem ser vistos na Seção 5.2.3. A Figura 5.5 apresenta a primeira base de dados *BDBra*.

Figura 5.5: BDBra: base de dados que contempla índices socioeconômicos do Brasil, no período de 2000 a 2019, considerada no experimento 2.

Data	cres_pop	per_capita	exp_vida	morte_5	escol_sec	Imuniza12-23	infla	investExterno	export	import
31/12/2000	1.4	3930.0	70.1	34.8	99.0	?	5.6	10.2	12.5	3.29
31/12/2001	1.4	3350.0	70.5	32.6	99.0	?	8.2	12.4	14.6	2.32
31/12/2002	1.3	3090.0	70.8	30.5	96.0	110.0	9.8	14.2	13.4	1.65
31/12/2003	1.3	2980.0	71.2	28.5	97.0	101.9	14.1	15.2	13.0	1.01
31/12/2004	1.2	3340.0	71.5	26.6	97.0	102.4	7.8	16.5	13.1	1.81
31/12/2005	1.1	4000.0	71.9	24.9	98.0	101.3	7.4	15.2	11.8	1.54
31/12/2006	1.1	4850.0	72.3	23.3	99.0	?	6.8	14.4	11.7	1.93
31/12/2007	1.0	6190.0	72.6	21.9	99.0	95.2	6.4	13.3	12.0	4.45
31/12/2008	1.0	7620.0	73.0	20.7	99.0	96.7	8.8	13.5	13.7	5.07
31/12/2009	1.0	8320.0	73.3	19.6	99.0	96.7	7.3	10.9	11.3	3.14
31/12/2010	0.9	9660.0	73.6	18.7	99.0	91.5	8.4	10.9	11.9	8.23
31/12/2011	0.9	11080.0	73.9	17.9	99.0	95.3	8.3	11.6	12.4	1.02
31/12/2012	0.9	12360.0	74.2	17.3	99.0	93.0	7.9	11.9	13.2	9.25
31/12/2013	0.9	12810.0	74.5	16.7	98.0	102.1	7.5	11.7	14.0	7.52
31/12/2014	0.9	12100.0	74.7	16.2	97.0	100.8	7.8	11.0	13.7	8.77
31/12/2015	0.8	10160.0	75.0	15.7	96.0	100.1	7.6	12.9	14.1	6.47
31/12/2016	0.8	8930.0	75.2	16.3	95.0	101.5	8.1	12.5	12.1	7.42
31/12/2017	0.8	8670.0	75.5	14.8	91.0	100.8	3.5	12.6	11.6	6.88
31/12/2018	0.8	9140.0	?	14.4	84.0	?	3.0	14.8	14.3	7.88
31/12/2019	1.0	8320.0	73.0	19.6	98.0	101.3	7.8	12.6	13.0	5.07

Fonte: Elaborada pelo autor.

Entretanto muitos destes índices são tão específicos a algumas localidades que têm valores muito esparsos, o que resulta em muitos dados faltantes. Foi feita uma escolha dos índices mais interessantes para compor as bases de dados *BDBra* e *BDSocio*, levando em consideração o fato que este conjunto é uma das combinações de índices (atributos) mais completa entre os países mais populares. Em outras palavras, foi feita a seleção dos índices que contém menos dados faltantes entre os países que compõem a base de dados. Os dez índices selecionados para compor *BDBra* são:

1. cres_pop: representa o crescimento populacional anual (em %);
2. per_capita: renda per capita anual (em U$);
3. exp_vida: expectativa de vida (em anos) para aquele período;
4. morte_5: mortalidade infantil (morte de crianças de até 5 anos), razão por cada 1000 nascimentos;
5. escol_sec: crescimento em % de estudantes no secundário;

6. imuniza12-23: % de crianças imunizadas entre os 13 e 23 primeiros meses de vida;
7. infla: inflação do país (em %) no ano;
8. investExterno: investimento externo em U$;
9. export: % do PIB (Produto Interno Bruto) de bens e serviços exportados;
10. import: % do PIB (Produto Interno Bruto) de bens e serviços importados.

Ainda que esse conjunto de dados seja o mais completo, note que a figura revela a ocorrência de alguns dados faltantes (identificados por um `"?"`).

Fato este que não implica em prejuízos aos resultados obtidos por meio do uso do ART-Q. O método aqui descrito lança mão de uma estratégia que ignora os valores (ou transações) faltantes na base de dados. O registro com o valor faltante simplesmente não comparece como um ponto de interesse do atributo.

5.2.2 Condução do experimento

O objetivo principal da condução deste experimento é o de provar que o ART-Q é capaz de lidar com valores de diversos cenários. Note que a base de dados descrita na seção anterior é pequena, com valores não normalizados e com dados faltantes, o que poderia representar o caos em algumas estratégias de mineração de dados.

Os parâmetros do método ART-Q, para lidar com a *BDBra* foram definidos como segue: *Min_sup*: 0,22; *Min_conf*: 0,5; *MWI (janela de intervalos)*: 370 (dias); e *MWR (janela de relações)*: 750 (dias). Com essa configuração, um intervalo pode ser constituído para quaisquer pontos temporais consecutivos, visto que os valores são anuais. Uma relação pode ser considerada entre dois intervalos de interesse distantes, uma da outra, por um período de até 2 anos. A admissão como um padrão para uma relação que compareça em, pelo menos, 22% da base de dados de relações temporais ($BDRelT$).

Trata-se de uma busca com espaçamento temporal considerável. Entretanto como a finalidade do uso do método ART-Q é encontrar informações implícitas, neste experimento, as janelas destes tamanhos fazem-se necessárias.

Para a definição dos comportamentos de interesse dos atributos, a premissa que valores *acima do normal* são os momentos a serem identificados ($\mu + m * \sigma | m = 1$) foi levada em consideração, para quase todos os índices que compõem a base de dados. Excluem-se

dessa premissa, os índices referentes à inflação anual (*infla*) e taxa de mortalidade infantil (*morte_5*). Para estes, os valores de maior interesse são aqueles que se encontram *abaixo do normal* ($\mu - m * \sigma$), ou seja, quando apresentam quedas.

A Figura 5.6 apresenta uma ilustração dos intervalos de interesse, para cada um dos atributos de $BDBra$, identificados pelo ART-Q configurado com os valores de parâmetros descritos acima. A visualização permite múltiplas interpretações de forma simples e intuitiva. Pode-se, por exemplo, ser fonte da afirmação que o Brasil teve seus melhores períodos de investimento externo (atributo $investExterno$) durante o período compreendido entre os anos de 2003 até o final do ano de 2005. A melhor renda per capita do brasileiro (relativa) foi durante os anos de 2011 até o fim do ano de 2015. O intervalo temporal de destaque no crescimento populacional do Brasil ocorreu no período compreendido pelos anos 2000, 2001, 2002 e 2003. Durante o período de 2010 a 2012, o Brasil importou mais que o normal.

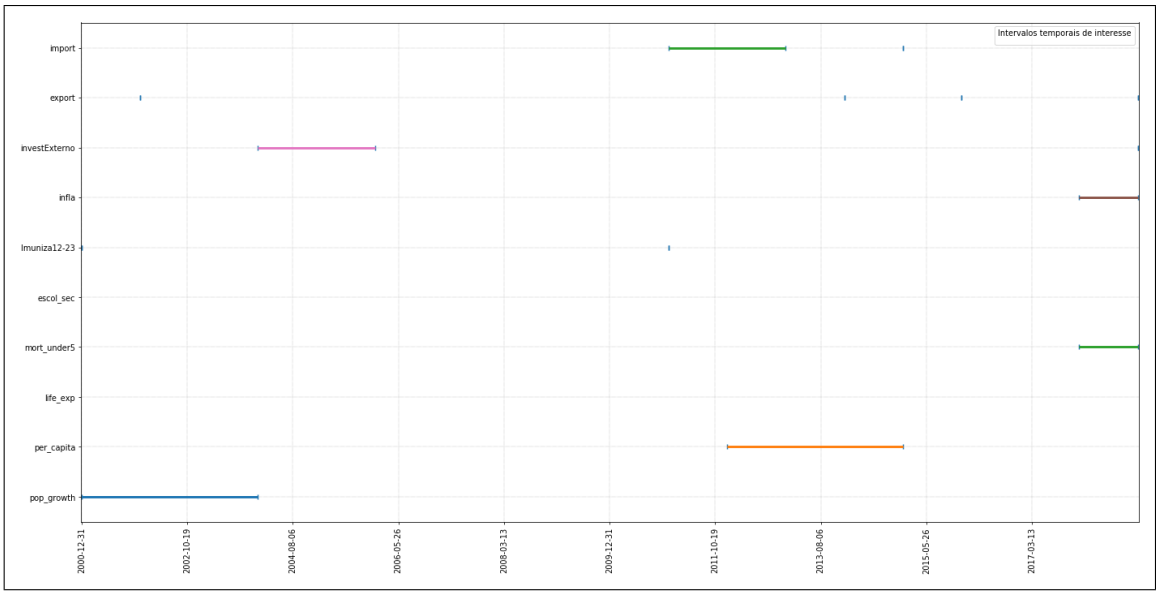
Como comentado por Martello (2013), em entrevista ao portal de notícias G1 o Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC) declarou que a balança comercial brasileira (exportações menos importações) registrou uma queda de 34,75% no ano de 2012. O saldo positivo do ano anterior também representou o menor superavit da balança comercial brasileira desde 2002. Esse fato, no entanto, não é um interesse positivo, mas permite provar que o ART-Q identificou períodos de destaque da economia brasileira.

Durante o período de 2018 até 2019 (fim da linha do tempo), ocorreram simultaneamente períodos de destaque para os índices inflação e mortalidade infantil. Revela, portanto, que o Brasil teve suas menores taxas de inflação e mortalidade infantil nesse período. Esse fato pode levantar o questionamento de que por ocorrerem de forma simultânea, pode existir algum fator externo ou acontecimento de importância no Brasil. Esse é um dos motivadores da elaboração deste método: revelar ocorrências não triviais. A simultaneidade dos intervalos, entretanto é detalhe da próxima etapa de execução do ART-Q, a identificação das relações temporais.

Note que existem representações de intervalos que se concentram em apenas um ponto, como pode ser observado com os atributos $export$(2001, 2013, 2015 e 2019), $imuniza12-23$(2010), $import$(2014) e $investExterno$(2019). Esse fato demonstra o que descrevem as Definições 4.3 e 2.11: um intervalo de interesse pode ser constituído, simplesmente, por um ponto temporal. Esses intervalos, portanto, podem se relacionar com quaisquer outros intervalos de interesse, desde que se repeite a janela máxima para se considerar

uma relação temporal (MWR, previamente estipulada). Desta forma é possível afirmar, também, que no ano de 2018 tanto a importação, quanto a exportação apresentaram valores acima da normalidade.

Figura 5.6: Intervalos de interesses dos índices socioeconômicos do Brasil (BDBra), - experimento 2.



Fonte: Elaborada pelo autor.

A partir dos intervalos de interesse para cada atributo de $BDBra$ identificado, a próxima etapa de execução do ART-Q concentra-se na construção da base de dados de relações temporais. No total, 21 relações foram identificadas pelo ART-Q entre os intervalos de interesse dos atributos de $BDBra$. As relações compõem 9 registros da base de dados de relações temporais $BDRelT$. A Tabela 5.4, abaixo, apresenta a $BDRelT$ construída nesta etapa de execução.

Ao observar a Figura 5.6 é possível argumentar que semanticamente a relação EQUAL(morte_5;in seja idêntica à EQUAL(infla; morte_5), portanto se houvesse ordenação alfabética entre os termos das relações, a frequência de relações como esta seriam incrementadas e potencializariam a probabilidade de serem consideradas padrões.

Entretanto por se iniciarem e terminarem no mesmo ponto temporal, ambos descrevem exatamente a mesma relação (mesma semântica). Desta forma, a não ordenação dos atributos impede o ART-Q de cometer um equívoco. Se fossem ordenados seus termos, teriam seus valores de suporte elevados, o que não seria reflexo da realidade. O mesmo ocorre com as relações $EQUAL(export; import)$ e $EQUAL(import; export)$.

Devido ao fato da $BDRelT$ compreender, neste experimento, apenas nove registros,

Tabela 5.4: Base de dados (*BDRelT*) de relações entre os intervalos temporais de interesse dos atributos de *BDBra* - experimento 2.

t_i	Relações da AIA no tempo t_i
1	STARTS(cres_pop;Imuniza12-23), CONTAINS(cres_pop;export), MEETS(cres_pop,investExterno)
2	STARTS(Imuniza12-23;cres_pop), BEFORE(Imuniza12-23;export)
3	STARTS(Imuniza12-23;import), BEFORE(Imuniza12-23;per_capita)
4	STARTS(import;Imuniza12-23), OVERLAPS(import;per_capita), BEFORE(import;export)
5	CONTAINS(per_capita;export), FINISHES(per_capita;import), BEFORE(per_capita;export)
6	EQUAL(morte_5;infla), FINISHES(morte_5;investExterno), FINISHES(morte_5;export)
7	EQUAL(infla;morte_5), FINISHES(infla;investExterno), FINISHES(infla;export)
8	EQUAL(investExterno;export)
9	EQUAL(export;investExterno)

Fonte: Elaborada pelo autor.

para uma relação ser considera padrão é importante que compareça em, pelo menos 2 dos 9 registros de $BDRelT$. O valor de $Min\_sup = 0,1$ não faz sentido ser considerado, uma ve que se uma relação comparece em 1 dos 9 registros, apresenta frequência igual a 0,11 (11%). Ou seja, toda relação seria considerada padrão. Entretanto, ao considerar $Min\_sup = 0,22$ nenhuma relação, ou combinação de relações, retrata um padrão de verdade.

Ainda que essa configuração não tenha permitido a construção de regras de associação, as relações temporais encontradas entre os intervalos de interesse dos atributos de $BDBra$ são ricas em detalhes não triviais e podem revelar informações não triviais, tais como:

- No passado, é possível identificar que um período de destaque nas exportações de bens e serviços ocorreu enquanto era presente um crescimento populacional acima da normalidade, precedendo períodos de crescimento no investimento externo.
- Períodos de alta na importação de bens e serviços que se sobrepuseram ao crescimento na renda per capita. Durante esses períodos de crescimento na renda per capita, houveram destaques na exportação de bens e serviços.

Esta interpretação dos resultados parciais gerados a partir do ART-Q revela o método é, de fato, capaz de contribuir de forma efetiva ao processo de análise de dados por

meio da incorporação do aspecto temporal de forma explícita e a consideração dos dados quantitativos contínuos.

Como evidenciado por algumas vezes neste texto, o método ART-Q foi projetado de maneira a proporcionar flexibilização em sua execução, por meio de combinações de parâmetros. O que proporciona mais de uma forma de análise dos dados. Uma simples modificação na definição dos interesses dos atributos de $BDBra$, por exemplo, muda o foco da análise.

Na Figura 5.7(a) é possível observar todos os intervalos de interesses dos atributos de $BDBra$ quando na definição dos parâmetros, os comportamentos de interesse dos atributos sejam o **normal**. Já a Figura 5.7(b) ilustra uma outra visão, na qual os intervalos de interesse dos atributos são **fora do normal**, ou seja, o comportamento de interesse dos atributos é definido para quando assumem valores fora da normalidade. É possível observar, na figura, que os intervalos de interesse dos atributos, ilustrados nas imagens (a) e (b) se complementam.

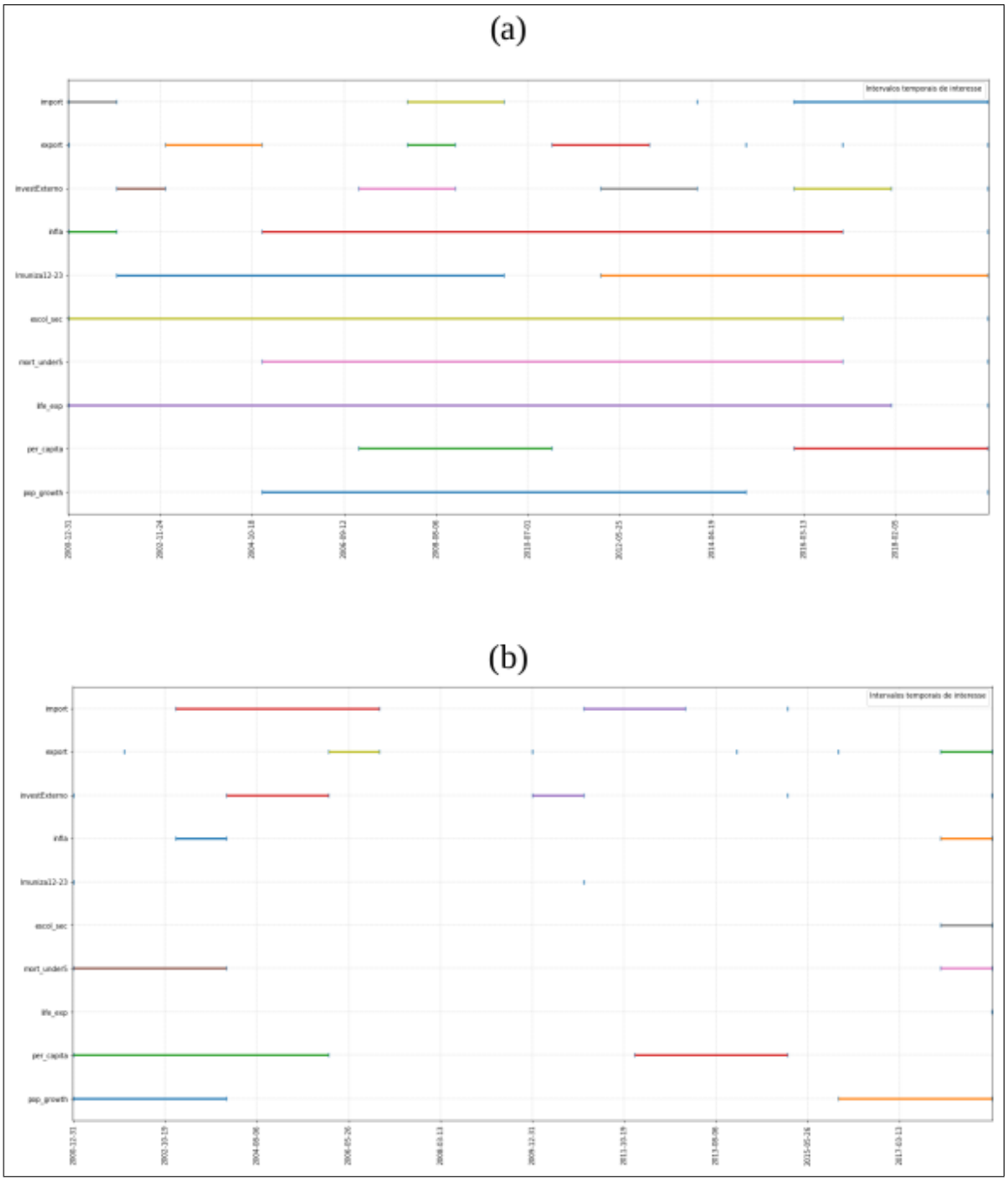
Por fim, a própria permissibilidade do ART-Q em poder testar diferentes comportamentos de interesses para cada um dos atributos promove uma forma de traçar comparações entre os intervalos de interesse dos atributos, sem a necessidade da identificação de padrões e construção de regras de associação.

A Figura 5.7 confirma esta afirmação. Quando configurado para uma análise com a visão ilustrada pela Figura 5.7 (a), ou seja, comportamentos de interesse definido para valores dentro da normalidade, foram encontrados 261 relações temporais (dispostas em 30 registros da $BDRelT$ construída). A partir das relações, 4 padrões temporais puderam ser identificados, entretanto nenhuma regra de associação foi construída pois os padrões são unitários.

Já para a visão representada pela Figura 5.7 (b), 115 relações temporais foram encontradas e compuseram 25 registros da $BDRelT$. Após a execução da etapa de identificação de padrões, 1 padrão temporal foi identificado. A Tabela 5.5 relaciona os padrões encontrados segundo as visões (a) e (b). Note que o maior valor de suporte (frequência) encontrado é o de 0,1333, o que representa a ocorrência em 4 dos 30 registros que compõem $BDRelT$, para a visão (a). Os demais padrões identificados, de acordo com a visão (a), apresentam valor de suporte em 0,1 (10%, ou 3 dos 30 registros de $BDRelT$).

Ainda que 0,133 seja um valor baixo para o suporte, o reduzido tamanho da base de dados $BDBra$ (20 registros anuais) deve ser levado em consideração. Isso indica

Figura 5.7: Intervalos de interesses dos índices socioeconômicos do Brasil (BDBra) quando os intervalos de interesse são definidos (a) normal e (b) fora do normal - experimento 2.



Fonte: Elaborada pelo autor.

que analisando apenas 20 registros de índices socioeconômicos. É possível afirmar que houveram situações suficientes, no passado, para configurar um padrão no comportamento dos índices exportação de bens e serviços e investimento externo. Segundo revelado pelo ART-Q:

...é comum que períodos de estabilidade nos valores de exportação sejam precedidos de períodos de estabilidade nos valores que descrevem o investimento externo.

Desta forma, ao observar o término de um período de estabilidade das exportações, pode-se esperar após um curto período de tempo que os índices de investimento externo se mantenham estáveis, também.

Tabela 5.5: Padrões encontrados pelo ART-Q em *BDRelT* a partir de duas visões: (a) comportamento de interesse do atributo é normal e (b) comportamento de interesse do atributo é fora do normal - experimento 2.

Visão (a): $\mu - 1 * \sigma \leq v_{ia} \leq \mu + 1 * \sigma$ 261 relações temporais - 30 registros	
Padrão	Suporte
BEFORE(export;export)	0,1
BEFORE(export;investExterno)	0,133
BEFORE(import;export)	0,1
BEFORE(investExterno;export)	0,1
Visão (b): $(v_{ia} \leq \mu - 1 * \sigma) \bigcup (v_{ia} \geq \mu + 1 * \sigma)$ 115 relações temporais - 25 registros	
Padrão	Suporte
BEFORE(export;import)	0,12

Fonte: Elaborada pelo autor.

A análise segundo a visão (b) proporciona, entretanto apenas um padrão frequente: $BEFORE(export; import)$ com valor de suporte igual a 0,12 (3 dos 25 registros). O que representa a seguinte informação:

é comum que períodos de exportação fora do normal sejam precedidos (não imediatamente) de períodos de importação de bens e serviços fora do normal.

Ainda que regras de associação não tenham sido geradas, mais uma vez, fica claro que o ART-Q cumpre o seu propósito de identificar relacionamentos interessantes em dados quantitativos e contínuos quando a temporalidade em sua forma explícita é admitida.

5.2.3 Evolução do Experimento 2 - incorporação de mais países

As execuções do ART-Q sob a base de dados *BDBra* revelaram-se contributivas, como era esperado. Entretanto devido ao reduzido conjunto de dados iniciais e também

das relações identificadas, regras de associação não puderam ser construídas. Ainda que esse fato não impossibilite o ART-Q de prover um resultado satisfatório quanto à análise temporal dos dados quantitativos contínuos, uma nova base de dados foi considerada. Esta, porém, constituída por índices socioeconômicos de mais países, além do Brasil.

Foram selecionados a compor a base de dados deste experimento, alguns países que representam fortes potências econômicas mundiais e países geograficamente semelhantes ao Brasil. São, portanto, referências para comparações que intentam revelar informações implícitas que deem ao Brasil uma posição de destaque. A nova base de dados, intitulada *BDSocio*, é constituída pelo mesmo conjunto de indicadores e o mesmo intervalo (20 registros que representam 20 anos de índices) índices socioeconômicos que compõem a *BDBra*, porém referentes aos países Brasil, Argentina, Chile, China e Estados Unidos.

A justificativa pela escolha destes países dá-se pelo simples fato que os países Estados Unidos e China se destacam, atualmente, como duas das maiores potências econômicas do mundo. O primeiro deles sempre se manteve acima da media mundial, enquanto o segundo cresceu de forma considerável e, segundo descrevem Zmoginski, Inohara e Yong (2020), se manteve com crescimento em seu PIB (Produto Interno Bruto) por 4 décadas, podendo atualmente apresentar a primeira retração desde 1976, conforme indica o Escritório Nacional de Estatísticas da China (ENE).

Argentina e Chile compartilham, com o Brasil, os melhores indicadores econômicos da América latina. Em 2019, último ano de publicação do Relatório do Desenvolvimento Humano, como descreve ONU (2019), divulgado no site da ONU, apenas dois países latino-americanos estavam entre os 50 primeiros do mundo no Índice de Desenvolvimento Humano (IDH): Chile, em 44º, e Argentina, em 48º. Portanto, as relações temporais entre os indicadores destes países tendem a ser mais contributivas às comparações envolvendo o Brasil. A Figura 5.8 apresenta uma visão geral da base de dados *BDSocio* utilizada neste experimento, os marcadores ARG, BRA, CHIL, CHIN e EUA representam, respectivamente, os países Argentina, Brasil, Chile, China e Estados Unidos.

As bases de dados selecionadas foram mantidas quase que na totalidade em sua forma bruta, como podem ser obtidas do site do Banco Mundial. As únicas tarefas de pré-processamento de dados realizadas foram: (a) a unificação das bases de dados de cada país em uma base única, a *BDSocio* e (b) a transposição das linhas e colunas, uma vez que para servirem como entrada do ART-Q, as bases de dados devem apresentar os atributos dispostos nas colunas e os registros dispostos em linhas (busca vertical para cada atributo). Originalmente, as bases de dados adquiridas apresentavam como colunas

Figura 5.8: BDSocio: base de dados que contempla índices socioeconômicos dos países Argentina, Brasil, Chile, China e Estados Unidos, no período de 2000 a 2019 - experimento 2.

Data	cres_pop	per_capita	exp_vida	morte_5	escol_sec	Imuniza12-23	infla	investExterno	export	import	ARG_cres_pop	ARG_per_capita		EUA_ export
31/12/2000	1.4	3930.0	70.1	34.8	99.0	?	5.6	10.2	12.5	3.29	1.1	7470		14.4
31/12/2001	1.4	3350.0	70.5	32.6	99.0	?	8.2	12.4	14.6	2.32	1.1	7000		13.2
31/12/2002	1.3	3090.0	70.8	30.5	96.0	110.0	9.8	14.2	13.4	1.65	1.1	4040		13
31/12/2003	1.3	2980.0	71.2	28.5	97.0	101.9	14.1	15.2	13.0	1.01	1.1	3650		13.4
31/12/2004	1.2	3340.0	71.5	26.6	97.0	102.4	7.8	16.5	13.1	1.81	1.1	3370		14.7
31/12/2005	1.1	4000.0	71.9	24.9	98.0	101.3	7.4	15.2	11.8	1.54	1	4260		15.5
31/12/2006	1.1	4850.0	72.3	23.3	99.0	?	6.8	14.4	11.7	1.93	1	5480		16.2
31/12/2007	1.0	6190.0	72.6	21.9	99.0	95.2	6.4	13.3	12.0	4.45	1	6510		16.5
31/12/2008	1.0	7620.0	73.0	20.7	99.0	96.7	8.8	13.5	13.7	5.07	1	7670	...	17.4
31/12/2009	1.0	8320.0	73.3	19.6	99.0	96.7	7.3	10.9	11.3	3.14	1	7800		13.7
31/12/2010	0.9	9660.0	73.6	18.7	99.0	91.5	8.4	10.9	11.9	8.23	0.8	9270		15.7
31/12/2011	0.9	11080.0	73.9	17.9	99.0	95.3	8.3	11.6	12.4	1.02	1.2	10710		17.3
31/12/2012	0.9	12360.0	74.2	17.3	99.0	93.0	7.9	11.9	13.2	9.25	1.1	11890		17
31/12/2013	0.9	12810.0	74.5	16.7	98.0	102.1	7.5	11.7	14.0	7.52	1.1	12870		16.5
31/12/2014	0.9	12100.0	74.7	16.2	97.0	100.8	7.8	11.0	13.7	8.77	1.1	12350		16.4
31/12/2015	0.8	10160.0	75.0	15.7	96.0	100.1	7.6	12.9	14.1	6.47	1.1	12600		15.3
31/12/2016	0.8	8930.0	75.2	16.3	95.0	101.5	8.1	12.5	12.1	7.42	1.1	12220		14.6
31/12/2017	0.8	8670.0	75.5	14.8	91.0	100.8	3.5	12.6	11.6	6.88	1	13120		15
31/12/2018	0.8	9140.0	?	14.4	84.0	?	3.0	14.8	14.3	7.88	1	12390		15.3
31/12/2019	1.0	8320.0	73.0	19.6	98.0	101.3	7.8	12.6	13.0	5.07	1.1	13221		15.5

Fonte: Elaborada pelo autor.

os anos e em suas linhas, os índices.

As tarefas de unificação e transposição das bases de dados foram realizadas por meio de um *script* em linguagem *Python* (versão 3.7), com auxílio da ferramenta *Orange Canvas*. O Algoritmo 6, abaixo, descreve em alto nível o processo de unificação e transposição das bases de dados dos países selecionados para compor a *BDSocio*. O *script*, basicamente, seleciona cada uma das N bases de dados de países (originais) e incorpora a BD_{temp}, uma base de dados temporária. No processo de incorporação de cada base de dados (ds_i), a intersecção entre ds_i e BD_{temp} representa o atributo *Data* (temporal).

Algoritmo 6 Procedimento de unificação e transposição das bases de dados que compõem a BDSocio.

1: **procedure** CONSTRÓI BDSOCIO($ds_1, ds_2, ..., ds_N$)
2: $BDUnificada \leftarrow \emptyset$;
3: $BD_{temp} \leftarrow \emptyset$;
4: **for** $i \leftarrow 1;\ i < N;\ i++$ **do**
5: $BD_{temp} \leftarrow BD_{temp} \bigcup (ds_i - (BD_{temp} \bigcap ds_i))$;
6: **end for**
7: $BDUnificada \leftarrow BD_{temp}^{T}$; ▷ Transposta de BD_{temp}
8: **return** $BDUnificada$;
9: **end procedure**

5.2.4 Condução da evolução do experimento 2

Para a condução do experimento que envolve a base de dados *BDSocio*, quanto aos interesses dos atributos que a compõem, os seguintes cenários foram idealizadas:

(a) O Brasil se destaca no mundo pois apresenta valores em seus índices socioeconômicos acima da normalidade, ou seja, apresenta crescimento econômico enquanto os demais países se mantém na normalidade;

(b) O Brasil apresenta valores em seus índices socioeconômicos acima da normalidade enquanto os demais países também se encontram na mesma situação - crescimento global.

Em outras palavras, para o cenário (a), todos os índices do Brasil (exceto os referentes à inflação e mortalidade infantil) tem seus pontos de interesse definidos **acima do normal**. Para todos os demais índices, de todos outros países, os valores de interesse são **normais**. Já no cenário (b), para todos os países, todos os índices (exceto os referentes à inflação e mortalidade infantil) tem seus pontos de interesse definidos para valores **acima do normal**. Para ambos os cenários (a e b), foram testadas várias combinações para os parâmetros Min_sup, Min_conf, MWI e MWR. Os resultados obtidos dessa bateria de execuções podem ser observados na Tabela 5.6, juntamente com as configurações consideradas para atingir o resultado descrito.

A partir da primeira coluna, mais à esquerda, a tabela apresenta um identificador para a execução do ART-Q (ex_i), seguido pelos valores de Min_sup, Min_conf, MWI e MWR, respectivamente. Logo em seguida encontram-se os resultados da execução ex_i, descritos pela quantidade de relações temporais identificadas entre os intervalos de interesse, a quantidade de registros que compõem a $BDRelT$ (a base de dados de relações temporais), quantos padrões temporais foram identificados e quantas regras de associação puderam ser geradas a partir dos padrões identificados.

Observe, na tabela, que as combinações de parâmetros nem sempre contemplam os mesmos valores para Min_sup. Nos experimentos os quais os parâmetros se mantiveram iguais, com exceção do parâmetro Min_conf, ou seja, visaram observar os resultados obtidos com o decréscimo deste parâmetro, tiveram como ponto de interrupção os momentos em que ao decrementar o valor de Min_conf a quantidade de padrões e regras de associação se mantiveram sem alterações. Este fato pode ser observado nas execuções ex_8 / ex_9, também em ex_{10} / ex_{11}, ex_{18} / ex_{19} e ex_{20} / ex_{21}.

Naturalmente, pela grande quantidade de regras de associação geradas pelas execuções do método ART-Q sob a base de dados $BDSocio$, é inviável discutir todas elas. Entretanto, algumas delas se destacam:

R_i BEFORE(EUA_import;CHIN_infla) $\rightarrow$ STARTS(EUA_import;EUA_export):

Tabela 5.6: Resultados das execuções do ART-Q na base de dados *BDSocio* quando considerados dois cenários para a análise dos dados - experimento 2.

#	Resultados							
	Min_sup	Min_conf	MWI	MWR	Relações	|BDRelT|	Padrões	Regras
				Cenário (a)				
ex_1	0,1	0,5	370	750	4122	115	0	0
ex_2	0,05	0,5	370	750	4122	115	0	0
ex_3	0,02	0,5	370	750	4122	115	37	15
ex_4	0,01	0,5	370	750	4122	115	3755	10587
ex_5	0,1	0,5	750	750	2955	89	0	0
ex_6	0,05	0,5	750	750	2955	89	0	0
ex_7	0,02	0,5	750	750	2955	89	205	201
ex_8	0,02	0,4	370	750	4122	115	37	15
ex_9	0,02	0,3	370	750	4122	115	37	15
ex_{10}	0,02	0,4	750	750	2955	89	205	201
ex_{11}	0,02	0,3	750	750	2955	89	205	201
				Cenário (b)				
ex_{12}	0,1	0,5	370	750	1712	88	0	0
ex_{13}	0,05	0,5	370	750	1712	88	1	0
ex_{14}	0,02	0,5	370	750	1712	88	195	173
ex_{15}	0,1	0,5	750	750	1216	68	0	0
ex_{16}	0,05	0,5	750	750	1216	68	0	0
ex_{17}	0,02	0,5	750	750	1216	68	23	7
ex_{18}	0,02	0,4	370	750	1712	88	195	182
ex_{19}	0,02	0,3	370	750	1712	88	195	182
ex_{20}	0,02	0,4	750	750	1216	68	23	7
ex_{21}	0,02	0,3	750	750	1216	68	23	7

Fonte: Elaborada pelo autor.

$Sup = 0,026|Conf = 1,0|Lift = 38,333|Conv = 0,0$;

R_j BEFORE(import;ARG_infla) $\rightarrow$ BEFORE(export;CHIL_escol_sec): $Sup = 0,022|Conf = 0,667|Lift = 29,333|Conv = 2,935$;

R_k BEFORE(CHIL_escol_sec;CHIN_infla) $\rightarrow$ FINISHES(CHIL_escol_sec;export): $Sup = 0,0227|Conf = 0,667|Lift = 29,333|Conv = 2,935$.

Que podem ser interpretadas da seguinte maneira:

R_i: *Ao observar um período de estabilidade na inflação da China após um crescimento nos índices de importações de bens e serviços dos Estados Unidos é esperado (com frequência de 2,6% e 100% de confiança nesta afirmação) que se observe, também, períodos de crescente importação e exportação que se iniciam simultaneamente nos Estados Unidos*; - Cenário (a)

R_j Períodos de crescimento na importação de bens e serviços do Brasil que antecedem (não imediatamente) períodos de estabilidade na inflação da Argentina podem indicar (com 66,7% de confiança nesta afirmação) que períodos de estabilidade nos índices de escolarização secundária do Chile ocorram após um período de crescimento nas exportações do Brasil; - Cenário (a)

R_k Períodos de crescimento nos índices de escolaridade secundária do Chile que antecedem (não imediatamente) períodos de estabilidade na inflação chinesa podem indicar (com 66,7% de confiança) que um período de bons índices de exportação de bens e serviços do Brasil se finalize ao mesmo tempo que se finaliza um período de crescimento nos índices de escolaridade secundária do Chile. - Cenário (b)

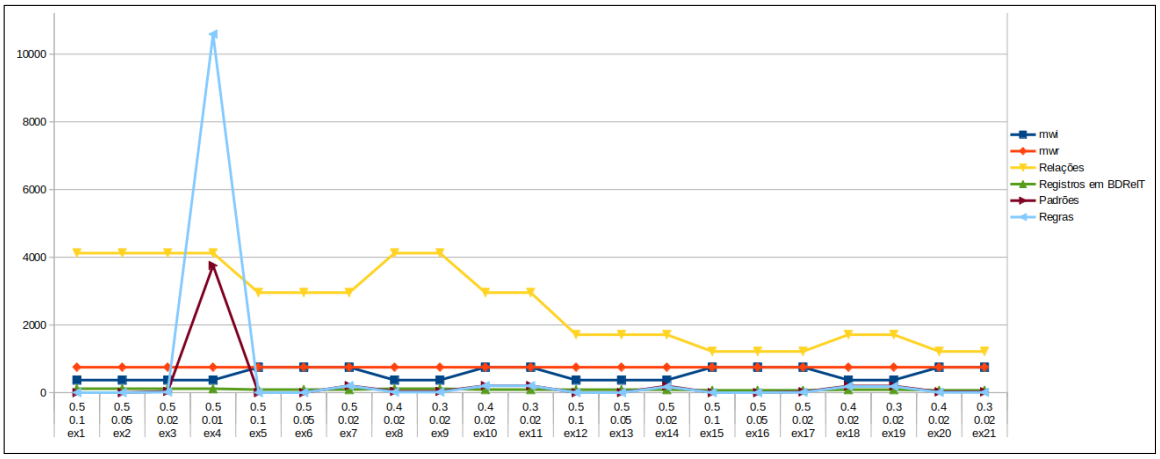
A execução ex_{13} teve como resultado apenas um único padrão, como pode ser observado na Tabela 5.6, o padrão *BEFORE(CHIL_infla; CHIL_infla)*, com valor de suporte igual a $0,568$. Semanticamente, o padrão pode ser interpretado como: *"Chile apresenta recorrências (não imediatas) de períodos nos quais seus seus índices de inflação são controlados (dentro da normalidade), enquanto tem-se um cenário de crescimento global."*. Essa informação é presente em 56,8% dos dados analisados.

A fim de permitir uma visualização diferente para os resultados obtidos com o ART-Q neste experimento, a Figura 5.9 foi incluída nesta discussão. Ela ilustra as informações contidas na Tabela 5.6 a respeito das execuções do ART-Q considerando os dois cenários (a) e (b) sob a base de dados *BDSocio*.

É interessante ressaltar que, com exceção da configuração considerada na execução ex_4, as quantidades de relações entre intervalos de interesse, padrões temporais e regras de associação não sofrem com a explosão da dimensionalidade. A quantidade de relações (linha amarela) acompanha as curvas traçadas pelos valores de MWI na maioria das execuções.

Assim como também ocorre com as linhas que descrevem a quantidade de padrões temporais identificados e regras de associação geradas. A figura demonstra, ainda, que uma boa escolha nos valores dos parâmetros facilita o trabalho de análise dos resultados obtidos. As visualizações dos intervalos de interesse identificados para os diferentes cenários (**a** e **b**), quando consideram diferentes janelas para a construção de um intervalo estão apresentadas no Apêndice A.

Figura 5.9: Gráfico a respeito das combinações de parâmetros para as execuções do ART-Q sob a base de dados *BDSocio* - experimento 2.



Fonte: Elaborada pelo autor.

5.3 Experimento 3 - Dados meteorológicos

O terceiro experimento idealizado para defender as contribuições que o método ART-Q pode prover, considera uma base de dados meteorológicos. Trata-se de uma base de dados, intitulada *BDTempo*, que é composta por informações climatológicas de quatro cidades do estado de São Paulo, a saber: Franca, São Carlos, São Paulo e Votuporanga. Os dados foram obtidos do BDMEP (Banco de Dados Meteorológicos para Ensino e Pesquisa) do INMET (Instituto Nacional de Meteorologia) e compreendem o período entre 01/01/2015 a 31/12/2019. No total, a *BDTempo* é composta por 1633 registros dispostos em um arquivo do tipo *csv*.

A escolha por estas cidades levou em consideração o fato que elas estão geograficamente localizadas em regiões climatologicamente diferentes. São Paulo, localizada mais à leste do estado apresenta clima mais frio, assim como São Carlos (mais ao centro do estado), Votuporanga à oeste e Franca ao norte, ainda que geograficamente em regiões distantes, são conhecidas por apresentarem temperaturas mais elevadas. Outras cidades que atendem ao mesmo critério não foram selecionadas para compor a base de dados, tais como Presidente Prudente e Campos do Jordão, que tem climas bem definidos, devido ao fato que, para ambas as cidades, a base de dados do BDMEP não contém dados nos mesmos períodos que as demais cidades consideradas. De fato esse cenário não impede o correto funcionamento do ART-Q, mas para garantir que nenhum viés pudesse interferir no processo, a opção pela adoção de uma base de dados completa (sem dados faltantes) foi feita.

A Tabela 5.7 apresenta uma descrição dos atributos que compõem a *BDTempo*. A partir da esquerda, é possível ver na tabela a coluna que representa a descrição de cada um dos atributos, seguida pelas colunas que relacionam cada uma das cidades. Cada registro da tabela descreve um atributo e como cada atributo pode ser identificado (nome) na *BDtempo*. Por exemplo, o atributo que descreve a umidade relativa do ar é identificado como *Frumi* para a cidade de Franca, *Scumi, Spumi* e *Vtumi* para as cidades de São Carlos, São Paulo e Votuporanga, respectivamente.

Tabela 5.7: Descrição dos atributos que compõem a base de dados *BDTempo* - experimento 3.

Descrição	Franca	São Carlos	São Paulo	Votuporanga
Temperatura máxima	FrtempMax	SctempMax	SptempMax	VttempMax
Insolação	Frinso	Scinso	Spinso	Vtinso
Evaporação do piche	Frevap	Scevap	Spevap	Vtevap
Umidade relativa do ar	Frumi	Scumi	Spumi	Vtumi
Precipitação de chuva	Frprec	Scprec	Spprec	Vtprec
Temperatura Mínima	Frtempmin	Sctempmin	Sptempmin	Vttempmin

Fonte: Elaborada pelo autor.

O atributo temporal que descreve o momento da coleta das informações que compõem cada um dos registros de *BDTempo* não é contemplado pela Tabela 5.7. Isso se deve pelo fato que a base de dados *BDTempo* foi constituída a partir do mesmo procedimento descrito no experimento anterior. Isto é, para compor a base de dados, um conjunto de dados de cada cidade selecionada foi extraído do BDMEP e, por meio da estratégia descrita pelo Algoritmo 6 tais conjuntos de dados foram unificados para compor uma única base de dados, a *BDTempo*. Portanto, os registros foram unificados de acordo com a mesma data de coleta dos dados. Assim somente um atributo data é necessário para descrever os valores de todas as cidades.

5.3.1 Condução do experimento e Resultados obtidos

A estratégia utilizada para a condução deste terceiro experimento foi a de testar diversas possibilidades de combinações dos parâmetros do ART-Q. Desta forma, a quantidade de relações temporais, padrões e regras de associação podem ser observadas e o conjunto dos melhores parâmetros, então, selecionado. Entretanto para a definição dos comportamentos de interesse dos atributos, em todas as execuções deste experimento, foi considerado que os valores de interesse são aqueles que ocorrem acima, ou fora, do normal. Obviamente essa não é a única visão de análise que pode ser conduzida com esta base de

dados, entretanto para este experimento, foi decidido que a anormalidade é foco da busca de informações não triviais.

Devido ao fato que os registros da $BDTempo$ são diários, optou-se pela adoção de três possibilidades quanto às janelas máximas para consideração de um intervalo temporal de interesse (o parâmetro MWI). Inicialmente os testes foram realizados com MWI definido com o valor 7, ou seja, para um atributo, quando apresenta mais de um valor de interesse dentro do período de 7 dias, um intervalo de interesse é configurado. Em outras palavras, essa escolha foi motivada pelo fato que, uma vez que os comportamentos de interesse dos atributos foram definidos para acima, ou fora da normalidade, se um valor (por exemplo) de chuva fora da normalidade se repete dentro de um período de 7 dias, pode-se dizer que a semana é uma semana atípica.

Posteriormente, a janela máxima para configuração de um intervalo de interesse foi modificada para o valor 15. O que representa que a ocorrência de duas anormalidades em um período de 15 dias configura um intervalo de interesse. Por fim, MWI foi reajustado para o valor 3. Dessa forma, somente são considerados para compor um intervalo interessante aqueles valores que se encontram fora da normalidade quase que consecutivamente (no máximo 3 dias de distanciamento). Acredita-se que padrões nas relações temporais compostas por comportamentos de interesse que consideram esse período podem ser os mais contributivos para previsões futuras.

As Tabelas 5.8, 5.9 e 5.10, apresentadas abaixo, tecem um comparativo entre os resultados das execuções do ART-Q quando o parâmetro MWI foi definido com os valores 7, 15 e 3, respectivamente.

A partir da coluna mais à esquerda, nas três tabelas, encontram-se as colunas que descrevem, respectivamente, (1) o identificador da execução realizada, os valores de (2) suporte mínimo para constituir um padrão, (3) o valor mínimo de confiança para uma regra ser dita forte, (4) qual o comportamento de interesse dos atributos, (5) a quantidade de relações temporais encontradas entre os intervalos de interesse dos atributos, (6) qual o tamanho (quantidade de registros) da base de dados $BDRelT$, (7) a quantidade de padrões identificadas considerando o valor de Min_sup e (8) a quantidade de regras de associação fortes geradas a partir dos padrões identificados e consideradas fortes pelo parâmetro Min_conf.

Tabela 5.8: Resultados das execuções do ART-Q quando considera MWI = 7 e MWR = 7 - experimento 3.

#	Min_sup	Min_conf	Interesse	Relações	Registros	Padrões	Regras
ex_1	0,1	0,5	acima	12388	1235	0	0
ex_2	0,02	0,5	acima	12388	1235	26	0
ex_3	0,01	0,5	acima	12388	1235	265	34
ex_4	0,01	0,4	acima	12388	1235	265	35
ex_5	0,1	0,5	fora	21882	1401	0	0
ex_6	0,02	0,5	fora	21882	1401	4	0
ex_7	0,01	0,5	fora	21882	1401	571	235
ex_8	0,01	0,4	fora	21882	1401	571	236

Fonte: Elaborada pelo autor.

Tabela 5.9: Resultados das execuções do ART-Q quando considera MWI = 15 e MWR = 7 - experimento 3.

#	Min_sup	Min_conf	Interesse	Relações	Registros	Padrões	Regras
ex_9	0,1	0,5	acima	5731	521	0	0
ex_{10}	0,02	0,5	acima	5731	521	2	0
ex_{11}	0,01	0,5	acima	5731	521	205	45
ex_{12}	0,01	0,4	acima	5731	521	205	45
ex_{13}	0,1	0,5	fora	6912	418	0	0
ex_{14}	0,02	0,5	fora	6912	418	67	15
ex_{15}	0,01	0,5	fora	6912	418	6331	17618
ex_{16}	0,01	0,4	fora	6912	418	6331	17751

Fonte: Elaborada pelo autor.

Tabela 5.10: Resultados das execuções do ART-Q quando considera MWI = 3 e MWR = 7 - experimento 3.

#	Min_sup	Min_conf	Interesse	Relações	Registros	Padrões	Regras
ex_{17}	0,1	0,5	acima	25933	2115	0	0
ex_{18}	0,02	0,5	acima	25933	2115	126	9
ex_{19}	0,01	0,5	acima	25933	2115	949	405
ex_{20}	0,01	0,4	acima	25933	2115	949	556
ex_{21}	0,1	0,5	fora	71777	3442	0	0
ex_{22}	0,02	0,5	fora	71777	3442	1389	1188
ex_{23}	0,01	0,5	fora	71777	3442	22584	27933
ex_{24}	0,01	0,4	fora	71777	3442	22584	38189

Fonte: Elaborada pelo autor.

Como era previsto, o volume de informações resultante nas execuções que consideram $MWI = 3$ é muito superior àqueles dos demais. Consequentemente, para chegar ao resultado, o ART-Q tomou mais tempo. Foram 48 minutos e 13 segundos para a conclusão das execuções ex_{23} e ex_{24} enquanto a média das demais foi de 3 minutos e 35 segundos.

A observação das tabelas revela que mesmo quando são considerados valores relativamente baixos para Min_sup (entre $0,02$ e $0,01$), uma grande quantidade de registros de $BDRelT$ devem contemplar um item (ou conjunto de itens) para que ele seja considerado um padrão. Tome como exemplo as execuções ex_{19} e ex_{23}. Ainda que os valores de Min_sup definidos em $0,01$ possam parecer baixos, eles indicam que uma relação temporal (ou um conjunto de relações) é um padrão pois ocorre 22 e 35 vezes, respectivamente. O que, visto deste ângulo, demonstra não ser um valor baixo. Ainda mais quando se recorda que uma relação temporal de interesse contempla intervalos temporais de duração de vários dias. As imagens dos intervalos de interesse para as configurações consideradas neste experimento são apresentadas no Apêndice B.

Naturalmente é inviável a apresentação e discussão de todos os padrões e regras de associação resultantes de cada execução da evolução deste experimento 3. Entretanto alguns exemplos foram selecionados entre os resultados e são discutidos a seguir.

A partir da execução ex_4 do ART-Q, o padrão mais frequente (aquele com maior valor de suporte) foi o *BEFORE(Spevap;Spumi)*, com suporte igual a 0,0315. O que pode ser semanticamente interpretado como a recorrência, na cidade de São Paulo, de períodos (semanas) acima da normalidade da umidade relativa do ar após de valores elevados de evaporação do piche. Esse fato foi observado no passado 39 vezes (1235 registros e $suporte = 0,0315$).

Quanto às regras de associação geradas, a regra $R: BEFORE(SptempMax; Frprec) \rightarrow BEFORE(SptempMax; Spprec) | Sup: 0,0137, Conf: 0,81, Lift: 33,325, Conv: 5,135$ selecionada pode ser vista sob a seguinte semântica:

> *As semanas nas quais os valores de temperatura máxima na cidade de São Paulo ocorrem acima da normalidade antecedendo (no máximo uma semana) semanas com valores de precipitação de chuva elevados podem indicar, com confiança de 81% nesta afirmação, semanas de temperaturas máximas elevadas na cidade de São Paulo antecedendo (não imediatamente) semanas de precipitações de chuva elevadas na cidade de Franca.*

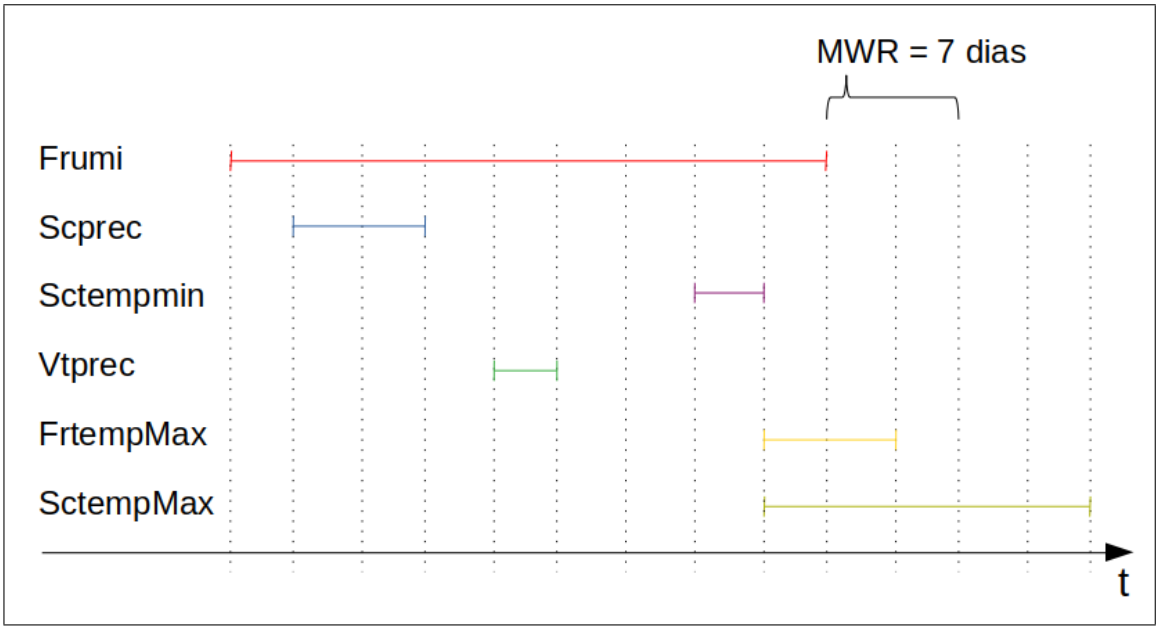
A execução ex_{16} revelou que no passado, períodos de, pelo menos 15 dias, fora do normal (acima ou abaixo) na precipitação de chuvas na cidade de São Paulo ocorreram, com frequência, junto com períodos fora do normal para a temperatura mínima. Fato este que ocorreu, pelo menos 13 vezes (padrão $CONTAINS$ ($Sptempmin$; $Spprec$) | $suporte = 0,0311$).

A regra $R: CONTAINS\ (Frumi; Scprec),\ CON\ TAINS\ (Frumi; Sctempmin),$ $CONTAINS(\ Frumi\ ;\ Vtprec), OVERLAPS\ (Frumi; FrtempMax) \rightarrow OVERLAPS$ $(Frumi; SctempMax)\ |\ Sup: 0,0119,\ Conf: 1,0,\ Lift: 69,66,\ Conv: 0,0,$ por sua vez, pode ser entendida como:

> *Períodos quinzenais de precipitação de chuva fora do normal (excesso ou falta) na cidade de São Carlos que ocorrem durante períodos quinzenais de valores fora do normal na umidade relativa do ar na cidade de Franca, quando identificados juntamente com períodos quinzenais fora da normal de valores de temperatura mínima na cidade de São Carlos e precipitação na cidade de Franca e, na cidade de Franca, os valores de umidade relativa do ar se encontram fora do normal por um período que se sobrepõe a um período de temperaturas máximas fora do normal na mesma cidade, implica na ocorrência de um período de valores anormais para temperaturas máximas na cidade de São Carlos.*

A fim de simplificar a compreensão desta explicação, a Figura 5.10 apresenta uma visualização da regra acima discutida.

Figura 5.10: Visualização de uma regra de associação identificada pelo ART-Q na base de dados *BDTempo* - experimento 3.



Fonte: Elaborada pelo autor.

A descrição da regra acima citada foi propositalmente selecionada para demonstrar a quantidade de informação que apenas uma regra de associação gerada pelo ART-Q pode representar. Observe a regra $R: BEFORE(\ Spinso; Frevap),\ BEFORE(Spinso;$

$FrtempMax$), $BEFORE(Spinso; Spevap)$, $BEFORE\ (Spinso; Spinso) \rightarrow BEFORE($ $Spinso; Spumi)$ | $Sup : 0,010$, $Conf : 0,947$, $Lift : 27,87, Conv : 18,22$, que relaciona 10 intervalos de interesse de 5 atributos em 2 cidades.

A regra acima apresentada foi obtida a partir da execução ex_{24}. Esta regra leva em consideração que um intervalo de interesse pode ser constituído sempre que dois pontos temporais de interesse estiverem a, não mais que, 3 dias distantes um do outro.

5.3.2 Evolução do experimento 3 - fenômeno *El Niño*

El Niño[2] é um fenômeno natural classificado como atmosférico-oceânico que é principalmente caracterizado por ocasionar um aquecimento anormal das águas no Oceano Pacífico Tropical, como descreve BBC (2019). Como consequência do seu surgimento, o clima regional muda o padrão de ventos e chuvas.

Segundo especialistas em entrevista à *BBC News Mundo*, "O *El Niño* altera os padrões de circulação da atmosfera e causa eventos extremos pelo mundo. De inundações na Índia ou na Austrália à costa oeste da América do Sul"(BBC, 2019). No Brasil o *El Niño* tem como consequências de seu surgimento, um moderado aumento das temperaturas médias, principalmente nas regiões sudeste e sul. Também precipitações abundantes e chuvas mais intensas no inverno.

Ainda de acordo com o portal BBC News (2019) durante o período de 1901 a 2017 houveram 33 ocorrências do *El Niño*. Quando o fenômeno é muito agressivo, como ocorreu em 2015, é denominado *Super El Niño*. Entretanto só existem três registros deste fenômeno no mesmo período, em 1982, 1998 e 2015.

Para a realização deste experimento, uma base de dados foi obtida contendo os registros do ano que mais pode revelar informações que descrevem o fenômeno. A base de dados $BDNino$ foi constituída a partir dos registros de $BDTempo$ referentes ao ano de 2015, somente. O intuito desta abordagem é construir um conjunto de regras de associação e intervalos de interesse que possa ser comprados aos intervalos de interesse e regras de associação de um ano sem incidência do fenômeno como o ano de 2015. Para compor a base de dados a ser comparada com a $BDNino$, os registros de $BDTempo$ referentes ao ano de 2018 foram selecionados para compor a intitulada $BDTempo2018$. Isso porque 2018 é um ano tido como de pouca, ou nenhuma, intensidade do *El Niño*.

[2]O nome *El Niño* faz referência ao *Menino Jesus*. Pescadores da costa do Peru o denominaram assim devido ao fato que, pelo aquecimento anormal das águas na época do Natal, cria-se um ambiente propício ao aumento dos peixes na região.

No total, a $BDNino$ é composta por 265 registros, enquanto $BDTempo2018$ contempla 354. A diferença entre a quantidade de registros das bases de dados indica que alguns dias do ano não tiveram seus valores climatológicos incorporados à base de dados do BDMEP. O que serve como forma de ressaltar que para o método ART-Q, lacunas de registros da base de dados não distorcem o resultado obtido, uma vez que a temporalidade explícita é quem vai indicar se registros podem, ou não, compor um mesmo intervalo temporal.

5.3.3 Condução da evolução do experimento 3

A estratégia adotada neste experimento segue o raciocínio que os intervalos dentro do comportamento padrão de um ano de forte intensidade do fenômeno *El Niño* no Brasil são os momentos de interesse. Uma vez que já se configura a anormalidade pela simples presença do fenômeno no ano. Portanto, os parâmetros do ART-Q foram definidos como segue:

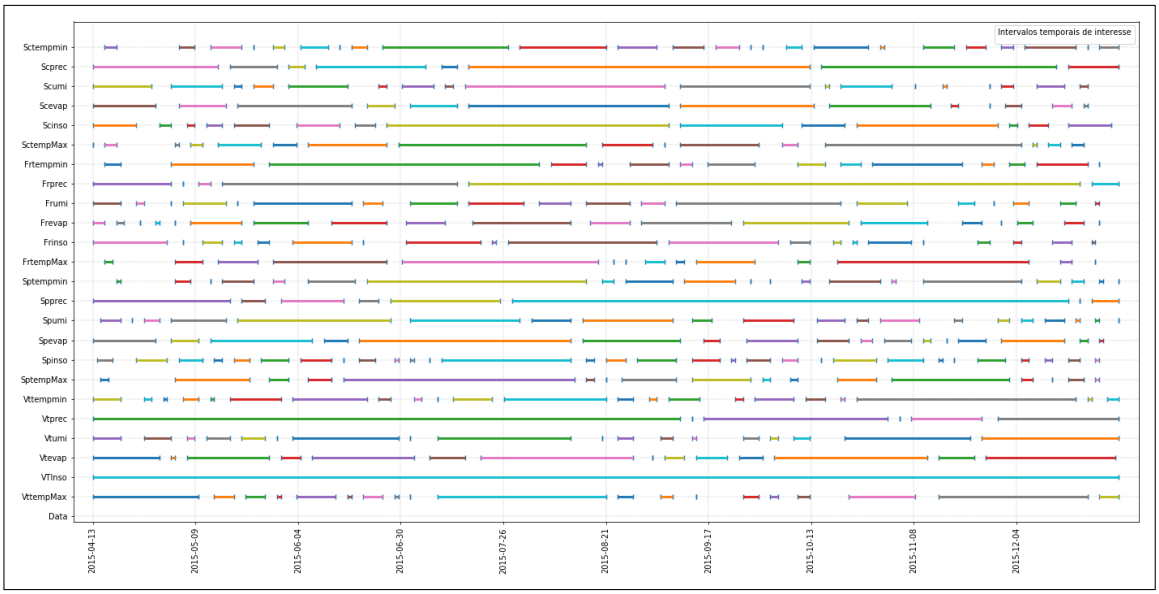
- $Min\_sup = 0,02$;
- $Min\_conf = 0,5$;
- $MWI = 3$;
- $MWR = 5$;
- Comportamentos de interesse = normal.

No total, 9689 relações temporais de interesse foram identificadas pelo ART-Q e dispostas em 378 registros da $BDRelT$. Após a etapa de identificação dos padrões e construção das regras de associação, o ART-Q resultou em 212 padrões e 80 regras de associação. Ao considerar a base de dados com registros do ano de 2018 (sem incidência do *El Niño*, 13258 relações foram identificadas, organizadas em 508 registros da $BDRelT$. Ao final da execução com a mesma configuração, foram identificados 167 padrões e 49 regras de associação.

Abaixo, as figuras 5.11 e 5.12 apresentam os intervalos de interesse dos atributos nos períodos em que o El Niño foi considerado de forte intensidade e de fraca intensidade, respectivamente.

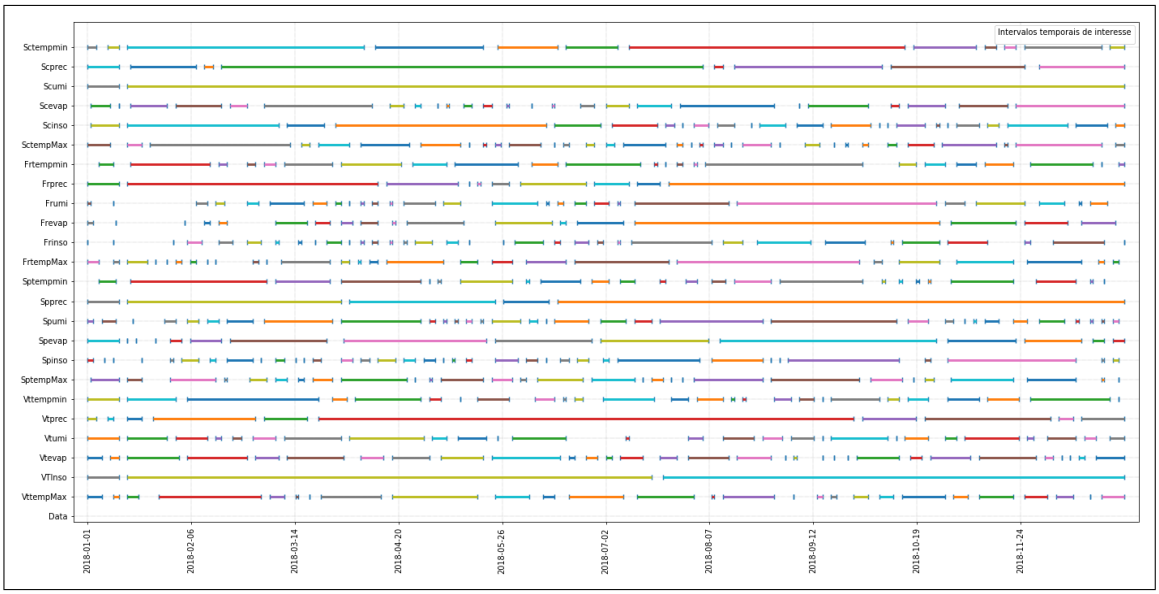
Nota-se, a partir dos resultados obtidos no ano de 2018 (Figura 5.12), que os intervalos de interesse são, no geral, mais longos e em menor quantidade. Em 2015 (Figura 5.11)

Figura 5.11: Intervalos de interesse dos atributos de *BDNino*, que contempla índices do ano de 2015 com comportamento de interesse normal- experimento 3.



Fonte: Elaborada pelo autor.

Figura 5.12: Intervalos de interesse dos atributos de *BDTempo*2018, que contempla índices do ano de 2018 com comportamento de interesse normal- experimento 3.



Fonte: Elaborada pelo autor.

a maior quantidade, mesmo que sutil, de intervalos de interesse, com durações menores, revela mais instabilidade no clima.

Ainda que as imagens se assemelhem, é válido ressaltar que ambas descrevem o comportamento padrão somente de um ano. Pode ser que os valores assumidos pelos atributos sejam distintos nas bases de dados.

As janelas MWI e MWR foram definidas com os valores descritos acima (3 e 5, respectivamente) pois acredita-se que acima destes valores, a distância temporal pode considerar eventos esporádicos como intervalos de interesse e abaixo deste valor, muitas regras podem ser geradas, o que poderia levar à confusão.

Quanto à escolha do valor para o suporte mínimo, vários testes foram realizados a fim de identificar o melhor valor. A adoção de $Min\_sup = 0,01$ proporcionou uma situação na qual relações que comparecem em 4 registros de $BDRelT$ já são consideradas padrões. Uma execução considerando este valor para o parâmetro gerou como resultado um conjunto de 24449 padrões e 94297 regras de associação. O que comprova que é inviável essa definição. Acima de 0,03, por outro lado, o valor de Min_sup não permitiu a identificação de nenhum padrões que pudesse subsidiar a construção de regras de associação.

Quando somente os padrões que consideram o atributo temperatura máxima são selecionados, a execução do ART-Q sob a $BDNino$ revelou um total de 82 padrões, enquanto a execução com os dados de 2018 revelou 114. O que também reforça o fato que existe mais instabilidade no ano que o *El Niño* incide.

5.3.4 Validação dos resultados

Do total de 80 regras de associação geradas pelo ART-Q sob a base de dados $BDNino$ (ao considerar $Min\_sup = 0,02$, $Min\_conf = 0,5$, $MWI = 3$ e $MWR = 5$), 26 delas contemplam o atributo temperatura máxima, o que representa 32,5% do total dos resultados. Regras como $R: BEFORE(Spinso; SctempMax)$, $BEFORE(Spinso; Spinso)$, $\rightarrow$ $BEFORE(Spinso; Sptempmin)$ | $Sup: 0,0211$, $Conf: 0,889$, $Lift: 24,0$, $Conv: 8,675$.

Uma execução sob os mesmo parâmetros, entretanto com a mudança na definição dos comportamentos de interesses, de normal para acima do normal, revelou que 7 dos 10 padrões com maior valor de suporte entre os identificados pelo ART-Q, contemplam o atributo referente à insolação na cidade. Do total de 155 regras de associação geradas, 25 delas contemplam o atributo temperatura máxima e 116 contemplam o atributo insolação.

Essa discussão à respeito dos resultados nestas condições é interessante pois evidencia a capacidade do método ART-Q em revelar informações implícitas. Neste caso, corrobora com as afirmações de especialistas (apresentadas na descrição deste experimento 3) sobre as influências do *El Niño* nos valores de temperatura na região sudeste do Brasil.

Considerações finais

Este capítulo apresentou os experimentos realizados com o método desenvolvido ART-Q. Mais especificamente, foram discutidos três possibilidades de uso do método, com a finalidade de reforçar a afirmativa que o ART-Q foi projetado e desenvolvido para lidar com contextos distintos, ou seja, ser apto a contribuir com análises em quaisquer bases de dados numéricas. O primeiro dos experimentos considerou o uso de uma base de dados sintéticos, com atributos que assumem valores de diferentes comportamentos. Manualmente algumas informações foram inseridas à base de dados para que fosse verificada a capacidade do ART-Q revelá-las. A partir do sucesso deste experimento é possível confirmar a habilidade do método em lidar com valores quantitativos e identificar informações implícitas. O segundo dos experimentos realizados teve seu foco em índices socioeconômicos de países selecionados. Ainda que o contexto selecionado seja de alta complexidade de interpretação e dependa de um especialista para a interpretação de muitas relações entre os intervalos de interesse dos atributos, os resultados deste segundo experimento foram obtidos a partir de uma grande quantidade de combinações de parâmetros e cenários de análises. Os cenários descrevem alternativas para a extração de conhecimento da base de dados e provam que o método ART-Q é flexível quanto à forma como uma análise pode ser conduzida por meio dele. Nas execuções propostas o método apresentou como resultados: conjuntos de intervalos de interesses de acordo com a definição de comportamento de interesse de cada atributo, conjuntos de padrões temporais que envolvem relações intervalares dos períodos de interesse dos atributos e conjuntos de regras de associação, construídas a partir dos padrões identificados. O último experimento conduzido retoma a estratégia do experimento 1 ao colocar o ART-Q à prova da capacidade de identificar informações implícitas na base de dados. Entretanto neste experimento o ART-Q foi submetido ao processamento de uma base de dados real, composta por dados climatológicos. A evolução deste experimento levou, ainda, em consideração informações a respeito da influência do fenômeno El Niño na região sudeste do Brasil. Mais especificamente em cidades do estado de São Paulo. Mais uma vez o método ART-Q foi capaz de revelar informações que especialistas afirmaram sobre o tema. A partir da condução destes três experimentos, o método ART-Q provou atingir o objetivo de sua construção, o de contribuir com a área de mineração de regras de associação, ao incorporar ao processo a temporalidade explícita dos dados e a consideração de atributos que assumem valores quantitativos contínuos.

Capítulo 6

Conclusões e trabalhos futuros

Este capítulo apresenta quais as conclusões observadas ao final do desenvolvimento e utilização do método ART-Q. São descritas quais as contribuições que este trabalho pôde fazer à literatura, especialmente à área de mineração de regras de associação, o que envolve, além da disponibilização do método em si, contribuições literárias e conceituais. Por fim, a seção de trabalhos futuros apresenta algumas vertentes que podem ser exploradas para a continuidade deste trabalho, algumas já em curso. Serve, assim, tanto para indicar as projeções e planos dos autores do ART-Q, quanto como um norte a quem queira contribuir com a mineração de regras de associação temporais que consideram dados quantitativos contínuos.

6.1 Conclusões

O estado da arte demonstra que várias técnicas distintas têm surgido a cada dia com o intuito de contribuir das mais diversas formas à análise e extração de informação das bases de dados. A consideração da temporalidade associada aos dados, por exemplo, é uma área que apresenta crescente atração da atenção de pesquisadores, uma vez que a quantidade de dados armazenados cresce de forma ligeira e os dados podem não representar informações válidas com o passar do tempo.

Dentre os métodos e ferramentas que compartilham das ideologias defendidas neste trabalho, o ART-Q se destaca ao incorporar à busca de conhecimento em bases de dados, um novo tipo de padrão, mais complexo que simplesmente um item (ou *itemset*) que compõe a base de dados de entrada. Este padrão possui semântica que incorpora uma grande quantidade de informações: cada item da base de dados $BDRelT$ é, na verdade, uma relação temporal entre dois intervalos de interesse de um atributo quantitativo con-

tínuo. Interesse que é definido pelo usuário, i.e., uma forma muito flexível de analisar dados contínuos.

Consequentemente, a regra de associação construída pelo método ART-Q incorpora, também, uma grande quantidade de informação além da implicação que a regra traz. Entretanto a complexidade dos padrões e regras resultantes não impede que um usuário com menos domínio do contexto realize análises em dados das mais diversas áreas. Isso se dá pelo fato que o ART-Q provê uma forma simples e intuitiva de visualizar os intervalos de interesse dos atributos. A partir de tal visualização os padrões e as regras de associação podem ser mais facilmente compreendidos e explorados.

A estratégia de validação dos resultados obtidos pelo ART-Q, descrita pela Seção 5.1.3, foi essencial para comprovar que o ART-Q é capaz de revelar informações implícitas nos dados. Foi por meio dela que o método mostrou-se confiável para a condução dos próximos experimentos realizados, que consideram bases de dados reais. No experimento 3, por exemplo, a identificação de informações que compartilham semântica com afirmações de especialistas da área põe em prática tal habilidade do ART-Q, apresentada pela validação dos resultados no experimento 1.

Até o presente momento, nenhum trabalho na literatura constrói informações que se assemelham ao padrão buscado pelo ART-Q. Consequentemente, nenhum trabalho gera regras de associação com semântica parecida. Talvez esta possa ser considerada a maior contribuição que este trabalho provê para a literatura. Aliado à esta afirmação, é importante ressaltar que a busca pelo comportamento de interesse dos atributos é uma inovação, também, realizada de forma flexível e intuitiva, uma vez que na literatura não existem trabalhos que constroem padrões constituídos pela semântica (de importância de valores, neste caso) dos atributos.

Uma comparação direta com qualquer outra estratégia presente na literatura não é possível, pelo simples fato que nenhuma outra estratégia emprega a identificação dos pontos e intervalos de interesse dos atributos. Porém, uma bateria de execuções com dados sintéticos para tecer uma breve comparação com outros algoritmos como o Apriori, em sua versão original e o FP-growth quanto à etapa de identificação dos padrões e construção das regras de associação revelou que os tempos de execução obtidos pelo ART-Q se comportam exatamente iguais ao Apriori em sua versão original e, com o prejuízo esperado de um algoritmo incremental, quando comparado ao FP-growth, um algoritmo da linhagem do crescimento de padrões. Portanto, a etapa de identificação dos pontos de interesse, construção dos intervalos de interesse e indicação das relações da AIA é que

demonstram que a contribuição viabilizada pelo ART-Q é mais interessante que o leve aumento de complexidade que essas tarefas acarretam.

Além disso, o ART-Q foi capaz de responder a questão evidenciada pela hipótese deste trabalho. Pois sim, a consideração da temporalidade de forma explícita e a manutenção dos dados quantitativos contínuos contribui com a descoberta de informações não triviais nos processos de análises por meio da mineração de dados.

Conclui-se, então, que uma nova forma de mineração de dados compõe a gama de possibilidades que a área provê. Esta que é mais robusta ao permitir a fácil identificação de registros faltantes (pela informação temporal explícita), uma vez que não sofre com lacunas temporais nos dados e considera dados quantitativos contínuos, em sua forma mais original, mantendo nuances que poderiam ser omitidas caso processos de discretização dos dados fossem aplicados.

Ainda que exista a carência de um especialista para cada domínio em que o ART-Q foi submetido, o método se mostrou flexível o suficiente para prover várias formas de análise sobre os dados, apresentadas neste trabalho como cenários de análise. Não se mantém, portanto, restrito a apenas um domínio de aplicação.

Por fim, é possível afirmar, então, que o método é flexível e não sensível à transações e dados faltantes, reduz a necessidade da realização de muitas tarefas de pré processamento dos dados, permite definir e identificar comportamentos de interesse dos atributos, visualizar intervalos de interesse e provê uma nova forma de padrão e regra de associação, diferente das que a literatura já contempla.

6.2 Contribuições

Ao atender os objetivos, específicos e geral, que norteiam este trabalho, as seguintes contribuições foram possíveis:

- Desenvolvimento de uma estratégia para a obtenção dos valores do comportamento de interesse para os atributos quantitativos que considera a distribuição normal de probabilidade;
- Desenvolvimento de uma estratégia para a construção de intervalos temporais nos quais os atributos assumem os valores de seus respectivos comportamentos de interesse;

- Desenvolvimento de uma representação dos intervalos temporais de interesse dos atributos;
- Desenvolvimento de uma estratégia para a identificação dos relacionamentos entre os intervalos temporais e suas vizinhanças. Nesta etapa a Álgebra Intervalar de Allen (AIA) foi considerada para descrever as relações temporais;
- Definição e implementação do método ART-Q de mineração de regras de associação envolvendo dados quantitativos contínuos a partir da união dos itens acima descritos.

Além de disponibilizar tal método, ao contemplar todos os objetivos (geral e específicos) acima descritos, este trabalho em nível de doutorado também foi responsável pelas seguintes produções científicas:

- João, R. S. et al. A New Approach to Classify Sugarcane Fields Based on Association Rules. Advances in Intelligent Systems and Computing. 14. ed. Cham: Springer International Publishing, 2018, v. 558, p. 475-483;
- João, R. S.; Ribeiro, M. X. Identifying relational temporal intervals of interests: A new strategy to deal with temporal data. - em processo de submissão;
- João, R. S.; Ribeiro, M. X. Mining temporal association rules on quantitative continuos data - em processo de submissão.

Em mais uma forma de contribuição, o método ART-Q despertou interesse de uma empresa da cidade de Birigui que atua na inclusão da tecnologia nas questões sociais de cidades do Brasil. A proposta do uso do ART-Q nas tarefas realizadas pela empresa é a de revelar informações implícitas de índices sociais e suas relações temporais quando assumem valores em seus respectivos comportamentos de interesse, para visões e cenários distintos.

6.3 Trabalhos Futuros

Naturalmente, ao passo que este trabalho foi conduzido, várias possibilidades de estender o método no futuro foram identificadas. O ART-Q demonstrou ser versátil o suficiente para atuar em diversos contextos e flexível o bastante para proporcionar várias análises, por cenários que descrevem diferentes visões do usuário. Isso inspira a continuidade do trabalho por meio das seguintes possibilidades:

- Incorporar ao processo, medidas que levam em consideração, também, atributos categóricos e/ou de linguagens naturais, como a distância de Hamming;
- Auto adaptação de valores de parâmetros. Desta forma, o ART-Q descartará aqueles parâmetros que levam a grandes quantidades de padrões e regras de associação, para simplificar as análises;
- Prover uma forma de visualização das regras de associação, à semelhança da visualização provida para intervalos de interesses - em desenvolvimento;
- Considerar o uso de estratégias de algoritmos não iterativos, como o FP-growth Han, Pei e Yin (2000), que realizam a busca de padrões por meio da estratégia de crescimento de padrões e não a de geração de candidatos (como no Apriori);
- Incluir, também, as relações inversas da AIA, que não foram contempladas por esta versão do ART-Q;
- Incorporar a definição de comportamentos de interesse que levam em conta outras distribuições de probabilidade, diferentes da normal.

Considerações finais

Este capítulo apresentou as conclusões obtidas a partir do desenvolvimento e teste do método ART-Q e as contribuições que o método desenvolvido neste trabalho pode fazer à área da mineração de dados. Especificamente sobre as contribuições deste trabalho, além da construção e disponibilização de uma nova forma de minerar regras de associação, o trabalho introduziu uma nova estratégia, que permite ao usuário definir quais os valores que mais interessam nos atributos quantitativos contínuos para a análise que deve ser conduzida. Essa permissibilidade flexibiliza a análise dos dados e proporciona uma forma de visualização daqueles que são os momentos de maior revelação de informações nos atributos, os intervalos de interesse. Além disso, por meio da álgebra intervalar de Allen, um novo tipo de padrão pôde ser identificado nos dados, mais complexo por revelar uma quantidade maior de informações que os padrões das estratégias tradicionais na literatura. O que, consequentemente, resultou em regras de associação que são mais ricas em semântica. São apresentadas, também, as contribuições à literatura, por meio da realização do levantamento bibliográfico do estado da arte e a confecção de artigos científicos. Por fim, os trabalhos futuros são apresentados como uma forma de indicar quais as possibilidades de extensão deste trabalho que os autores pretendem investigar e, também, prover um

norte a quem intencione contribuir com a área da mineração de regras de associação temporais que envolvem dados quantitativos contínuos.

Referências

ADHIKARY, D.; ROY, S. Trends in quantitative association rule mining techniques. In: Recent Trends in Information Systems (ReTIS), 2., 2015, Kolkata. **Proceedings...** [S.l.]: IEEE, 2015. p. 126–131.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: International Conference on Very Large Data Bases (VLDB), 20., 1994, Santiago. **Proceedings...** São Francisco: Morgan Kaufmann, 1994. p. 487–499.

AGRAWAL R.; IMIELINSK, T. S. A. Mining association rules between sets of items in large databases. In: International Conference on Management of Data, 1993, Washington. **Proceedings...** Nova Iorque: ACM SIGMOD, 1993. p. 207–216.

AKHLAGH, M. M.; TAN, S. C.; KHAK, F. Temporal data classification and rule extraction using a probabilistic decision tree. In: International Conference on Computer Information Science (ICCIS), 2., 2012, Kuala Lumpur. **Proceedings...** [S.l.]: IEEE, 2012. p. 346–351.

ALE, J. M.; ROSSI, G. H. An approach to discovering temporal association rules. In: ACM Symposium on Applied computing, 2000, Como. **Proceedings...** Nova Iorque: ACM, 2000. p. 294–300.

ALLEN, J. F. An Interval-based Representation of Temporal Knowledge. In: International Joint Conference on Artificial Intelligence (IJCAI'81) - Volume 1, 7., 1981, Vancouver. **Proceedings...** São Francisco: Morgan Kaufmann, 1981. p. 221–226.

ALLEN, J. F. Maintaining knowledge about temporal intervals. **Communications of the ACM**, Nova Iorque, v. 26, n. 11, p. 832–843, nov. 1983.

ALLEN, J. F.; FERGUSON, G. Actions and events in interval temporal logic. **Journal of logic and computation**, Oxford, v. 4, n. 5, p. 531–579, abr. 1994.

ALLEN, J. F.; HAYES, P. J. A common-sense theory of time. In: International Joint Conference on Artificial Intelligence (IJCAI'85) - Volume 1, 9., 1985, Los Angeles. **Proceedings...** São Francisco: Morgan Kaufmann, 1985. p. 528–531.

ALVAREZ, V. P.; VAZQUEZ, J. M. An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization. **Expert Systems with Applications**, v. 39, n. 1, p. 585–593, 2012.

AMARAL, T. **Mineração de Regras de Exceção em Séries Temporais Multivariadas**. 2020 . Tese (Mestrado em Computação) — Instituto de Ciências

Matemáticas e de Computação, Universidade de São Paulo, São Carlos, São Carlos, 2020.

AUMANN, Y.; LINDELL, Y. A statistical theory for quantitative association rules. **Journal of Intelligent Information Systems**, v. 20, n. 3, p. 255–283, 2003.

BANK, T. W. **DataBank: World Development Indicators**. 2020. Disponível em: <https://databank.worldbank.org/source/world-development-indicators>. Acesso em: 12 fev. 2020.

BBC, N. **Aquecimento global: Super El Niño, a perigosa versão do fenômeno climático cada vez mais frequente no Pacífico**. 2019. Disponível em: <https://www.bbc.com/portuguese/geral-50207541>. Acesso em: 10 fev. 2020.

BELLINI, P.; MATTOLINI, R.; NESI, P. Temporal logics for real-time system specification. **ACM Computing Surveys (CSUR)**, Nova Iorque, v. 32, n. 1, p. 12–42, 2000.

BOHLEN, M. H.; BUSATTO, R.; JENSEN, C. S. Point-versus interval-based temporal data models. In: International Conference on Data Engineering, 14., 1998, Orlando. **Proceedings...** Orlando: IEEE, 1998. p. 192–200.

BRUCE, B. C. A model for temporal references and its application in a question answering program. **Artificial intelligence**, v. 3, n. 1, p. 1–25, 1972.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 8. ed. São Paulo: Saraiva, 2013.

CHAMAZI, M. A.; MOTAMENI, H. Finding suitable membership functions for fuzzy temporal mining problems using fuzzy temporal bees method. **Soft Computing**, Salerno, v. 23, p. 3501–3518, maio 2019.

CHAN, K. C.; AU, W.-H. An effective algorithm for mining interesting quantitative association rules. In: ACM symposium on Applied computing, 1997, San Jose. **Proceedings...** Nova Iorque: ACM, 1997. p. 88–90.

CHATFIELD, C. **The analysis of time series**: an introduction. 6. ed. Boca Raton: Chapman e Hall, 2003. (Texts in Statistical Science, 59).

CHEN, C.-H. et al. Mining fuzzy temporal association rules by item lifespans. **Applied Soft Computing**, Amsterdã, v. 41, n. C, p. 265–274, abr. 2016.

CHITTARO, L.; MONTANARI, A. Temporal representation and reasoning in artificial intelligence: Issues and approaches. **Annals of Mathematics and Artificial Intelligence**, Kluwer, v. 28, n. 1-4, p. 47–106, out. 2000.

CORCORAN, A. L.; SEN, S. Using real-valued genetic algorithms to evolve rule sets for classification. In: IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence, 1., 1994, Orlando. **Proceedings...** Orlando: IEEE, 1994. p. 120–124.

DELGADO, M.; OTHERS. Fuzzy association rules: General model and applications. **IEEE Transactions on Fuzzy Systems**, Piscataway, v. 11, n. 2, p. 214–225, abr. 2003.

DEMsAR, J.; CURK, T.; ERJAVEC, A.; GORUP Črt; HOcEVAR, T.; MILUTINOVIc, M.; MOzINA, M.; POLAJNAR, M.; TOPLAK, M.; STARIc, A.; STAJDOHAR, M.; UMEK, L.; ZAGAR, L.; ZBONTAR, J.; ZITNIK, M.; ZUPAN, B. Orange: Data mining toolbox in python. **Journal of Machine Learning Research**, v. 14, p. 2349–2353, 2013. Disponível em: <http://jmlr.org/papers/v14/demsar13a.html>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, Nova Iorque, v. 39, n. 11, p. 27–34, nov. 1996.

FU, T.-C. A review on time series data mining. **Engineering Applications of Artificial Intelligence**, Washington, v. 24, n. 1, p. 164–181, fev. 2011.

FUKUDA, T. et al. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. In: International Conference on Management of Data, 25., 1996, Montreal. **Proceedings...** Nova Iorque: ACM SIGMOD, 1996. p. 13–23.

FURIA, C. A. et al. Modeling time in computing: A taxonomy and a comparative survey. **ACM Computing Surveys (CSUR)**, Nova Iorque, v. 42, n. 2, p. 6–6:59, fev. 2010.

GIUSTI, R. **Classificação de séries temporais utilizando diferentes representações de dados e ensembles**. 2017 . Tese (Doutorado em Computação) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, São Carlos, 2017.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. 1. ed. Boston: Addison-Wesley Longman, 1989.

GUPTA, A.; JING, G.; AGGARWAL, C. C.; HAN, J. Outlier detection for temporal data: A survey. **IEEE Transactions o knowledge and data engineering**, Sydney, v. 26, n. 9, p. 2250 – 2267, set. 2014.

HALDULAKAR, R.; AGRAWAL, J. Optimization of association rule mining through genetic algorithm. **International Journal on Computer Science and Engineering (IJCSE)**, Otteri, v. 3, n. 3, p. 1252–1259, mar. 2011.

HAN, J.; PEI, J.; KAMBER, M. **Data mining**: concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2012.

HAN, J.; PEI, J.; YIN, Y. Mining frequent patterns without candidate generation . In: International conference on Management of data, 2000, Dalas. **Proceedings...** [S.l.]: ACM SIGMOD, 2000. p. 1–12.

HARMS, S. K.; DEOGUN, J. S. Sequential association rule mining with time lags. **Journal of Intelligent Information Systems**, Dordrecht, v. 22, n. 1, p. 7–22, jan. 2004.

HIRANO, S.; TSUMOTO, S. Detection of risk factors as temporal data mining. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008): New Frontiers in Applied Data Mining , 2008, Osaka. **Proceedings...** Berlin: Springer, 2008. p. 143–156.

HOLTE, R. C. Very simple classification rules perform well on most commonly used datasets. **Machine learning**, v. 11, n. 1, p. 63–90, abr. 1993.

HÖPPNER, F. Learning temporal rules from state sequences. In: International Joint Conference on Artificial Intelligence - Workshop on Learning from Temporal and Spatial Data , 17., 2001, Seatle. **Proceedings...** [S.l.], 2001. p. 25–31.

JOãO, R. S.; NICOLETTI, M. C.; MONTEIRO, A. M. Dealing with temporality when inducing association rules from a retail database. In: International Conference on Intelligent Systems Design and Applications (ISDA), 15., 2015, Marrakech. **Proceedings...** [S.l.]: IEEE, 2016. p. 19–24.

JOãO, R. S. et al. A new approach to classify sugarcane fields based on association rules. In: International Conference on New generations (ITNG), 14., 2017, Las Vegas. **Proceedings...** Cham: Springer, 2017. p. 475–483.

KANG, G.-M. et al. Bipartition techniques for quantitative attributes in association rule mining. In: TENCON - IEEE Region Conference, 10., 2009, Singapura. **Proceedings...** Singapura: IEEE, 2009. p. 1–6.

KAYA, M.; ALHAJJ, R. Novel approach to optimize quantitative association rules by employing multi-objective genetic algorithm. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems - IEA/AIE : Innovations in Applied Artificial Intelligence, 18., 2005, Bari. **Proceedings...** Berlin: Springer, 2005. p. 560–562.

KIDA, T.; SAITO, T.; ARIMURA, H. Flexible framework for time-series pattern matching over multi-dimension data stream. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) - New Frontiers in Applied Data Mining: Workshops, 2008, Osaka. **Proceedings...** Berlin: Springer, 2009. p. 1–12.

KIRCHGäSSNER, G.; WOLTERS, J.; HASSLER, U. **Introduction to modern time series analysis**. 2. ed. Berlin: Springer, 2007. (Springer Texts in Business and Economics).

KNIGHT, B.; MA, J. Time representation: A taxonomy of temporal models. **Artificial Intelligence Review**, v. 7, n. 6, p. 401–419, nov. 1993.

KOH, J.-L.; CHOU, P.-M. Incrementally mining recently repeating patterns over data streams. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) - New Frontiers in Applied Data Mining, 2008, Osaka. **Proceedings...** Berlin: Springer, 2009. p. 26–37.

LADKIN, P. B. Time representation: A taxonomy of internal relations. In: AAAI National Conference on Artificial Intelligence , 50., 1986, Filadélfia. **Proceedings...** [S.l.]: Association for the Advancement of Artificial Intelligence (AAAI), 1986. p. 360–366.

LAROSE, D. T. **Discovering Knowledge in Data**: An introduction to data mining. 1. ed. Nova Jersey: Wiley-Interscience, 2004.

LAXMAN, S.; SASTRY, P. S. A survey of temporal data mining. **Sãdhanã**, Nova Delhi, v. 31, n. 2, p. 173–198, abr. 2006.

LEE, C.-H.; CHEN, M.-S.; LIN, C.-R. Progressive partition miner: An efficient algorithm for mining general temporal association rules. **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, v. 15, n. 4, p. 1004–1017, jul. 2003.

LI, D. et al. A new approach of self-adaptive discretization to enhance the apriori quantitative association rule mining. In: International Conference on Intelligent System Design and Engineering Application (ISDEA), 2., 2012, Sania. **Proceedings...** Sanya: IEEE, 2012. p. 44–47.

LI, Y. et al. Generating market basket data with temporal information. In: SIGKDD Workshop on Temporal Data Mining, 1., 2001, São Francisco. **Proceedings...** São Francisco: ACM, 2001.

LIAN, W.; CHEUNG, D. W.; YIU, S. An efficient algorithm for finding dense regions for mining quantitative association rules. **Computers & Mathematics with Applications**, Kidlington, v. 50, n. 3-4, p. 471–490, ago. 2005.

LIN, M.-Y.; LEE, S.-Y. Fast discovery of sequential patterns by memory indexing. **Lecture notes in computer science**, Berlin, n. 2454, p. 150–160, set. 2002.

LIN, W.; ORGUN, M. A.; WILLIAMS, G. J. An overview of temporal data mining. In: The Australasian Data Mining Workshop in conjunction with The 15th Australian Joint Conference on Artificial Intelligence, 1., 2002, Canberra. **Proceedings...** Sidney: University of Technology Sydney, 2002. p. 83–89.

LIU, H.; SETIONO, R. Chi2: feature selection and discretization of numeric attributes. In: International Conference on Tools with Artificial Intelligence, 70., 1995, Herndon. **Proceedings...** Herndon: IEEE, 1995. p. 388–391.

LU, H.; FENG, L.; HAN, J. Beyond intratransaction association analysis: mining multidimensional intertransaction association rules. **ACM Transactions on Information Systems (TOIS)**, Nova Iorque, v. 18, n. 4, p. 423–454, out. 2000.

MANI, I.; PUSTEJOVSKY, J.; SUNDHEIM, B. Introduction to the special issue on temporal information processing. **ACM Transactions on Asian Language Information Processing (TALIP) - Special Issue on Temporal Information Processing**, Nova Iorque, v. 3, n. 1, p. 1–10, mar. 2004.

MARTELLO, A. **Balança comercial registra em 2012 pior desempenho em 10 anos**. 2013. Disponível em: <http://g1.globo.com/economia/noticia/2013/01/balanca-comercial-registra-em-2012-menor-superavit-em-dez-anos.html>. Acesso em: 08 jan. 2020.

MARTIN, D. et al. A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. **IEEE Transactions on Evolutionary Computation**, Hong Kong, v. 18, n. 1, p. 54–69, fev. 2014.

MARTINS, M. E. G. **Introdução à probabilidade e à estatística com complementos de Excel**. Lisboa: Departamento de Estatística e Investigação Operacional da FCUL; Sociedade Portuguesa de Estatística, 2005.

MCCUE, C. **Data mining and predictive analysis**: intelligence gathering and crime analysis. Amsterdã: Butterworth-Heinemann, 2015.

MILLER, R. J.; YANG, Y. Association rules over interval data. In: ACM SIGMOD International Conference on Management of Data, 1997, Tucson. **Proceedings...** Nova Iorque: ACM, 1997. p. 452–461.

MITSA, T. **Temporal Data Mining**. 1. ed. Boca Raton: LLC, 2010.

MOSLEHI, F.; HAERI, A.; MARTíNEZ-ÁLVAREZ, F. A novel hybrid ga–pso framework for mining quantitative associationrules. **Soft Computing**, Salerno, v. 24, p. 4645–4666, jul 2019.

NI, J. et al. Artar: Temporal association rule mining algorithm based on attribute reduction. In: International Conference on Computer Communication and the Internet (ICCCI), 1., 2016, Wuhan. **Proceedings...** Wuhan: IEEE, 2016. p. 350–353.

NICOLETTI, M. C.; LISBOA, F. O. S. d. S.; HRUSCHKA-JUNIOR, E. R. Automatic learning of temporal relations under the closed world assumption. **Fundamenta Informaticae**, Amsterdã, v. 124, n. 1-2, p. 133–151, jan. 2013.

NONATO, R. T.; OLIVEIRA, S. R. de M. Técnicas de mineração de dados para identificação de áreas com cana-de-açúcar em imagens landsat 5. **Engenharia Agrícola**, Jaboticabal, v. 33, n. 6, p. 1268–1280, dez. 2013.

OLSON, D. L.; DELEN, D. **Advanced Data Mining Techniques**. 1. ed. Berlin: Springer, 2008.

ONU. **Relatório do Desenvolvimento Humano 2019**: Além do rendimento, além das médias, além do presente:desigualdades no desenvolvimento humano no século xxi. Nova Iorque: Programa das Nações Unidas para o Desenvolvimento, 2019. Disponível em: <http://hdr.undp.org/sites/default/files/hdr_2019_pt.pd>. Acesso em: 06 mai. 2020.

OZDEN, B.; RAMASWAMY, S.; SILBERSCHATZ, A. Cyclic association rules. In: International Conference on Data Engineering, 14., 1998, Orlando. **Proceedings...** Orlando: IEEE, 1998. p. 412–421.

RIBEIRO, M. X.; TRAINA, A. J.; TRAINA, C. A new algorithm for data discretization and feature selection. In: ACM symposium on Applied computing, 2008, Fortaleza. **Proceedings...** Nova Iorque: ACM, 2008. p. 953–954.

ROMANI, L. A. S. et al. A new time series mining approach applied to multitemporal remote sensing imagery. **IEEE Transactions on Geoscience and Remote Sensing**, Piscataway, v. 51, n. 1, p. 140 – 150, jan. 2013.

RUMSEY, D. **Estatística para leigos**. 8. ed. Boston: Alta books, 2009.

SANTIAGO, A. D.; ROSSETTO, R. **Doenças causadas por fungos**. 2009. Ageitec: Agência Embrapa de Informação Tecnológica <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_79_22122006154841.html>. Acesso em: 08 ago. 2017.

SANTIAGO, A. D.; ROSSETTO, R. **Doenças causadas por vírus**. 2009. Ageitec: Agência Embrapa de Informação Tecnológica <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_80_22122006154841.html>. Acesso em: 08 ago. 2017.

SCHUSTER, A. On the periodicities of sunspots. **Philosophical Transactions of the Royal Society of London. Mathematical, Physical and Engineering Sciences**, v. 206, n. 402-412, p. 69–100, jan. 1906.

SRIKANT, R.; AGRAWAL, R. Mining quantitative association rules in large relational tables. In: ACM SIGMOD international conference on Management of data, 1996, Quebec. **Proceedings...** Nova Iorque: ACM, 1996. p. 1–12.

STAM, R. B.; SNODGRASS, R. **A bibliography on temporal databases**. Carolina do Norte: Defense Technical Information Center, 1988.

STEVENS, S. S. **Handbook of experimental psychology**. Nova Iorque: Wiley, 1951. (A Wiley publication in psychology).

STEVENSON, W. J. **Estatística Aplicada à Administração**. São Paulo: Harbra, 1986.

TAN, P.-N.; KUMAR, V.; STEINBACH, M. **Introduction to data mining**. 1. ed. Boston: Pearson Addison-Wesley, 2006.

TELIKANI, A.; GANDOMI, A. H.; SHAHBAHRAMI, A. A survey of evolutionary computation for association rule mining. **Information Sciences**, Alberta, v. 524, p. 318–352, jul 2020.

TRIOLA, M. F. **Introdução à Estatística**. Rio de Janeiro: Livros Técnicos e Científicos, 1998.

WANG, W.; YANG, J.; MUNTZ, R. Tar: temporal association rules on evolving numerical attributes. In: International Conference on Data Engineering, 17., 2001, Heidelberg. **Proceedings...** Heidelberg: IEEE, 2001. p. 283–292.

WIEDERHOLD, G.; FRIES, J. F.; WEYL, S. Structured organization of clinical data bases. In: national computer conference and exposition (AFIPS), 1975, Anaheim. **Proceedings...** Nova Iorque: ACM, 1975. p. 479–485.

WINARKO, E.; RODDICK, J. F. Armada – an algorithm for discovering richer relative temporal association rules from interval-based data. **Data and Knowledge Engineering**, Amsterdã, v. 63, n. 1, p. 76–90, out. 2007.

XUE, Y.-J. et al. Soil quality assessment using weighted fuzzy association rules. **Pedosphere**, Nanquim, v. 20, n. 3, p. 334 – 341, maio 2010.

YANG, G. Mining association rules from data with hybrid attributes based on immune genetic algorithm. In: International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 70., 2010. **Proceedings...** Yantai: IEEE, 2010. p. 1446–1449.

YANG, J.; FENG, Z. An effective algorithm for mining quantitative associations based on subspace clustering. In: International Conference on Networking and Digital Society (ICNDS), 2., 2010, Wenzhou. **Proceedings...** Wenzhou: IEEE, 2010. p. 175–178.

ZAMBONI, A. et al. StArt uma ferramenta computacional de apoio à revisão sistemática. In: Congresso Brasileiro de Software (CBSoft'10) - Sessão de Ferramentas, 17., 2010, Salvador. Anais... [S.l.]: SBC, 2010. p. 91–96.

ZHANG, W. Mining fuzzy quantitative association rules. In: IEEE. International Conference on Tools with Artificial Intelligence, 11., 1999, Chicago. **Proceedings...** Chicago: IEEE, 1999. p. 99–102.

ZHENG, H. et al. Optimized fuzzy association rule mining for quantitative data. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2014, Pequin. **Proceedings...** [S.l.]: IEEE, 2014. p. 396–403.

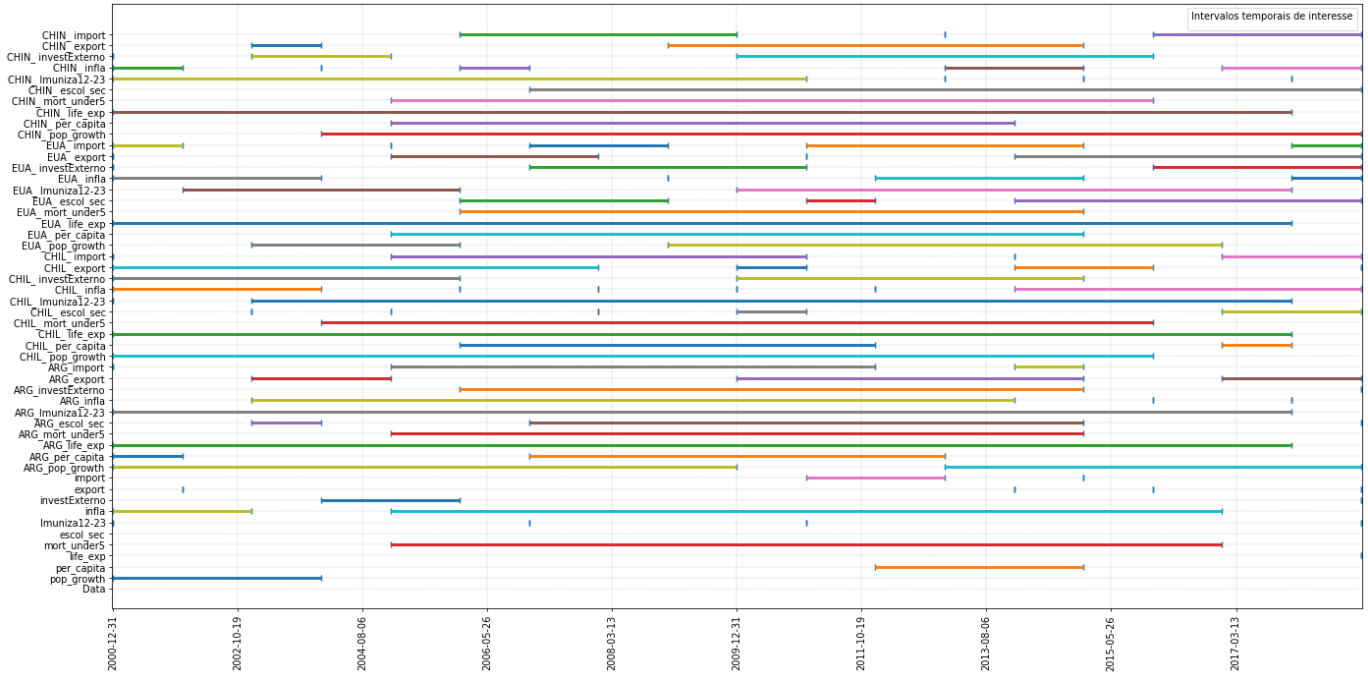
ZMOGINSKI, P. F.; INOHARA, A.; YONG, Z. **Perspectivas do desenvolvimento econômico chinês pós-Covid-19 e impactos para a economia brasileira**. 2020. Inovasia <https://ebook.inovasia.com.br/l/aJBrb0ABF1177>. Acesso em: 20 mar. 2020.

Apêndice A

Intervalos temporais de interesse dos atributos que compõem *BDSocio* para cada cenário que considera uma diferente visão.

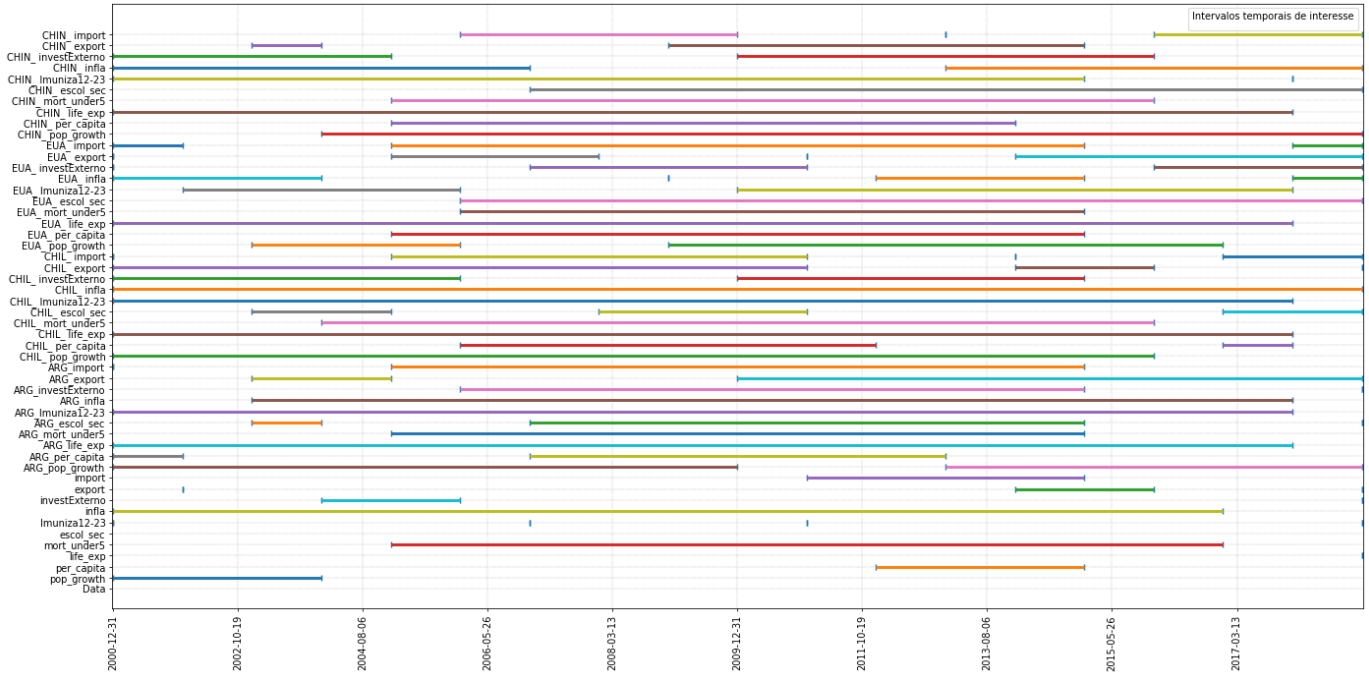
Cenário (a) O Brasil se destaca no mundo pois apresenta valores em seus índices socioeconômicos acima da normalidade, ou seja, apresenta crescimento econômico enquanto os demais países se mantém na normalidade;

Janela temporal para se considerar um intervalo: 370 unidades de tempo (dias);
Janela temporal para se considerar uma relação entre dois intervalos temporais: 750 unidades de tempo (dias).



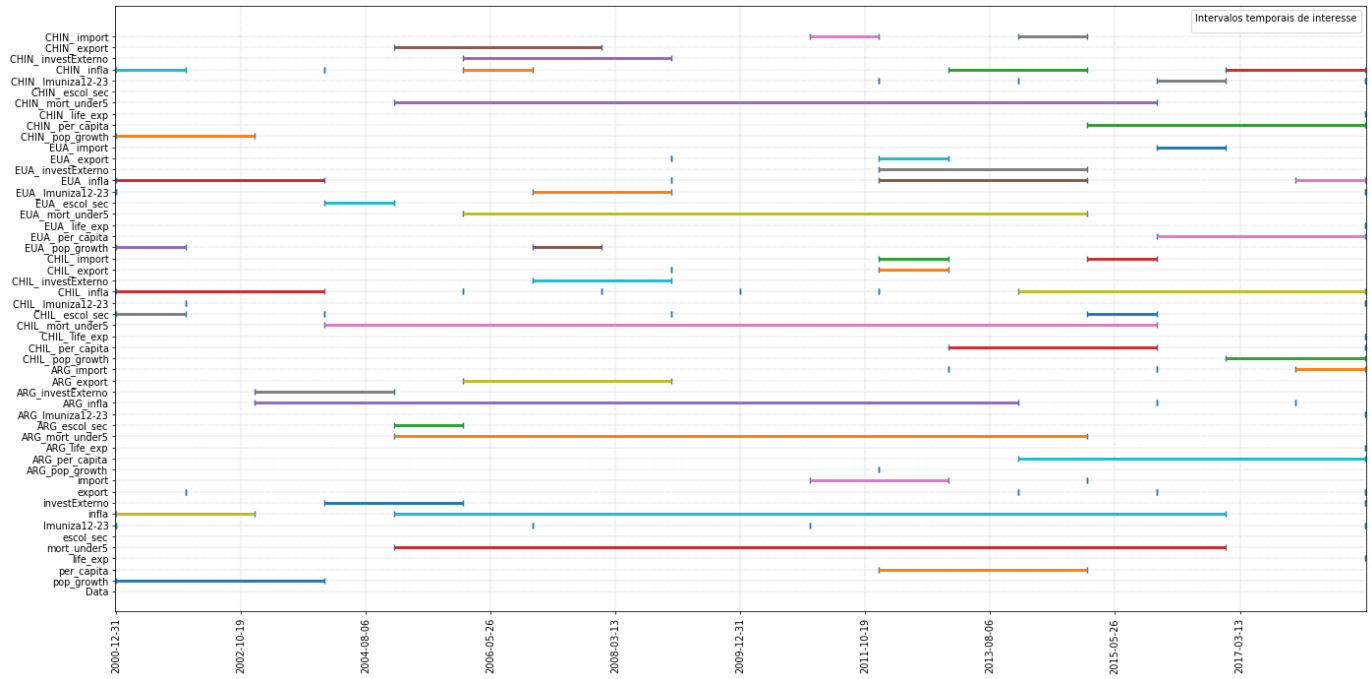
Cenário (a) O Brasil se destaca no mundo pois apresenta valores em seus índices socioeconômicos acima da normalidade, ou seja, apresenta crescimento econômico enquanto os demais países se mantém na normalidade;

Janela temporal para se considerar um intervalo: 750 unidades de tempo (dias);
Janela temporal para se considerar uma relação entre dois intervalos temporais: 750 unidades de tempo (dias).



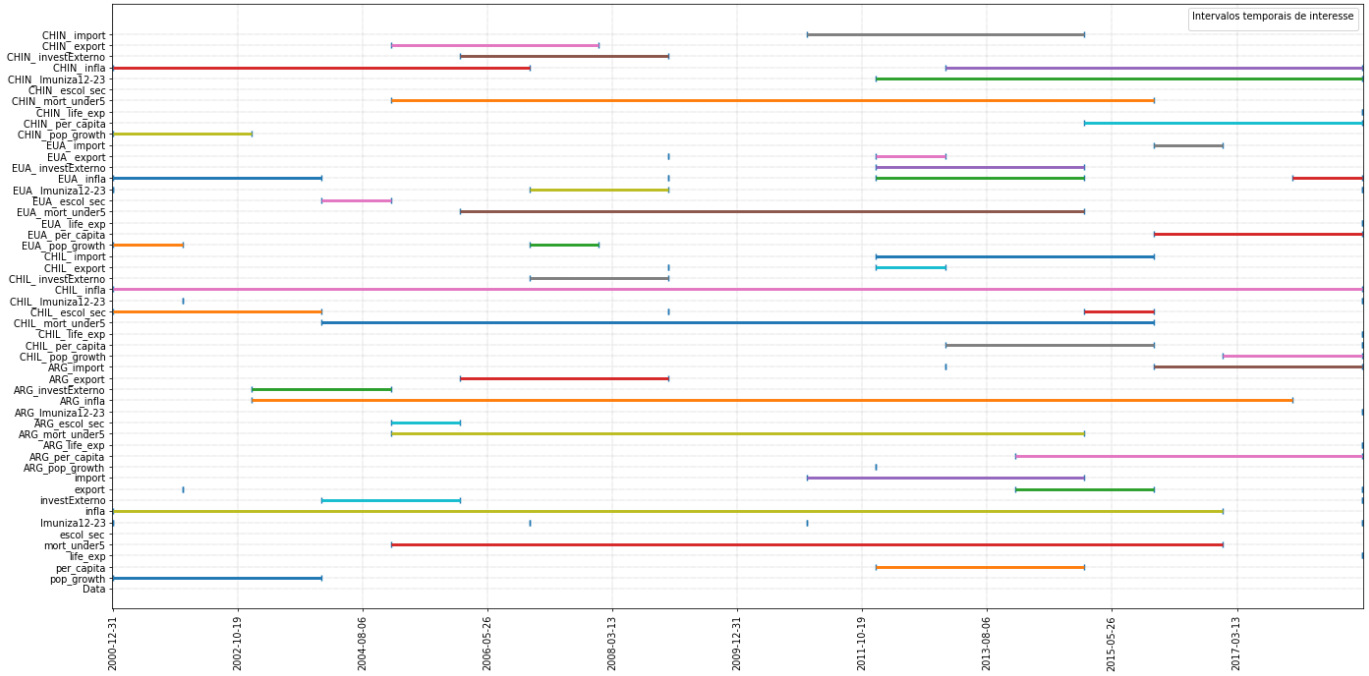
Cenário (b) O Brasil apresenta valores em seus índices socioeconômicos acima da normalidade enquanto os demais países também se encontram na mesma situação - crescimento global.

Janela temporal para se considerar um intervalo: 370 unidades de tempo (dias);
Janela temporal para se considerar uma relação entre dois intervalos temporais: 750 unidades de tempo (dias).



Cenário (b) O Brasil apresenta valores em seus índices socioeconômicos acima da normalidade enquanto os demais países também se encontram na mesma situação - crescimento global.

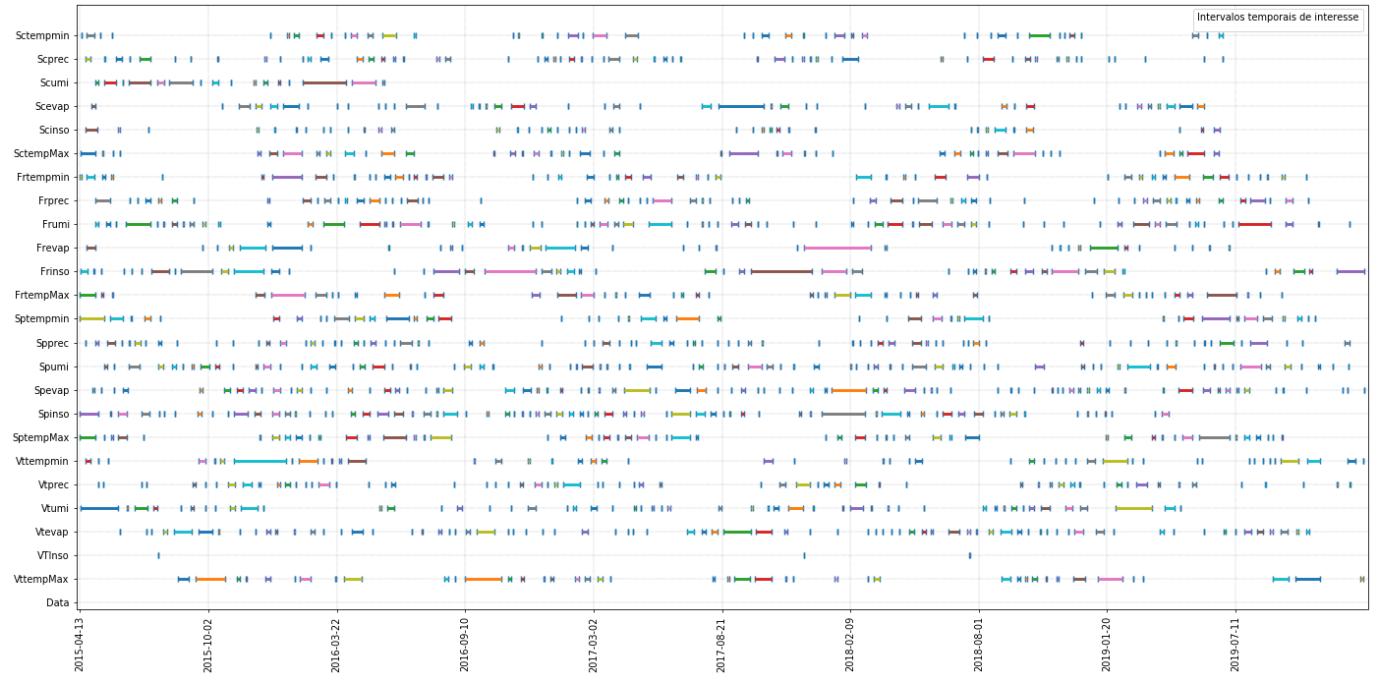
Janela temporal para se considerar um intervalo: 750 unidades de tempo (dias);
Janela temporal para se considerar uma relação entre dois intervalos temporais: 750 unidades de tempo (dias).



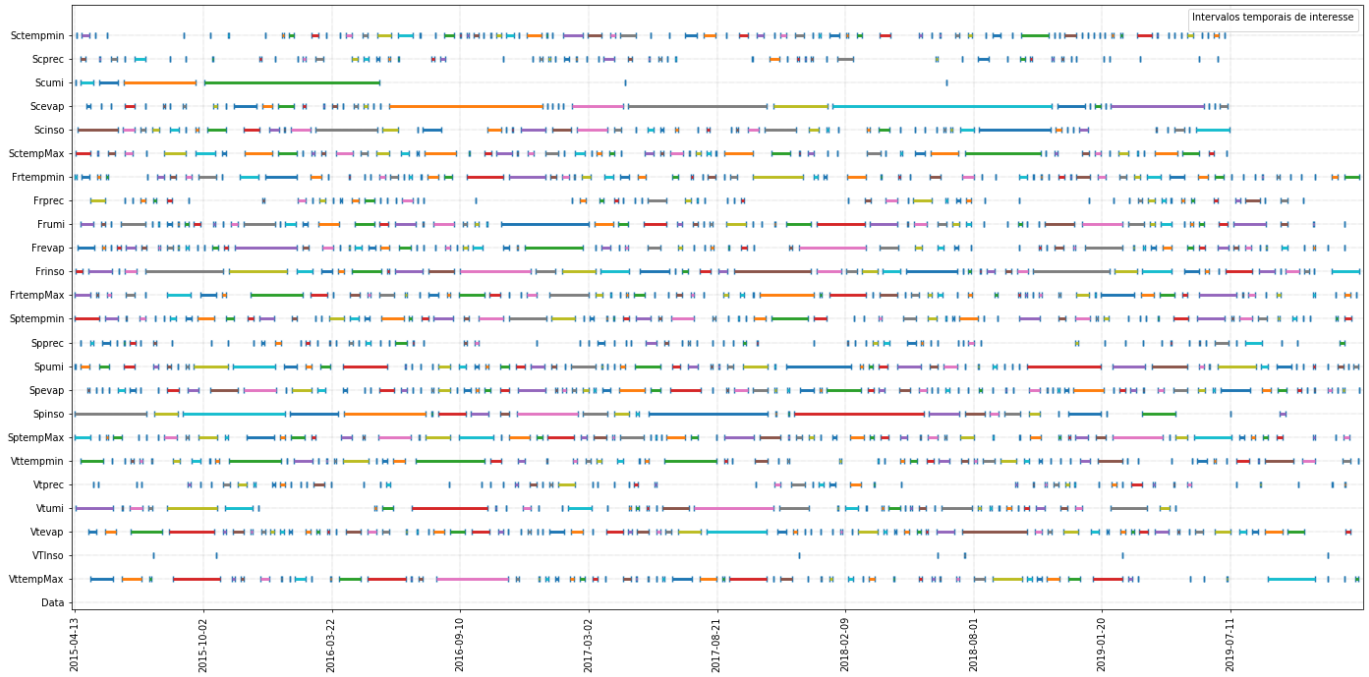
Apêndice B

Intervalos temporais de interesse gerados pelo ART-Q no experimento 3.

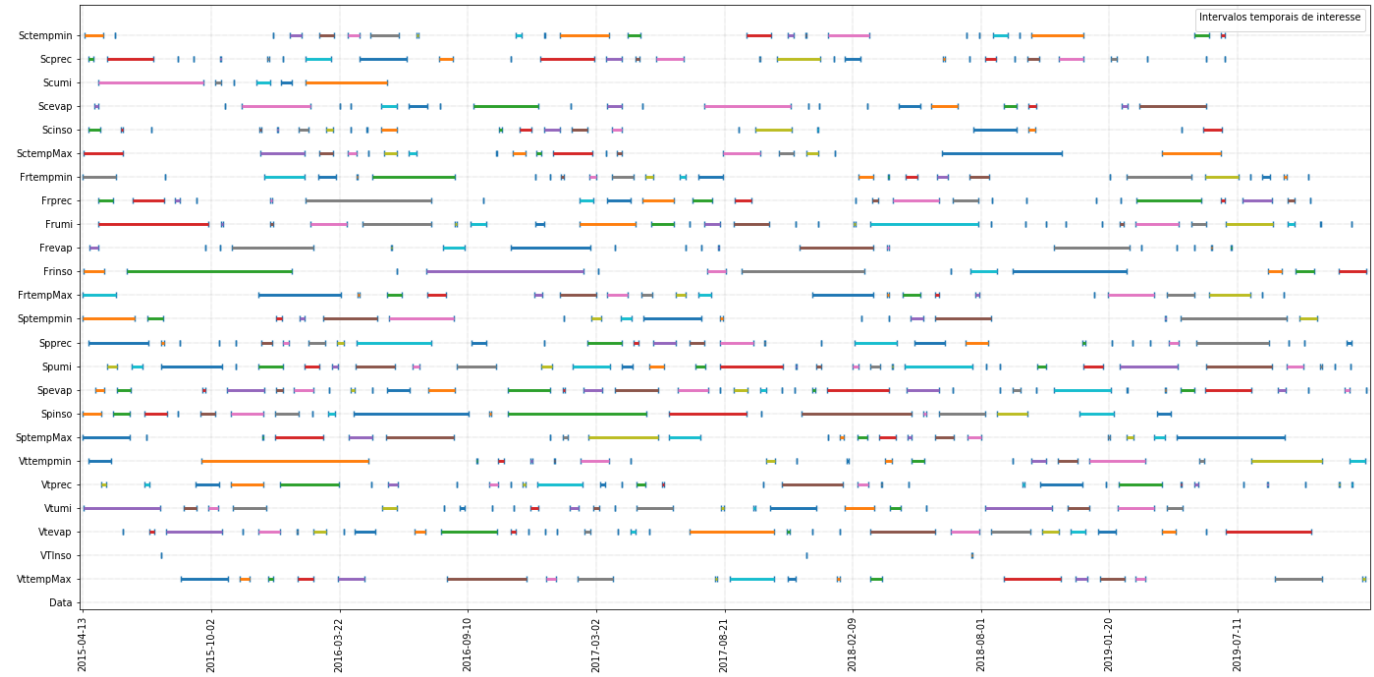
	Comportamento de interesse	Relações identificadas	Registros	Padrões	Regras	MWI	MWR
ex_4	Acima do normal	12388	1235	265	35	7 dias	7 dias



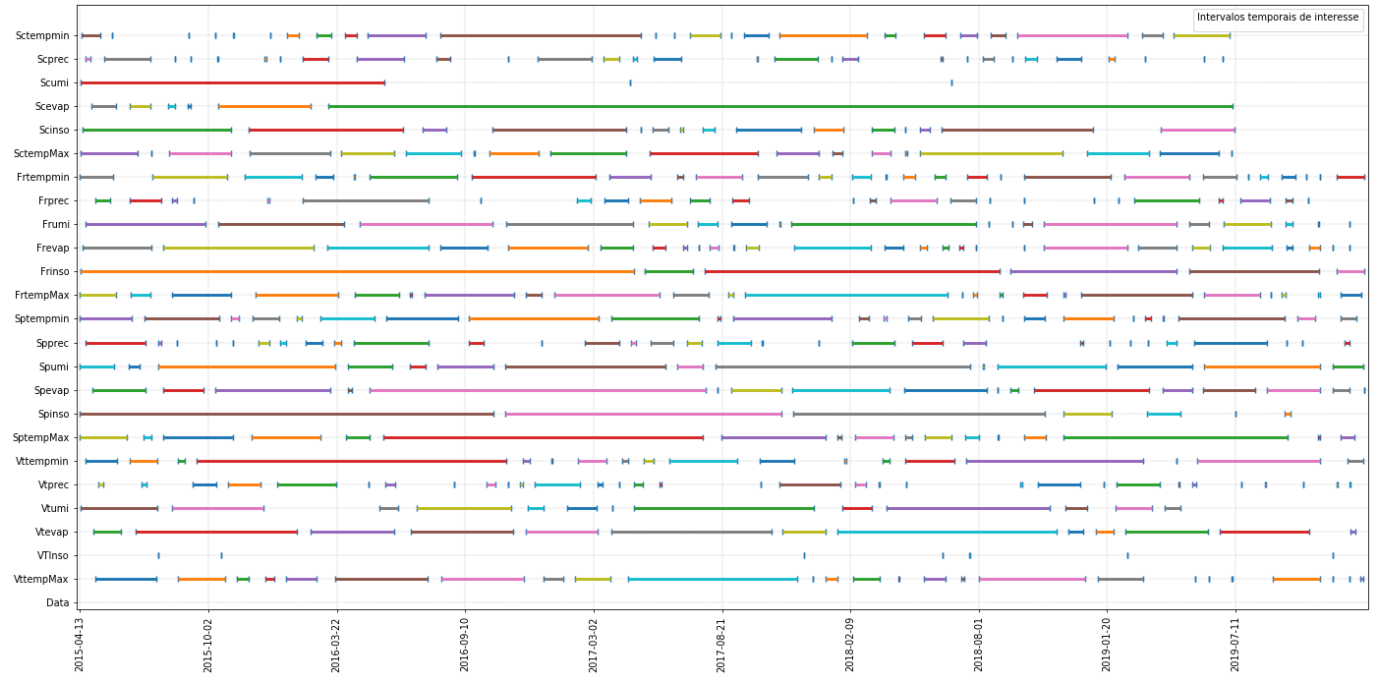
	Comportamento de interesse	Relações identificadas	Registros	Padrões	Regras	MWI	MWR
ex_8	Fora do normal	21882	1401	571	236	7 dias	7 dias



	Comportamento de interesse	Relações identificadas	Registros	Padrões	Regras	MWI	MWR
ex_{12}	Acima do normal	5731	521	205	45	15 dias	7 dias



	Comportamento de interesse	Relações identificadas	Registros	Padrões	Regras	MWI	MWR
ex_{16}	Fora do normal	6912	418	6331	17751	15 dias	7 dias



Apêndice C

Revisão da literatura apoiada pela ferramenta stArt.

A fim de garantir a condução coesa na etapa de identificação dos estudos relacionados ao tema desta pesquisa, optou-se pela utilização de uma ferramenta que auxilia no processo de revisão sistemática, a *StArt (State of the Art through Systematic Review)*, idealizada por Zamboni et al. (2010). Ainda que desenvolvida para a condução de revisões sistemáticas, ferramenta StArt foi considerada, neste trabalho, como facilitadora na identificação, seleção e extração de trabalhos relacionados ao tema da pesquisa, as demais etapas não foram executadas com auxílio da ferramenta, mas sim de forma manual. No total, 4 fontes de informação disponíveis na Internet foram consultadas, sem restrições quanto à data de publicação, para o levantamento inicial dos trabalhos correlatos ao tema desta pesquisa de doutorado, são elas: (1) ACM digital library [1], (2) IEEE [2], (3) Scopus [3] e (4) Science direct [4]. Cada uma das 4 fontes de dados listadas foi consultada por meio de uma mesma *string* de busca elaborada com os termos chave da pesquisa:

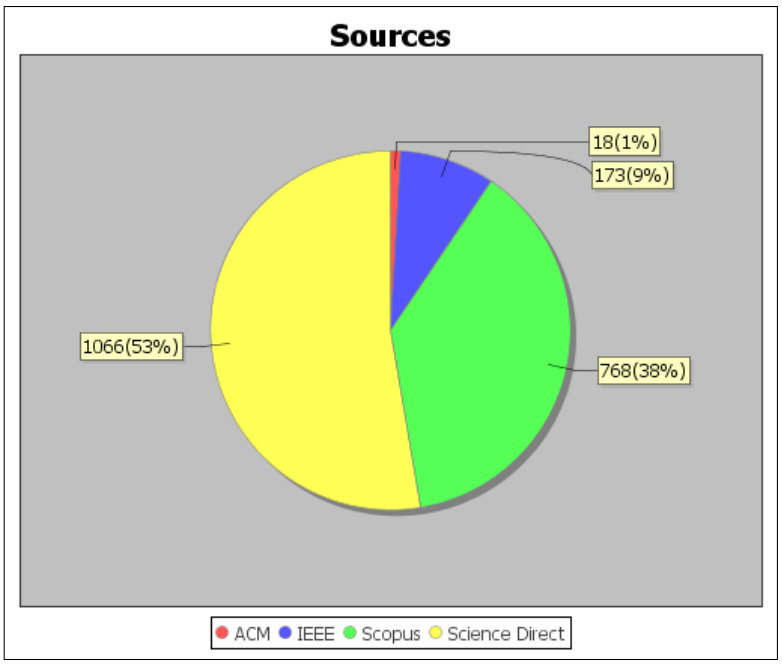
$$\begin{aligned} &\text{"}\textit{temporal association rules}\text{"} \quad OR \quad \text{"}\textit{continuous association rules}\text{"} \; OR \\ &\text{"}\textit{mining temporal association rules}\text{"} \quad OR \quad \text{"}\textit{mining continuous association rules}\text{"} \end{aligned} \tag{C.1}$$

No total, 2025 trabalhos foram encontrados nas consultas em (1), (2), (3) e (4) com a *string* C.1. A Figura C.1 ilustra em um gráfico a quantidade dos trabalhos que foram identificados em cada uma das fontes anteriormente citadas. É possível observar no gráfico

[1] http://dl.acm.org/
[2] http://ieeexplore.ieee.org/Xplore/
[3] https://www.elsevier.com/solutions/scopus
[4] http://www.sciencedirect.com/

que a grande maioria dos trabalhos identificados (53% do total dos resultados obtidos) é proveniente de uma única fonte (4), enquanto (3) é responsável por 38% dos resultados, (2) por 9% e (1) apenas 1% (18 trabalhos).

Figura C.1: Resultado das consultas realizadas para identificação dos trabalhos na literatura que atendem aos termos da *string* C.1.



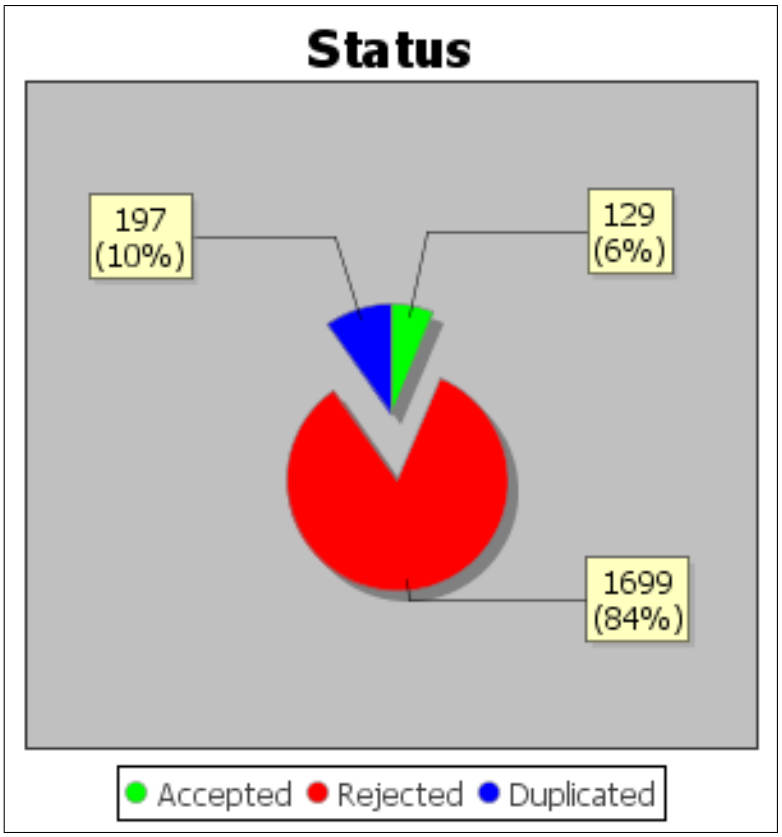
Fonte: Elaborada pelo autor.

Após a identificação dos trabalhos que atendem aos termos descritos pela *string* C.1, a próxima etapa realizada foi a de seleção daqueles que realmente são relacionados ao tema desta pesquisa de doutorado.

A seleção dos trabalhos foi feita por meio da análise de seus respectivos títulos. Para tal, alguns critérios de seleção e exclusão foram definidos, tais como: (1) inclusão: lidar com dados quantitativos contínuos, (2) inclusão: considerar a temporalidade, (3) inclusão: regras de associação temporais, (4) inclusão: regras de associação quantitativas e (5) exclusão: não ter relação ao tema da pesquisa.

A Figura C.2 ilustra o resultado desta etapa de análise. Nela é possível observar que 1699 trabalhos (cerca de 84% do total) foram descartados pela análise dos títulos, enquanto 197 (10%) foram identificados como trabalhos duplicados, i.e., trabalhos que estão presentes nos resultados de consultas em mais do que uma fonte de dados. Apenas 129 trabalhos (cerca de 6% dos 2025 no total) foram selecionados pela análise dos títulos.

Figura C.2: Resultado da seleção de trabalhos pela análise dos títulos. Do total, 129 (6%) foram aceitos; 1699 (84%) foram descartados pela análise e 197 (10%) referem-se a trabalhos duplicados.



Fonte: Elaborada pelo autor.

Em seguida, a etapa de extração foi conduzida, pela qual os resumos de cada um dos 129 trabalhos, selecionados na etapa anterior, foram analisados e aqueles que contemplam a mineração de regras de associação temporais ou a mineração de regras de associação que envolvem dados quantitativos contínuos foram aceitos. Ao final deste processo, 12 trabalhos foram selecionados como os mais correlatos ao tema abordado por este trabalho. Os selecionados descrevem o atual estado da arte na área selecionada.

A seguir, são apresentadas três tabelas (C.1, C.2 e C.3) que resumem os trabalhos acima comentados. Nelas são listados seus respectivos títulos, referências e características importantes que sintetizam o trabalho. Na Tabela C.1 é possível observar a listagem dos trabalhos que lidam com dados quantitativos contínuos (aqueles descritos em maior nível de detalhes na Seção 3.2).

Na Tabela C.2 estão presentes os trabalhosque consideram o aspecto temporal no processo de mineração de regras de associação. Por fim, na Tabela C.3 são listados os trabalhos que abordam as duas temáticas discutidas neste trabalho, i.e., que lidam com

dados quantitativos contínuos e consideram a temporalidade no processo de mineração de regras de associação. Os trabalhos presentes na Tabela C.3 são aqueles que lidam com dados quantitativos contínuos ao mesmo passo que consideram a temporalidade no seu desenvolvimento.

Tabela C.1: Algoritmos que lidam com dados quantitativos contínuos, suas referências e características importantes.

Título/Referência	Características importantes
A Statistical Theory for Quantitative Association Rules por Aumann e Lindell (2003)	• Regras de associação estatísticas; • Um subconjunto interessante: apresenta média e variância de seus valores diferente do restante; • Regras do tipo: (1) apenas atributos quantitativos e (2) atributos categóricos em seu antecedente e no consequente, quantitativo; • Aplica o teste Z para a validação do valor de média; • Avaliação do especialista: de grande importância.
Optimization of Association Rule Mining through Genetic Algorithm por Haldulakar e Agrawal (2011)	• Valores contínuos são convertidos em binários; • Algoritmo genético: seleciona pontos de corte e define intervalos para os atributos contínuos; • Algoritmo Apriori para gerar regras de associação; • Dupla validação das regras: suporte e aptidão - Filtra regras que seriam descartadas; • Teste em base de dados sintética - quantidade pré-definida de regras a serem buscadas; • Resultados comparados ao Apriori original: menor número de regras obtidas com a mesma representatividade.
	Continua

Tabela C.1 Continuação

Título/Referência	Características importantes
An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization por Alvarez e Vazquez (2012)	• Algoritmo evolutivo não supervisionado; • Atributos contínuos são normalizados; • Utiliza função de aptidão para selecionar indivíduos - incorpora mais valores do que apenas o suporte; • Antecedente/consequente das regras escolhidos aleatoriamente e com múltiplos atributos; • Validação quanto a intervalos mais representativos, bases ruidosas, sobreposição de regras, e tamanho das populações - em Várias bases sintéticas e uma real; • Conclusão: algoritmo escalável - Os atributos que não formam regras não são armazenados em memória.
Mining Association Rules from Data with Hybrid Attributes Based on Immune Genetic Algorithm por Yang (2010)	• Algoritmos genéticos - Sistema imunológico para evitar o problema da solução ótima local; • Medidas de aptidão e concentrações dos indivíduos; • Integra a discretização de valores contínuos, redução de atributos e a extração de regras ao mesmo tempo; • Considera os cromossomos compostos por três segmentos - demais trabalhos consideram com dois; • Experimentos realizados em base de dados sintética; • Conclusão: mais veloz que o Apriori - Extrai mais regras que o algoritmo SGA (seu inspirador).
	Continua

Tabela C.1 Conclusão

Título/Referência	Características importantes
Optimized Fuzzy Association Rule Mining for Quantitative Data por Zheng et al. (2014)	• Regras de associação Fuzzy; • Antecedente e consequente da forma: (atributo, pertinência Fuzzy). • Considera medida de confiança, convicção, interesse e fator de certeza. • Iteração de dois níveis: otimiza o conjunto inicial de itens frequentes e regras de associação - refina partições dos conjuntos fuzzy, repetidamente. • Experimentos com 3 bases de dados sintéticas, combinando parâmetros (suporte, confiança e certitude); • Conclusão: OFARM supera seu inspirador quantitativamente e qualitativamente.

Tabela C.2: Algoritmos que lidam com a temporalidade, suas referências e características importantes.

Título/Referência	Características importantes
ARMADA - An algorithm for discovering richer relative temporal association rules from interval-based data por Winarko e Roddick (2007)	• Regras de associação entre estados com intervalos temporais pro Höppner (2001); • Relações entre padrões pela álgebra de Allen; • Visita a base de dados apenas uma vez - copia a base toda para a memória; • Associa padrões identificados à tabela de índices na memória; • Assume temporalidade implícita; • maximum_gap: janela temporal para considerar relações temporais; • Utiliza base de dados sintética para avaliação; • Conclusão: apesar da dependência de memória, executou todas as tarefas propostas.
	Continua

Tabela C.2 Conclusão

Título/Referência	Características importantes
Dealing with temporality when inducing association rules from a retail database por João, Nicoletti e Monteiro (2016)	• Regras de associação temporais para busca de relações com a álgebra de Allen; • Realiza mineração de padrões na memória - apenas 1 varrida na base de dados; • Discretiza os dados e constrói séries temporais; • Considera base de dados real; • Conclusão:o algoritmo é flexível para lidar com dados temporais e não temporais - tempo de execução igual ao seu antecessor (MEMISP) que não incorpora o tratamento do aspecto temporal.
Temporal Data Classification and Rule Extraction Using a Probabilistic Decision Tree por Akhlagh, Tan e Khak (2012)	• Regras construídas por árvores de decisão temporais; • Temporalidade no sequenciamento dos dados (implícita); • Temporalismo: une transações para formar transações de tamanho w - relações inter transações;; • Teorema de Bayes para casos de dúvidas na classificação em folhas - nós folhas com classes multi-valoradas; • Testes com bases de dados sintéticas e reais; • Conclusão: temporalismo é melhor que árvores de decisão convencionais. O algoritmo provou ser útil para classificação não temporal, também. A abordagem Bayesiana melhorou a acurácia.
ARTAR: Temporal Association Rule Mining Algorithm Based on Attribute Reduction por Ni et al. (2016)	• Redução de atributos: Teoria de conjuntos aproximados; • Faz uso da computação paralela; • Mescla o algoritmo Apriori e TFP-growth para identificar padrões que compõem as regras de associação; • Temporalidade é assumida no espaço de busca das regras de associação; • Base de dados sintética; • Compara resultados com o algoritmo PPM e T-Apriori; • Identifica menos regras que ambos comparados em menos tempo; • Conclusão: a redução de atributos foi fundamental.

Tabela C.3: Algoritmos que lidam com dados quantitativos contínuos e a temporalidade, suas referências e características importantes.

Título/Referência	Características importantes
Mining fuzzy temporal association rules by item lifespans por Chen et al. (2016)	• Regras de associação temporais Fuzzy; • Embasamento no algoritmo Apriori - abordagem iterativa de geração de candidatos; • Considera o tempo de duração (lifespan) dos itens – já definidos na base de dados; • Refino das regras pela média de suporte e confiança; • Duas bases de dados sintéticas e duas reais; • Compara com seu sucessor (não contempla intervalos); • Conclusão: foram geradas mais regras fuzzy que seu antecessor, o FAR.
TAR: Temporal Association Rules on Evolving Numerical Attributes por Wang, Yang e Muntz (2001)	• Cada transação da base de dados é tratada como um objeto (ID+atributos); • Evolução temporal: mudanças temporais dos valores de atributos de um dado objeto; • Agrupamento feito pelas distâncias (evolução) entre os atributos; • Regra de associação vista como hipercubo – conjunto de objetos; • Emprega o algoritmo Apriori para mineração de regras; • Compara com outras duas estratégias: SR e LE; • 3 bases de dados sintéticas e 1 real; • Conclusão: Superou em termos de quantidade de regras e em tempo de execução os dois ao qual foi comparado.