

Carlos Roberto Silveira Junior

**Mineração de Regras de Associação
Espaço-Temporais Temáticas Aplicada a
Imagens de Explosões Solares**

São Carlos – Brasil

Agosto de 2018

Carlos Roberto Silveira Junior

Mineração de Regras de Associação Espaço-Temporais Temáticas Aplicada a Imagens de Explosões Solares

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: Engenharia de Software.

Universidade Federal de São Carlos

Departamento de Computação

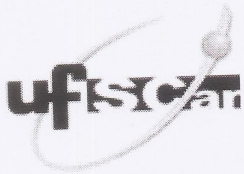
Programa de Pós-Graduação em Ciência da Computação

Orientadora: Profa. Dra. Marcela Xavier Ribeiro

Co-Orientadora: Profa. Dra. Marilde Terezinha Prado Santos

São Carlos – Brasil

Agosto de 2018

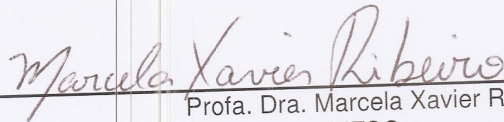


UNIVERSIDADE FEDERAL DE SÃO CARLOS

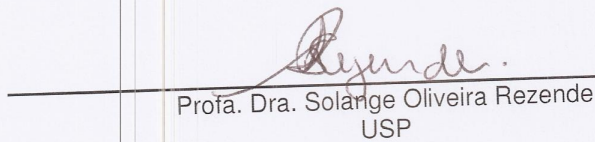
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

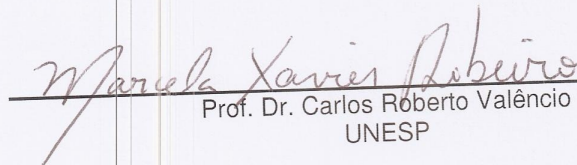
Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Carlos Roberto Silveira Junior, realizada em 14/09/2018:



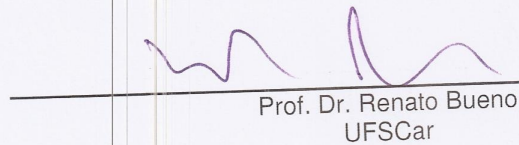
Prof. Dra. Marcela Xavier Ribeiro
UFSCar



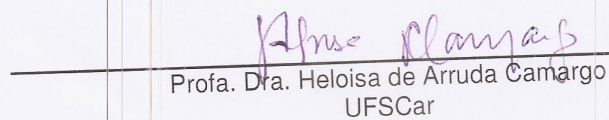
Prof. Dra. Solange Oliveira Rezende
USP



Prof. Dr. Carlos Roberto Valêncio
UNESP

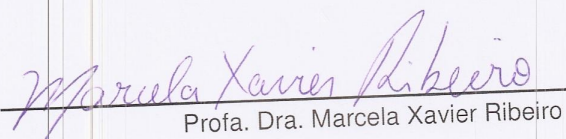


Prof. Dr. Renato Bueno
UFSCar



Prof. Dra. Heloisa de Arruda Camargo
UFSCar

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Carlos Roberto Valêncio e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.



Prof. Dra. Marcela Xavier Ribeiro

Aos meus pais e irmão

Agradecimentos

Agradeço aos meus pais, Carlos Roberto Silveira e Marli Aparecida Gonçalves Silveira, ao meu irmão Lucas Gonçalves Silveira e aos meus avós, Cláudio e Clarisse, Antônio e Laura (*in memoriam*), por todo o apoio e carinho que recebi durante essa jornada.

Agradeço às minhas orientadoras, Marcela Xavier Ribeiro e Marilde Terezinha Prado Santos, pela orientação que recebi e por todo os ensinamentos que me transmitiram.

Agradeço ao José Roberto Cecatto, especialista que norteou o projeto deste trabalho e ajudou durante a avaliação dos resultados.

Agradeço aos membros da banca avaliadora pelo tempo e pela dedicação de ler meu trabalho, é uma alegria tê-los comigo nesse momento.

Agradeço ao Programa de Pós-Graduação em Ciência da Computação por me propiciar essa oportunidade, agradeço aos seus integrantes por todo carinho nesses anos.

Agradeço aos membros Laboratório de Banco de Dados e Engenharia de Software pelo companheirismo.

Agradeço aos amigos e colegas de trabalho e vida pessoal, não os nomearei para não correr o risco de esquecer alguém, mas saibam que todos têm uma parcela neste trabalho.

Agradeço todos meus professores e em especial ao professor Orlando Thomas, parabéns, vocês são os responsáveis por este trabalho.

Por fim, um agradecimento especial a Universidade Federal de São Carlos da qual faço parte desde 2007.

“Begin at the beginning and go on till you come to the end; then stop.”
(Lewis Carrol - “Alice’s Adventures in Wonderland”)

Resumo

Introdução. A análise de clima espacial é uma tarefa complexa que envolve dados espaço-temporais provenientes de imagens de satélite somado a dados de boletins diários. Tais dados são caracterizados como séries temporais de imagens georeferenciadas e séries temporais de dados semânticos (dados alfanuméricos que descrevem as imagens), respectivamente. A mineração de regras de associação pode auxiliar na análise desses dados, como um mecanismo para a revelação de padrões novos e úteis para o especialista de domínio. No entanto, os métodos existentes de mineração de regras de associação espaço-temporais ainda são limitados e, em consequência disso, não atendem adequadamente às expectativas para extração de padrões que relacionam informações espaço-temporais em imagens e dados semânticos. **Objetivo.** Assim sendo, este trabalho tem por objetivo apoiar a análise do clima espacial a partir do desenvolvimento de um método de mineração de regras de associação espaço-temporais que permita relacionar dados solares semânticos e visuais. O foco são séries de imagens solares oriundas de satélites. **Contribuição científica.** Um novo método foi desenvolvido para a extração de padrões significativos de séries de imagens de satélite. Chamado de Solar Miner, esse método é composto por: um novo processo de Extração Transformação Carga de dados -direcionado ao domínio solar- capaz de trabalhar e relacionar dados espaço-temporais com processamento de imagens. Um novo algoritmo de mineração de regras de associação espaço-temporais temático, capaz de trabalhar com esse conjunto de dados em um tempo aceitável. E um novo classificador que utiliza as regras espaço-temporais para determinar o comportamento futuro de novos dados solares. O algoritmo de mineração proposto avança o atual estado da arte da área de mineração de regras de associação por dividir a aplicação das restrições espaço-temporais em duas etapas diferentes do processamento: a aplicação das restrições espaciais é feita durante a extração de *itemsets* frequentes e a aplicação das restrições temporais durante a geração das regras de associação espaço-temporais temáticas. Desta forma, é possível a obtenção de regras que representam a evolução de um determinado conjunto de eventos e como eles se relacionam entre si. Por fim, essas regras são utilizadas pelo classificador associativo que foi proposto neste trabalho para prever o comportamento solar com base em suas características visuais atuais. **Resultados.** O método proposto gerou regras que foram usadas para a classificação, apresentando uma precisão de até 87,3% na classificação de imagens solares, sendo que esse valor de precisão varia com o extrator de características utilizado para representar as imagens. A maior precisão (87,3%) foi obtida utilizando SURF como extrator de características e a menor precisão (82,7%) foi utilizado o Histograma como extrator de características. Os resultados obtidos foram analisados pelo especialista de domínio que avaliou como eficaz e válido o método proposto.

Palavras-chaves: Mineração de Dados; Mineração de Imagens; Regras de Associação Espaço-Temporais; Séries Temporais de Imagens Solares.

Abstract

Introduction. Space weather analysis is a complex task that involves spatiotemporal data from satellite images added to data from daily bulletins. These data are characterized as time series of georeferenced images and time series of semantic data (alphanumeric data describing the images), respectively. The mining of association rules can aid in the analysis of these data as a mechanism for revealing new and useful standards for the domain expert. However, existing spatiotemporal association rules mining methods are still limited and, as a consequence, they do not adequately meet expectations for extracting patterns that relate spatiotemporal information to images and semantic data. **Goal.** Therefore, this work aims to support the analysis of space climate from the development of a method of mining of spatiotemporal association rules that allows to relate solar data semantics and visual. The focus is a series of solar images from satellites. **Scientific contribution.** A new method was developed for extracting significant patterns from satellite imagery series. Called Solar Miner, this method is composed of: a new process of Extraction Transformation Data load - directed to the solar domain - able to work and relate spatiotemporal data with image processing. A new mining algorithm for spatiotemporal association rules, capable of working with this set of data in an acceptable time. And a new classifier that uses the space-time rules to determine the future behavior of new solar data. The proposed mining algorithm advances the current state-of-the-art mining area of association rules by dividing the application of spatiotemporal constraints into two different stages of processing: spatial constraints are applied during the extraction of frequent itemsets and application of temporal constraints during the generation of spatiotemporal association rules. In this way, it is possible to obtain rules that represent the evolution of a given set of events and how they relate to each other. Finally, these rules are used by the associative classifier that was proposed in this work to predict solar behavior based on its current visual characteristics. **Results.** The proposed method generated rules that were used for the classification, presenting a precision of up to 87.3% in the classification of solar images, being that this value of precision varies with the characteristic extractor used to represent the images. The higher precision (87.3%) was obtained using SURF as extraction of characteristics and the less precision (82.7%) was used the Histogram as extractor of characteristics. The results obtained were analyzed by the domain expert who evaluated how effective and valid the proposed method.

Keywords: Data Mining. Image Mining. Spatiotemporal Association Rules. Time Series of Solar Images.

Lista de ilustrações

Figura 1 – Processo de Descoberta de Conhecimento	33
Figura 2 – Árvore Evolutiva de Algoritmos	37
Figura 3 – Diagrama da Mineração Espaço-Temporal	43
Figura 4 – Mineração de Imagens em Alto Nível	51
Figura 5 – Exemplo de Execução do Omega	57
Figura 6 – Formalização do método SolarMiner	70
Figura 7 – Exemplo de imagens solares	71
Figura 8 – Classificação McIntosh visualmente explicada.	75
Figura 9 – Um exemplo de registro de dados solares	76
Figura 10 – Processo SETL	77
Figura 11 – Exemplo de Registro Solar	77
Figura 12 – Vetor de características antes e depois de ser processado pelo Omega.	78
Figura 13 – Exemplo de Registro Solar após a extração de características	79
Figura 14 – Cálculo do Suporte	82
Figura 15 – Cálculo da Confiança	83
Figura 16 – Pré-análise: Geração de Regras Através dos Atributos	93
Figura 17 – Pré-análise: Geração de Sequências Através dos Dados Textuais	95
Figura 18 – Comparativo de desempenho entre os SGBDs	96
Figura 19 – Exemplo de Regras Extraídas da Base Histograma.	97
Figura 20 – Exemplo de Ocorrência de uma Regra.	98
Figura 21 – Regras Extraídas da Base Haralick.	99
Figura 22 – Regras Extraídas da Base SURF.	100
Figura 23 – Classificação para as regras extraídas da Base Histograma.	102
Figura 24 – Classificação para as regras extraídas da Base Haralick.	103
Figura 25 – Classificação para Regras Extraídas da Base SURF.	103
Figura 26 – Exemplo de classificação para novas imagens.	104
Figura 27 – Revisão Sistemática para Regras Espaço-Temporais	135
Figura 28 – Revisão Sistemática para Mineração Aplicada a Imagens	136
Figura 29 – Revisão Sistemática para Mineração Aplicada ao Classificador Associativo	137

Lista de tabelas

Tabela 1 – Exemplo de arquivo descritor da imagem	73
Tabela 2 – Trabalhos Correlatos	88

Lista de Termos

Antecedente: parte A da regra de associação $A \rightarrow B$, que também é chamada de cabeça ou causa;

Atributo temático: atributos de interesse, cujo valor pode ou não ser dependente do tempo e/ou espaço;

Base de dados: o conjunto de dados, armazenado no SGBD, usado nos experimentos;

Confiança da regra de associação tradicional: probabilidade do conseqüente de uma regra acontecer dado que o antecedente aconteceu;

Confiança da regra espaço-temporal (MiTSAI): dado que o antecedente da regra aconteceu, é a probabilidade do conseqüente acontecer após um determinado período de tempo máximo (restrição temporal);

Confiança da regra sumarizada: média do valor de confiança das regras que compõem a regra sumarizada;

Conseqüente: parte B da regra de associação $A \rightarrow B$, que também é chamada de cabeça ou conseqüência;

Dado semântico: dado adicional que descreve uma imagem, composto de valores alfanuméricos;

Itemset: conjunto de itens;

Registro: uma entrada da base de dados, também chamado de tupla, linha ou transação;

Regra sumarizada: uma regra construída a partir de uma ou mais regras que compartilham o mesmo antecedente em comum;

SolarMiner: método de extração de conhecimento proposto neste projeto, composto pelos seguintes métodos: SETL, MiTSAI e SSAC;

Suporte da regra de associação tradicional: frequência da regra em relação aos dados contidos na base de dados;

Suporte da regra espaço-temporal (MiTSAI): frequência com que um conjunto de itens próximos ocorre simultaneamente em imagens da base de dados (ocorrência simultânea obedecendo a restrição espacial);

Suporte da regra sumarizada: média do valor de suporte das regras que compõem a regra sumarizada;

Parte do disco solar: unidade espacial de distância solar. Uma unidade do disco solar consiste em $7,80237 \times 10^3 km$.

Lista de Abreviações

RA: Regra de Associação;

ETL: *Extraction, Transformation and Load*;

MD: Mineração de Dados;

MiTSAI: *Miner of Thematic Spatio-temporal Associations for Images*, minerador de regras de associação proposto neste projeto;

SETL: *Solar ETL*, processo de ETL de dados solares proposto neste projeto;

SITS: Séries Temporais de Imagens de Satélites, do inglês, *Satellite Image Time Series*;

SGBD: Sistema de Gerenciamento de Banco de Dados;

SSAC: *Solar Spatiotemporal Associative Classifier*, classificador associativo proposto neste projeto;

Sumário

1	INTRODUÇÃO	25
1.1	Objetivos, Hipóteses e Justificativas	25
1.2	Métodos	27
1.2.1	Caracterização deste Projeto de Pesquisa	27
1.2.2	Métodos Usados para Implementação	27
1.2.3	Método <i>Goals, Question e Metrics</i>	28
1.3	Organização do Trabalho	28
1.4	Considerações Finais	29
I	REFERENCIAL TEÓRICO	31
2	MINERAÇÃO DE DADOS E REGRAS DE ASSOCIAÇÃO ESPAÇO-TEMPORAIS	33
2.1	Mineração de Dados	34
2.2	Regra de Associação, Padrão Sequencial e Regra de Associação Temporal	35
2.2.1	Algoritmo <i>Incremental Miner of Stretchy Time Sequences</i>	38
2.2.2	Algoritmo <i>ARMADA</i>	40
2.3	Regras de Associação Espaço-Temporais	41
2.4	Estado da Arte em Mineração de Dados Espaço-Temporal	45
2.5	Considerações Finais	48
3	MINERAÇÃO DE DADOS APLICADA A IMAGENS	51
3.1	Pré-Processamento de Imagens	52
3.2	Extração de Características de Imagens	52
3.2.1	Descritor de Haralick	54
3.2.2	Extração de Características Baseado em Formas por meio do Algoritmo <i>Speeded Up Robust Features</i>	55
3.2.3	Algoritmo Omega	56
3.3	Estado da Arte em Mineração de Imagens	58
3.4	Considerações Finais	60
4	CLASSIFICADOR ASSOCIATIVO PARA REGRAS DE ASSOCIAÇÃO	61
4.1	Formalização da Classificação Associativa	62
4.2	Classificador Associativo Baseado em Voto	62

4.3	Estado da Arte em Classificação Associativa para Regras de Associação	63
4.4	Considerações Finais	65
II	DESENVOLVIMENTO	67
5	O MÉTODO PROPOSTO - <i>SOLARMINER</i>	69
5.1	Formalização da Proposta	69
5.1.1	Caracterização da Fonte dos Dados	71
5.1.2	O processo SETL	74
5.1.3	<i>Miner of Thematic Spatio-temporal Associations for Images</i>	80
5.1.4	O <i>Solar Spatiotemporal Associative Classifier</i>	84
5.1.5	Comparação do Método proposto com os Trabalhos Correlatos	88
5.2	Considerações Finais	89
6	EXPERIMENTOS	91
6.1	A Base de Dados Usada	91
6.2	O Especialista de Domínio	92
6.3	Experimentos Iniciais	93
6.3.1	Experimento Inicial para a Extração de Regras de Associação	93
6.3.2	Experimento Inicial para a Extração de Padrões Sequenciais	94
6.4	Experimentos com o Processo SETL	95
6.5	Experimentos com MiTSAI	97
6.6	Experimentos com o SSAC	101
6.7	Avaliação com o Método <i>Goal, Question, and Metrics</i>	105
6.8	Considerações Finais	107
III	FINALIZAÇÃO	109
7	CONCLUSÃO	111
7.1	Avaliação dos Resultados	112
7.2	Principais Contribuições	112
7.2.1	Principais Pontos Inovadores	112
7.2.2	Lista de Publicações	113
7.3	Trabalhos Futuros	114
7.4	Considerações Finais	115
	REFERÊNCIAS	117
A	REVISÃO SISTEMÁTICA	133

A.1	O Método de Revisão Sistemática	133
A.2	Regras de Associação Espaço-Temporais	134
A.3	Mineração de Dados em Imagens	135
A.4	Classificador Associativo	136

1 Introdução

Séries Temporais de Imagens de Satélites (SITS) possuem uma grande quantidade de informação implícita cuja análise pode trazer padrões interessantes para o entendimento do clima espacial (BOULILA *et al.*, 2011). O processamento de tais dados pode ainda auxiliar na predição de tempestades solares de alta intensidade, também nomeadas de explosões solares. Dentre os efeitos de uma tempestade solar de alta intensidade destacam-se: distúrbios temporários no campo eletromagnético da Terra e falhas em satélites artificiais podendo, com isso, acarretar problemas nos sistemas de telecomunicações e navegação; problemas em redes de energia elétrica e, em consequência, ocorrências de *blackouts*.

Ainda assim, a revisão sistemática da literatura indica que há uma lacuna de trabalhos que tem como alvo esse domínio. A mineração de dados aplicada aos SITS é uma tarefa ainda pouco explorada na literatura, conforme afirmado por Vatsavai *et al.* (2012). Em decorrência deste fato, existe uma carência no desenvolvimento de técnicas de mineração de séries de imagens que seja apropriada para a análise de clima espacial e para o monitoramento da atividade solar, que é foco desta pesquisa.

Esse domínio impõe desafios para as tarefas de extração de conhecimento que exigem atenção multidisciplinar. Por exemplo, a mineração de dados de SITS envolve o processamento de imagens, o pré-processamento de dados alfanuméricos, a mineração de imagens, e a mineração de dados espaço-temporais propriamente dita. Ao se trabalhar com SITS solares, a mineração de dados enfrenta um desafio adicional: aproveitar adequadamente outras informações que se integram ao conjunto de imagens, tais como, nível de radiação, clima espacial, intensidade da atividade solar etc.

A motivação para a realização deste trabalho é o auxílio ao entendimento de fenômenos solares envolvendo o clima espacial.

No presente capítulo, está assim organizado: na Seção 1.1, são apresentados os objetivos, hipóteses e justificativas deste trabalho de pesquisa; na Seção 1.2, são apresentados os métodos usados para a realização da pesquisa bibliográfica e validação da proposta; na Seção 1.3, é apresentada a organização desta monografia, e; na Seção 1.4, são apresentadas as considerações finais deste capítulo.

1.1 Objetivos, Hipóteses e Justificativas

O objetivo geral deste trabalho de pesquisa é oferecer subsídios para a análise e o entendimento do clima espacial. Como há uma lacuna da literatura de domínio de trabalhos que abordam a predição do comportamento solar, esta trabalho de pesquisa tem

como objetivo endereçar alguns desses pontos para assim oferecer subsídios a análise desse domínio.

Os objetivos específicos deste trabalho são:

- Projetar, desenvolver e validar um processo de ETL (Extração, Transformação e Carregamento) para a coleta dos dados considerados relevantes para a mineração de dados solares neste trabalho;
- Oferecer um método para extrair regras de associação espaço-temporais temáticas de séries de imagens de satélite;
- Projetar, desenvolver e validar um método de sumarização das regras espaço-temporais temáticas de séries de imagens de satélites.

O objetivo geral é sustentado pelas hipóteses:

- É possível a extração de regras de associação espaço-temporais temáticas a partir de dados solares. Esta hipótese é sustentada pela seguinte justificativa de que os dados solares são multi-disciplinares que possuem características espaço-temporais que demandam pré-processamentos para viabilizar a mineração. Além disso, não foi possível encontrar na literatura de domínio trabalhos que endereçam tanto características espaço-temporais quanto características visuais para processamento de séries de imagens de satélites.
- É possível dar suporte a previsão do clima espacial a partir dos dados solares. Esta hipótese é sustentada pela seguinte justificativa: a partir dos conjuntos de regras espaço-temporais temáticas é possível a geração de modelos preditivos visando dar suporte a previsão do clima espacial. Na literatura de domínio são encontradas classificadores associativos baseados em votos que geram modelos preditivos para conjunto de regras de associação não espaço-temporais. Desta forma, é possível um classificador que considere também essas características desse domínio.

As regras de associação espaço-temporais temáticas são padrões que retratam tanto a relação entre eventos que estão ocorrendo ao mesmo tempo quanto a evolução desses eventos e suas relações espaciais.

Ao se trabalhar com as séries de imagens solares, a ocorrência de uma mancha solar pode influenciar na ocorrência e/ou evolução de outras manchas solares; as regras de associação espaço-temporais temáticas podem auxiliar na identificação deste comportamento.

1.2 Métodos

Nesta seção são apresentados os métodos utilizados para guiar este trabalho de pesquisa. Na Subseção 1.2.1, é apresentada a definição da pesquisa para este trabalho. Na Subseção 1.2.2, são apresentadas as técnicas que foram usadas no desenvolvimento do método. O método *Goals, Question e Metrics*, conforme apresentado na Subseção 1.2.3, foi utilizado para estruturar o planejamento da avaliação para a validação do método desenvolvido neste trabalho de doutorado.

1.2.1 Caracterização deste Projeto de Pesquisa

Este trabalho de doutorado caracteriza-se por ser uma pesquisa experimental, focado na mineração de regras de associação espaço-temporais temáticas aplicada a Séries Temporais de Imagens de Satélite, no qual são realizados experimentos para a validação do cumprimento dos objetivos específicos. As conclusões têm por base os resultados dos experimentos, sendo assim considerado um trabalho com características empíricas, de acordo com Wazlawick (2014).

A execução dos experimentos permite a manipulação das variáveis de entrada do método aqui proposto. A manipulação dessas variáveis foi acompanhada pelo especialista de domínio. O especialista também foi o responsável pela análise dos padrões obtidos com o processo de mineração e, conseqüentemente, a determinação do grau de sucesso no cumprimento do objetivo geral.

Outros aspectos, detalhados na Seção 6.7, foram medidos para viabilizar a comparação com outros métodos e abordagens correlatas a esse trabalho.

1.2.2 Métodos Usados para Implementação

A implementação seguiu o método *Test Driven Development* em que se assegura a qualidade do protótipo implementado. A arquitetura da implementação seguiu *SOLID Principles* e boas práticas de código-limpo apresentadas em Martin (2009).

O controle de versão foi feito por meio do GIT e o código foi armazenado no BitBucket ¹ com acesso apenas pelo autor. O trabalho também está disponível com acesso público no GitHub do autor ².

As tecnologias que foram utilizadas são: Java e Python. A linguagem Python foi escolhida devido ao grande volume de *frameworks* que facilitam a manipulação de dados não estruturados. A linguagem Java foi escolhida devido à existência de bibliotecas de

¹ BitBucket do autor: <<https://bitbucket.org/carlossilveirajr/mitsai>>, último acesso 9 de fevereiro de 2019

² GitHub do autor: <<https://github.org/carlossilveirajr/mitsai>>, último acesso 9 de fevereiro de 2019

processamento de imagens e a possibilidade de adaptação do código para ecossistemas distribuídos com facilidade e tênues alterações nos algoritmos.

O protótipo gerado é independente do Sistema Gerenciador de Banco de Dados (SGBD), assim para armazenar as imagens e seus dados espaço-temporais foram adotados: o MongoDB, Cassandra, PostGIS e MariaDB; um dos experimentos deste trabalho de pesquisa tem por objetivo definir qual SGBD apresenta o melhor desempenho nesse domínio. SGBDs *NoSQL*, como é o caso do MongoDB, são flexíveis e viabilizam a manipulação de dados complexos, no geral, são *open source* e escaláveis. Já SGBDs relacionais, como é o caso do PostGIS e MariaDB, garantem a consistência e possuem boa performance.

1.2.3 Método *Goals, Question e Metrics*

O método *Goals, Question and Metrics* (GQM) foi desenvolvido por Caldiera e Rombach (1994). Seu diferencial é que pode ser aplicado para vários tipos de avaliações distintas, desde qualidade de software a revisões sistemáticas de trabalhos acadêmicos (SOLLINGEN *et al.*, 2002).

O GQM divide-se em três partes:

Goal (Objetivo): propósito da aplicação;

Questions (Questões): questões levantadas para a verificação do cumprimento do objetivo, e;

Metrics (Métricas): utilizadas para responder às questões verificando o cumprimento do objetivo.

Geralmente, esta estrutura aparece em diagramas nos quais o primeiro nível é o objetivo do trabalho, no segundo nível estão as questões que verificam o cumprimento do objetivo e, no terceiro nível, as métricas que são utilizadas para responder às questões. As questões relacionam-se com o objetivo e com as métricas; todas as questões estão relacionadas ao objetivo e todas as métricas relacionam-se com, pelo menos, uma questão. A avaliação ocorre verificando-se ao término do trabalho se todas as questões foram respondidas satisfatoriamente. A instanciação deste método para este trabalho de pesquisa é apresentada na Seção 6.7.

1.3 Organização do Trabalho

Esse trabalho está organizado da seguinte maneira:

Capítulo 1: foram apresentados os objetivos, motivações e considerações a respeito desse trabalho;

Capítulo 2: é apresentado o processo de mineração de dados com foco na extração de regras de associação e em dados espaço-temporais;

Capítulo 3: é apresentado o processo de mineração de imagens, em especial, imagens oriundas de satélites;

Capítulo 4: é apresentado o processo de classificação associativa com ênfase na mineração de imagens de satélites;

Capítulo 5: é apresentado o método proposto neste trabalho;

Capítulo 6: são apresentados os experimentos realizados com o método proposto, e;

Capítulo 7: são apresentadas as conclusões deste projeto de pesquisa.

Este documento também possui o seguinte material complementar:

Apêndice A: é apresentada a execução do processo de revisão sistemática aplicada sobre os assuntos principais deste trabalho.

1.4 Considerações Finais

Neste capítulo introdutório foram apresentados a motivação e o objetivo deste trabalho. Além disso, foram discutidos os métodos utilizados para a realização da pesquisa e também a metodologia de desenvolvimento da mesma. Por fim, foi apresentada a organização geral dos capítulos e os principais assuntos abordados nos mesmos.

Parte I

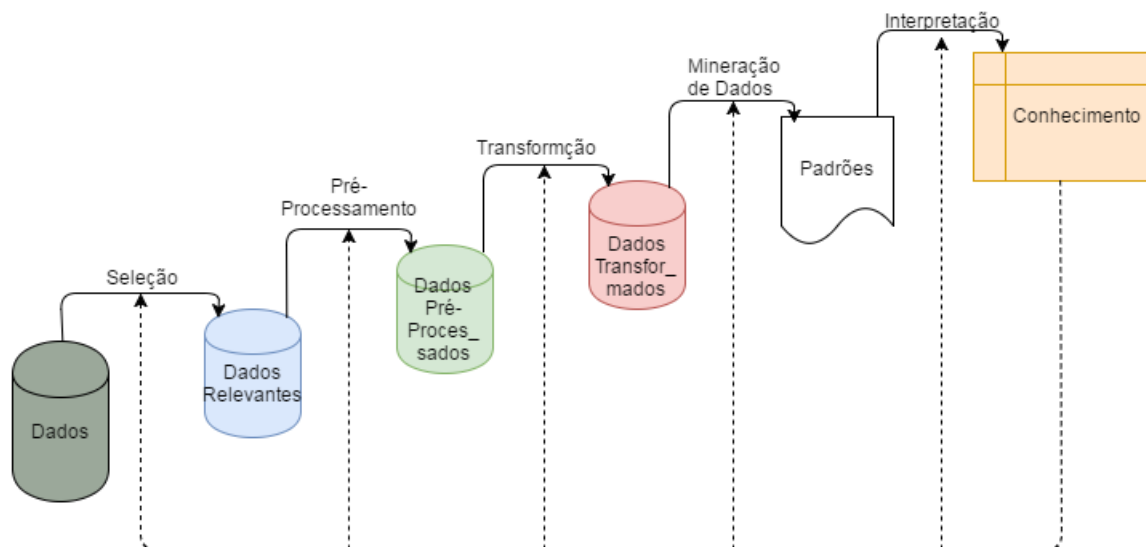
Referencial Teórico

2 Mineração de Dados e Regras de Associação Espaço-Temporais

Devido ao aumento na capacidade de armazenamento de dados, processos para automatizar a análise vêm ganhando relevância tanto no meio acadêmico quanto no meio comercial. A mineração de dados é uma alternativa que visa facilitar a análise de grandes conjuntos de dados por meio da extração de padrões relevantes.

Em Lu, Setiono e Liu (1995), a frase “dados ricos, porém conhecimento pobre” é citada para descrever o problema que o processo de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases*, KDD) aborda: desde a publicação deste trabalho, o grande acúmulo de dados já acarretava dificuldades de análises, consultas e manipulações.

Figura 1 – As cinco etapas do processo de Descoberta de Conhecimento em Bases de Dados (KDD).



Fonte: figura adaptada de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O processo de KDD foi introduzido por Fayyad, Piatetsky-Shapiro e Smyth (1996) e consiste na extração do conhecimento útil de um grande conjunto de dados. O processo de KDD é dividido em cinco etapas, como ilustrado na Figura 1: a Seleção tem como objetivo selecionar apenas os dados que são realmente relevantes para a análise pretendida; o Pré-Processamento visa a correção de inconsistências e eliminação de ruídos, nesta etapa é comum a aplicação de técnicas de limpeza de dados. A etapa Transformação tem como objetivo preparar os dados para servirem de entrada para o algoritmo minerador; em

seguida, a Mineração de Dados (MD) é aplicada para a extração dos padrões no montante de dados. Por fim, faz-se a interpretação destes padrões obtendo o conhecimento útil, a essa etapa dá-se o nome de Interpretação.

Desta forma, o KDD pode ser definido como um processo de extração de padrões novos, válidos, potencialmente úteis e compreensíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Este capítulo é dedicado à mineração de dados com foco na extração de regras de associação aplicada a dados espaço-temporais e está organizado da seguinte maneira na seção 2.1, é apresentada uma introdução sobre a etapa de Mineração de Dados do processo de KDD; na seção 2.2, é discutida a tarefa de Mineração de Dados relativa a extração de regras de associação tradicionais e regras de associação temporais; na seção 2.3, é apresentada a extração de regras de associação espaço-temporais; na seção 2.4, os trabalhos que trazem o estado da arte em mineração de dados espaço-temporais são apresentados, e; na seção 2.5, são apresentadas as considerações finais.

2.1 Mineração de Dados

A Mineração de Dados (MD) é uma das etapas do processo de KDD sendo responsável pela busca de padrões contidos nos dados. Assim sendo, a MD pode ser considerada como o núcleo deste processo e possui implicações diretas em seus objetivos (ELMASRI; NAVATHE, 2005; HAN; KAMBER, 2006; GALVÃO; MARIN, 2009; ABAR *et al.*, 2016; D'AMATO *et al.*, 2016; ANTONIE; ZAIANE; HOLTE, 2016).

Uma tarefa de MD pode ser de *predição*, quando o sistema encontra padrões que visam prever o comportamento de uma certa entidade baseando-se em seu histórico, e *descrição*, quando objetiva encontrar padrões que descrevem os dados contidos na base.

O foco deste trabalho é a mineração de **regras de associação** que consiste em um tipo de padrão descritivo do tipo antecedente e consequente (detalhado na seção 2.2). Porém, existem outros tipos de tarefas de mineração que são importantes:

Classificação é uma técnica de Aprendizado Supervisionado; consiste em criar um modelo de aprendizado a partir de um conjunto de instâncias de treino, cuja classe é previamente conhecida. O modelo de aprendizado é então usado para determinar a classe de novas instâncias cuja a classe é desconhecida. Como exemplo deste tipo de tarefa temos as árvores de decisões.

Agrupamento é uma técnica de *Aprendizado Não Supervisionado*. Consiste em agrupar instâncias de tal forma que as mais similares sejam colocados no mesmo grupo. O algoritmo mais utilizado, segundo Xindong *et al.* (2010), é o *K-means*.

Padrões Sequenciais representam um comportamento comum dos dados ao longo do tempo (HAN; PEI; YAN, 2005; SRINIVASAN; BHATIA; CHAKRAVARTHY, 2006). Devido à sua importância no contexto desse trabalho, são detalhados na seção 2.2.

2.2 Regra de Associação, Padrão Sequencial e Regra de Associação Temporal

Regras de Associação (RAs) são padrões do tipo antecedente e consequente que aparecem na seguinte forma:

$$\{Antecedente\} \implies \{Consequente\}$$

sendo que a intersecção dos conjuntos *Antecedente* e *Consequente* é vazia (HAN; KAMBER; PEI, 2012; MENG; SHYU, 2011; MOHAN; REVESZ, 2014), formalmente:

$$Antecedente \cap Consequente = \emptyset$$

Em alguns trabalhos da literatura o *antecedente* da regra é chamado de *corpo* ou *causa*, e o *consequente* é chamado de *cabeça* ou *consequência*.

Uma RA significa que ao acontecer o *Antecedente*, o *Consequente* provavelmente acontecerá.

O *Antecedente* e *Consequente* de uma regra são formados por *itemsets* (conjunto de itens). Um *itemset* nunca pode ser vazio.

Existem duas métricas relacionadas às RAs que refletem a sua relevância e frequência, suporte, cuja fórmula é apresentada na Equação 2.1; e, confiança, cuja fórmula é apresentada na Equação 2.2. Essas duas formulas utilizadas por diversos diversos mineração de regras de associação no atual estado da arte como Shu, Guo e Zhang (2009), Hammami, Turki e Faiz (2012), Lee, Cai e Lee (2014), Lo, Ding e Nazeri (2014), Raheja e Rajan (2012).

Seja D uma base de dados, $|D|$ o número de tuplas em D e uma RA $r : A \rightarrow B$:

$$suporte(r) = \frac{|Tuplas\ nas\ quais\ A \cup B\ aparecem|}{|D|} \quad (2.1)$$

$$confiança(r) = \frac{|Tuplas\ nas\ quais\ A \cup B\ aparecem|}{|Tuplas\ nas\ quais\ apenas\ A\ aparece|} \quad (2.2)$$

Geralmente, são definidos valores mínimos para essas métricas (*minSup* e *minConf*) e o padrão só é considerado **frequente** se apresentar pelo menos o mínimo definido para

o suporte e só é considerado **forte** se apresentar o mínimo definido pela confiança (ABUL; GAKASE, 2012; FISCH *et al.*, 2012; GUIL; MARIN, 2012; LI *et al.*, 2016; MEHTA *et al.*, 2016; TAYAL; RAVI, 2016).

Existem duas estratégias consolidadas para a extração de RAs na literatura: a estratégia **geração-e-teste de candidatos** e a estratégia **crescimento de padrões**. Essas estratégias são geralmente representadas pelos algoritmos que as empregam: algoritmos da Família Apriori (AGRAWAL; SRIKANT, 1994) representam a estratégia **geração-e-teste de candidatos**, e algoritmos da Família FP-Growth (HAN; PEI; YIN, 2000) representam a estratégia **crescimento de padrões**.

Os algoritmos baseados no Apriori geralmente dividem seu processamento em duas etapas: (i) geração de *itemsets* candidatos, por meio de operações de combinações: Os itens e *itemsets* frequentes são combinados gerando *itemsets* com um maior número de itens, nomeados *itemsets* candidatos; (ii) teste de *itemsets* candidatos: os *itemsets* gerados em (i) têm seus suportes calculados e os frequentes são mantidos. Após (ii), os *itemsets* frequentes realimentam a etapa (i). Essa iteração do algoritmo continua até que nenhum candidato gerado seja frequente (SRIKANT; AGRAWAL, 1995; AGRAWAL; SRIKANT, 1994; COFFI; MARSALA; MUSEUX, 2012). Com o conjunto de *itemsets* frequentes, as RAs são encontradas pelo cálculo do valor de confiança para cada combinação interna de todos os *-sub-itemsets* de cada *itemset* frequente (YANG; LIAO; ZHANG, 2013; CHENG; WANG, 2015).

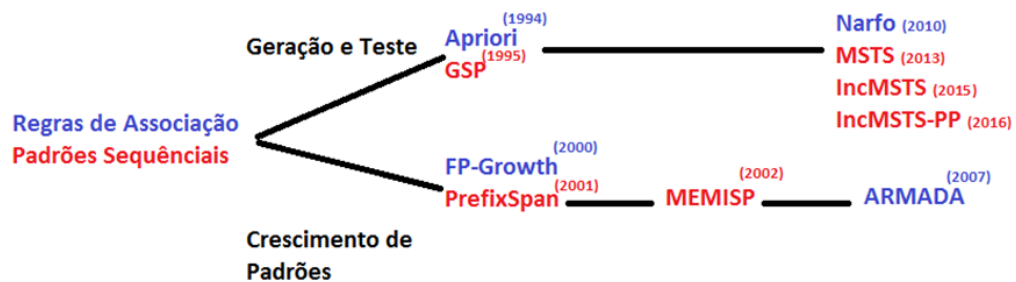
Os algoritmos baseados em FP-Growth (HAN; PEI; YIN, 2000; HAN; PEI; YAN, 2005) aplicam uma abordagem diferenciada para a redução do tempo de processamento em relação ao Apriori (PATHANIA; SINGH, 2015). A execução do FP-Growth tem as seguintes etapas: (i) o algoritmo conta o número de ocorrências de cada item e o armazena em uma tabela; (ii) o algoritmo utiliza essa tabela para a construção da árvore *FP-tree*, sendo que os itens não frequentes são descartados e os frequentes são comprimidos para perto da raiz, dessa forma, uma versão compactada da base de dados é formada; (iii) a árvore é recursivamente processada e novas árvores são criadas a partir de cada agrupamento, dessa forma, o espaço de busca é reduzido; (iv) o processo recursivo continua até não ser mais possível gerar candidatos frequentes, e; (v) através da utilização dos *itemset* frequentes, as regras são criadas. Desta forma, pela redução do espaço de busca a cada iteração, os algoritmos da família FP-Growth alcançam bons desempenhos.

No contexto de dados temporal a busca de padrões frequentes compreende a busca de padrões de sequências e não regras de associação. Uma sequência, ou **padrão sequencial**, é definida pela ocorrência de eventos temporalmente ordenados, por exemplo, seja $s = \langle i_1 \dots i_n \rangle$ para $n \geq 2$ e $i_1 \dots i_n$, eventos não necessariamente distintos, s é uma sequência, se e somente se, i_{j-1} anteceder i_j para todos valores de $j \in [2; n]$. Exemplos de algoritmos: GSP (SRIKANT; AGRAWAL, 1995; SRIKANT; AGRAWAL,

1996), PrefixSpan (HAN *et al.*, 2001), MSTs (SILVEIRA-JUNIOR; SANTOS; RIBEIRO, 2013) e IncMSTs (SILVEIRA-JUNIOR *et al.*, 2015).

Na Figura 2, é ilustrada uma árvore evolutiva de algoritmos de mineração de dados tanto para extratores de padrões sequenciais quanto para extratores de regras de associação. A razão de aparecerem juntos é que ambas as tarefas são bastante relacionadas como fica evidenciado pelos trabalhos de Fournier-Viger *et al.* (2012), Radhakrishna, Kumar e Janaki (2015), Refonaa, Lakshmi e Vivek (2015), Sharma e Bhatia (2016), Shekhar *et al.* (2018). Um ponto que evidencia a relação entre essas tarefas de mineração de dados é o fato de que suas estratégias para a extração de padrões são basicamente as mesmas: estratégia *geração-e-teste de candidatos* com o Apriori para a extração de regras de associação e o GSP para a extração de padrões sequenciais; estratégia de *crescimento de padrões* com o FP-Growth para a extração de regras de associação e PrefixSpan para a extração de padrões sequenciais.

Figura 2 – Árvore Evolutiva de Algoritmos para extração de Regras de Associação e Padrões Sequenciais. A árvore chega ao algoritmo ARMADA, o mesmo citado na seção 2.



Fonte: o próprio autor.

As **Regras de Associação Temporais** são sequências de dois eventos ou mais eventos que apresentam associação de causalidade ao longo do tempo. Assim, tem-se que os algoritmos extratores de Regras de Associação Temporais derivam dos algoritmos de extração de padrões sequenciais. Essa situação acontece com o algoritmo ARMADA proposto por Winarko e Roddick (2007). ARMADA é um algoritmo para a extração de regras de associação temporais que se baseia no algoritmo *MEMory Indexing for Sequential Pattern mining* (MEMISP), de Lin e Lee (2002), para a extração de padrões sequenciais. Sendo que o MEMISP é um algoritmo que tem por base o PrefixSpan.

Tanto o ARMADA quanto o MEMISP necessitam de apenas uma varredura na base de dados e também não demandam a geração de *itemset* candidatos nem a projeção da base de dados. A evolução dos algoritmos até o ARMADA é apresentada na Figura 2.

Pelo fato do ARMADA ser um algoritmo cujo desempenho é alto em relação aos seus antecessores, há outros trabalhos que o utilizam como base, por exemplo os trabalhos

de Moskovitch e Shahar (2009), Schluter e Conrad (2010), Martínez-Ballesteros *et al.* (2011), Pisón *et al.* (2012) e Nguyen e Vo (2015).

Os algoritmos IncMSTS e ARMADA são relevantes no contexto deste trabalho de pesquisa, portanto, são detalhados na Subseção 2.2.1 e na Subseção 2.2.2, respectivamente.

2.2.1 Algoritmo *Incremental Miner of Stretchy Time Sequences*

O algoritmo *Incremental Miner of Stretchy Time Sequences* (IncMSTS), que se baseia no *Miner of Stretchy Time Sequences* (MSTS) de Silveira-Junior, Santos e Ribeiro (2013), realiza a extração incremental de Padrões Temporais Elásticos (*Stretchy Time Patterns* –STP). STP é um padrão que apresenta lacunas temporais entre seus eventos e é definido pela Equação 2.3, na qual $i_1 \dots i_n$ são *itemsets* e $\Delta t_1 \dots \Delta t_{n-1}$ são lacunas temporais. Para cada ocorrência *oc* de uma sequência *s*, o total de lacunas temporais não pode ser maior que μ , configurado pelo usuário, $[\sum_{k=1}^{n-1} \Delta t_k^{oc}] \leq \mu$.

$$s = \langle i_1 \Delta t_1 i_2 \dots \Delta t_{n-1} i_n \rangle \quad (2.3)$$

O IncMSTS, apresentado no Algoritmo 1, recebe como entrada: (i) base de dados incremental *db*, é a base de dados completa, se é a primeira vez que está sendo processada, ou apenas um incremento de dados, caso contrário; (ii) o valor de suporte mínimo, *minSup*; (iii) o máximo de lacunas temporais, μ , que define o quão esparsos os padrões podem ser; (iv) δ valor que define padrões semi-frequentes, e; (v and vi) antigos padrões frequentes e semi-frequentes, *fs* e *sfs*, se não for a primeira vez que a base está sendo processada.

O IncMSTS encontra os *itemsets* frequentes no incremento da base de dados (linha 1) e, então, os antigos padrões frequentes e semi-frequentes são reconstruídos (linha 2). A reconstrução dos padrões antigos consiste em marcar a ocorrência desses padrões no incremento dos dados *db*, esse passo não é executado caso seja a primeira vez em que a base está sendo processada. O incremento dos dados *db* é minerado para encontrar novos padrões (linhas 3 até 5), se é a primeira vez que a base está sendo processada esse passo não é executado. Então, os antigos padrões (*fs* e *sfs*), que foram reconstruídos usando o incremento da base de dados, são minerados (linha 6 até 15). Para minerar os padrões antigos, o suporte de cada padrão é recalculado (linha 7), neste processo, as ocorrências antigas e as novas são consideradas.

Então, se o novo suporte for maior que ou igual a $minSup \times \delta$ ($0 \geq \delta \geq 1$), o padrão é considerado ao menos semi-frequente (linha 8 até 14), se o suporte for maior que ou igual a *minSup*, o padrão é frequente e generalizado (linha 10), senão o padrão é acionado ao conjunto dos semi-frequentes (linha 12). δ é um parâmetro definido pelo usuário cujo objetivo é determinar a frequência mínima de um *itemset* semi-frequente, seu valor pode variar de zero a um; o quão mais próximo a um, menos *itemsets* semi-frequentes

são minerados reduzindo a performance da mineração de incrementos; o quão mais próximo a zero, mais *itemsets* frequentes são gerados causando um aumento de padrões minerados.

Algoritmo 1: O algoritmo *Incremental Miner of Stretchy Time Sequences*.

Entrada: Conjunto incremental de dados db , $minSup$, μ , δ , conjunto de sequências frequentes fs , conjunto de sequências semi-frequentes sfs .

Saída: Conjunto de padrões sequenciais C , conjunto de padrões semi-frequentes.

- 1 $C \leftarrow$ padrões sequencias de com um único itemset formado pelos $\{itemset\ frequentes\} \in db$;
- 2 $reconstrói(fs \cup sfs, db)$;
- 3 **para cada** padrão sequencial $p \in C$ **faça**
- 4 | $C \leftarrow C \cup generalize(p, db, minSup, \mu, \delta)$;
- 5 **fim**
- 6 **para cada** padrão sequencial $p \in fs \cup sfs$ **faça**
- 7 | $sup_p \leftarrow \frac{númeroDeOcorrencias(p)}{Total\ de\ sequências}$;
- 8 | **se** $sup_p \geq minSup \times \delta$ **então**
- 9 | | **se** $sup_p \geq minSup$ **então**
- 10 | | | $C \leftarrow C \cup generalize(p, db, minSup, \mu, \delta)$;
- 11 | | **senão**
- 12 | | | $adicionaAosSemiFrequentes(p)$;
- 13 | | **fim**
- 14 | **fim**
- 15 **fim**

Fonte: algoritmo extraído de Silveira-Junior *et al.* (2015).

A função *generalize* é chamada duas vezes, na linhas 4 e 10, e é apresentada no Algoritmo 2. As entradas da função são: um padrão p que será generalizado, a base de dados db , valor de suporte mínimo $minSup$, o valor máximo de lacunas temporais μ e o valor δ . As saídas da função são: conjunto de padrões frequentes e conjunto de padrões semi-frequentes baseados no padrão p .

A função *generalize* encontra os *itemsets* frequentes na base de dados db (linha 1). Então o algoritmo combina o padrão p com cada *itemset* frequente (linha 3 até 14). O suporte para combinação é calculado (linha 5) e as combinações são classificadas em semi-frequente (linha 10), frequente (linha 8) ou não-frequente (descartado).

As combinações frequentes são adicionadas à *Result* e em outra iteração, linha 3, elas são recombinadas com os *itemsets* frequentes, desta forma, sequências maiores são criadas. Assim, todos os padrões com o prefixo padrão p são gerados. A função *generalize* termina quando não é mais possível gerar padrões maiores que sejam frequentes.

Fonte: algoritmo extraído de Silveira-Junior *et al.* (2015).

A função *verificaOcorrência*, chamada no Algoritmo 2 na linha 5, implementa o *Stretchy Time Windows* (STW). Esse método visa encontrar os padrões esparsos em uma base de dados com lacunas temporais. O método STW usa o parâmetro μ que define o

Algoritmo 2: A Função *Generalize*.

Entrada: Padrão Sequencial p , base de dados db , $minSup$, μ , δ .
Saída: Conjunto de padrões frequentes e semi-frequentes, *Resultado*, derivados de p .

```

1 Itemsets  $\leftarrow \{itemsets\ frequentes\} \in db$  ;
2 Resultado  $\leftarrow \{p\}$  ;
3 para cada padrão  $p' \in$  Resultado faça
4   para cada itemset  $\iota \in$  Itemsets faça
5      $sup_p \leftarrow \frac{verificaOcorrência(p', \iota, \mu)}{Total\ de\ sequências}$  ;
6     se  $sup_p \geq minSup \times \delta$  então
7       se  $sup_p \geq minSup$  então
8          $Resultado \leftarrow Resultado \cup \{p' \cup \iota\}$  ;
9       senão
10         $adicionaAosSemiFrequentes(p' \cup \iota)$  ;
11      fim
12    fim
13  fim
14 fim

```

máximo de lacunas temporais que o padrão pode conter.

O algoritmo IncMSTS possui uma complexidade quadrática, $\Theta(n^2)$, porém, diferente do MSTS e GSP, IncMSTS reaproveita a informação já processada.

2.2.2 Algoritmo ARMADA

O algoritmo ARMADA foi proposto por Winarko e Roddick (2007) e é apresentado pelo Algoritmo 3. Considere como entrada do ARMADA uma base de dados temporal, $D = \{t_1 \dots t_n\}$ sendo cada registro t_i composto por: ID do cliente, atributo temporal com os valores de *início* e *fim*, sendo que $início < fim$ e o período entre *início* e *fim* marca o intervalo em que o registro t_i é considerado válido. Seja um conjunto de estados possíveis denotado por S para os dados em D , tal que, um estado $s \in S$ cuja duração é $[início, fim)$ é denotado pela tupla $(início, s, fim)$. Por exemplo, em uma base médica o resultado de um exame é válido para um determinado período de tempo para um paciente, assim, o paciente é representado pelo ID do cliente, os atributos temporais, o período de validade do exame e o valor do exame. Nesse exemplo o valor do exame é um estado. Se s é único em S então s é um padrão temporal, denotado por $\langle s \rangle$.

O primeiro passo do algoritmo ARMADA (linha 1) consiste em ler a base de dados d e trazê-la para memória, isso é referido como *MDB*. Enquanto lê a base de dados, o algoritmo faz a contagem do suporte para cada estado, e então encontra os estados frequentes. *Estado* define o valor de uma entidade da base de dados em um determinado período de tempo.

O segundo passo do algoritmo é a criação do conjunto de índices, mostrado na

linha 4. Seja p_0 um padrão formado pela combinação do prefixo p com um estado frequente s . Seja $p_0.indice$ um conjunto de índices que contém o padrão p_0 . O conjunto de índice é utilizado para fazer a projeção da base de dados.

O terceiro passo do ARMADA, então, é a mineração de padrões por meio do conjunto de índices, mostrado na linha 5. Para tal, ARMADA encontra quais os estados frequentes no conjunto de índices e para cada estado é criado um subconjunto de índices que contém o padrão de entrada, prefixo, e o estado frequente p_1 . Recursivamente é feita a projeção da base até não ser mais possível gerar um padrão maior frequente.

Algoritmo 3: Algoritmo ARMADA.

Entrada: Base de dados Temporal D , $minSup$.

Saída: Todos os padrões temporais frequentes

- 1 Lê D em MDB (base de dados em memória) para encontrar os estados frequentes ;
 - 2 **para cada** *estados frequentes* s **faça**
 - 3 para os padrões $p \leftarrow \langle s \rangle$, envie para saída p ;
 - 4 construa $p.indice \leftarrow$ *crie conjunto de índice*(s, hi, MDB) ;
 - 5 minere conjunto índices($p, p.indices$) ;
 - 6 **fim**
-

Fonte: algoritmo extraído de Winarko e Roddick (2007).

Outro exemplo de trabalho que realiza a extração temporal é Janssoone (2015) cujos autores apresentam um método para prever sequências de sinais expressas por humanos durante uma conversa. Para tanto, o método utiliza a extração de regras de associações temporais: se um participante de uma conversa expressa um sinal, esse mesmo participante expressará outro sinal em um determinado período de tempo.

Há também trabalhos que realizam a extração de padrões considerando as restrições espaciais como Jayababu, Varma e Govardhan (2016) cujos autores apresentam algoritmo para mineração de regras de associação topológicas incrementais a partir de uma base de dados geográficos. No algoritmo, a base de dados é lida e o processo de minerar as regras de associação espaciais topológicas resulta em regras que expressam a relação topológica entre os objetos espaciais contidos na base de dados.

2.3 Regras de Associação Espaço-Temporais

Dados espaço-temporais são dados que podem ser organizados em função de propriedades tanto espaciais quanto temporais como evidenciado nos trabalhos Mennis e Guo (2009), Fang e Wu (2013), Madraky, Othman e Hamdan (2014), Mahmood *et al.* (2013), Spatenkovaa e Virrantausb (2013). Temos como exemplos de dados espaço-temporais: dados meteorológicos, exemplos de trabalhos que utilizam dados meteorológicos para a mineração de dados são Lausch, Schmidt e Tischendorf (2015), Sammouri *et al.* (2012), Sammouri *et al.* (2013); dados de sensores, exemplos de trabalhos que utilizam

dados de sensores para a mineração de dados são Silveira-Junior, Santos e Ribeiro (2013), Silveira-Junior *et al.* (2015), Kodeswaran *et al.* (2016); tráfego de rede, exemplos de trabalhos que utilizam dados de tráfego de rede para a mineração de dados são Rashid, Gondal e Kamruzzaman (2013), Rashid, Gondal e Kamruzzaman (2014); entre outros.

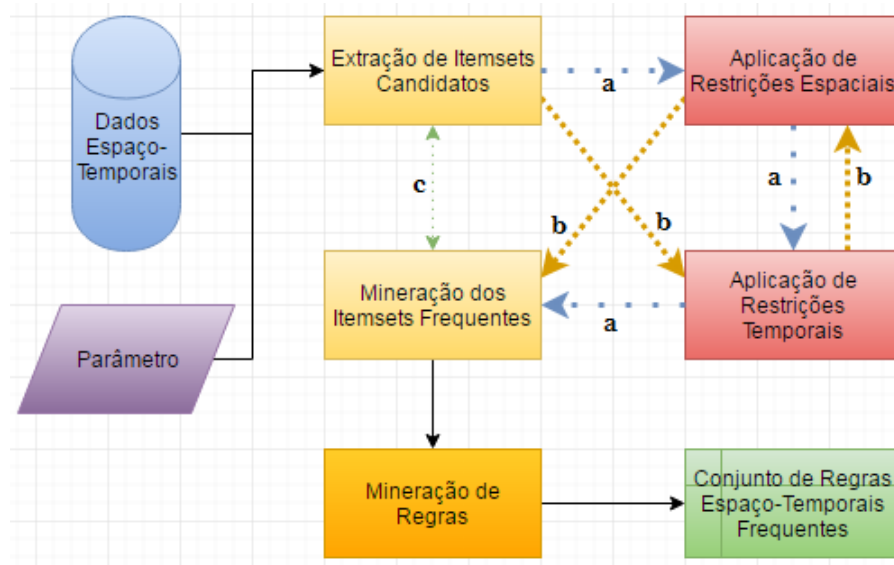
Uma base de dados D é espaço-temporal, se e somente se, seus itens possuem propriedades espaço-temporais (YIN, 2009; SHAHEEN; SHAHBAZ; GUERGACHI, 2013; YAZTI; KRISHNASWAMY, 2014). Portanto, se i é um item qualquer tal que $i \in D$ então i é definido pela quintupla $\{x, y, z, t, F\}$, sendo x, y, z coordenadas em um espaço cartesiano, t uma coordenada temporal e F um conjunto de atributos temáticos, isto é, atributos cujos valores são dependentes do tempo/espaço. Por exemplo, em uma base de dados que armazena o nível de precipitação em estações meteorológicas, seus itens podem ser definidos como $\{S11, W65, 0, 20150824, 0.123mm\}$, sendo as coordenadas 11 ao sul, 65 ao oeste, ao nível do mar, na data de 24 de agosto de 2015 e o valor de 0.123 mililitros de chuva. Os itens, também chamados de objetos, representados em bases espaço-temporais, também podem representar objetos com formas, sejam elas retas, polígonos, ou áreas de um plano.

Os algoritmos que fazem a extração de padrões devem considerar restrições de localidades e também restrições temporais. Na literatura, essas restrições podem aparecer tanto em tempo de pré-processamento, exemplos de trabalhos que aplicam restrições em pré-processamento são Su, Zhou e Shi (2004), Chen, Mao e Liu (2014), Kim *et al.* (2012), como em pós-processamento, exemplo de trabalho que aplica restrições em pós-processamento Zaragoza *et al.* (2012). A vantagem do pré-processamento está na diminuição do espaço de busca durante o processamento do algoritmo, tendendo a reduzir o tempo de execução. A vantagem do pós-processamento é haver um resultado prévio que ajuda a ajustar melhor as restrições que selecionarão os padrões mais interessantes para o domínio.

É também possível aplicar as restrições espaço-temporais durante o processo de mineração como no trabalho de Pillai *et al.* (2012), Obulesu e Reddy (2016), Aljawarneh *et al.* (2017). Na Figura 3, é ilustrada a aplicação das restrições espaço-temporais durante a geração dos *itemsets* frequentes. Como estes *itemsets* respeitam as restrições espaço-temporais são chamados de *itemsets* espaço-temporais.

Inicialmente o processo de Extração de *Itemsets* Candidatos geram os *itemsets* candidatos e então faz a Aplicação das Restrições Espaciais, seguindo o caminho a na Figura 3. Em seguida, os *itemsets* candidatos que respeitam as restrições espaciais vão para o processo de Aplicação das Restrições Temporais, gerando os *itemsets* candidatos espaço-temporais. Logo depois desse processo é feita a verificação de frequência dos *itemsets* espaço-temporais candidatos, resultando nos *itemsets* espaço-temporais frequentes. O algoritmo continua na iteração de geração e teste dos *itemsets* até não ser mais possível gerar *itemsets* frequentes (mesma estratégia do Apriori), seguindo o caminho c na Figura 3.

Figura 3 – Diagrama simplificado do processo de mineração de regras de associação espaço-temporal.



Fonte: o próprio autor com base no trabalho de Pillai *et al.* (2012).

Também é possível inverter a ordem na qual as restrições espaço-temporais são aplicadas, seguindo o caminho *b* na Figura 3. Desta forma, é inicialmente aplicada as restrições temporais e depois as restrições espaciais, gerando assim os *itemsets* espaço-temporais candidatos. Ademais, independentemente da ordem de aplicação das restrições espaço-temporais também é possível iterar entre o conjunto de restrições.

São encontrados na literatura, basicamente, três tipos de RA distintas para domínios espaço-temporais. Cada tipo de RA pode ser utilizada para atingir objetivos diferentes, são elas (RAO; GOVARDHAN; RAO, 2012):

Objetos Móveis: descreve o movimento de objetos entre regiões. Uma regra desse tipo significa que um objeto que satisfaz uma determinada condição *c* migrou da região r_1 para a região r_2 em um determinado período de tempo $[t_1, t_2]$. O formato da regra é $(r_1, t_1, c) \rightarrow (r_2, t_2) < sup, con >$, sendo *sup* o suporte e *con* a confiança da RA. O suporte é definido como o número de objetos que migraram da região r_1 para a região r_2 no período de tempo entre t_1 e t_2 dividido pelo número de objetos que satisfazem *c* no mesmo período. A confiança é definida como o número de objetos que migraram dividido pelo total de objetos na região r_1 no tempo t_1 . Exemplos de trabalhos são Wang, Ng e Chen (2012), Wang *et al.* (2013), Mohan e Revesz (2014), Honda e Mori (2015) e Lemmerich *et al.* (2016).

Relações Topológicas: são regras que envolvem topologia e predicados espaciais como sobreposição, intersecção, toque...; além de predicados temporais como sequencial, paralelo etc. Essas bases necessitam serem pré-processadas para encontrar as relações

topológicas e para organizar os dados em função destas relações. Somente após o pré-processamento é possível fazer a aplicação das técnicas de mineração de dados. O formato da regra é $R_1(obj_1, obj_2, t_1) \rightarrow R_2(obj_3, obj_4, t_2) < sup, con >$, sendo R_1 e R_2 relações espaciais, $obj_{1...4}$ características dos objetos que os diferencie nos períodos de tempo t_1 e t_2 ; por exemplo, *sobreposição(São Carlos, enchente, verão) → vizinhaça(São Carlos, alta vazão de rios, outono)*. Um exemplo de trabalho é Burbey e Martin (2012).

Regras Temáticas: são regras de associação que envolvem propriedades espaciais, propriedades temporais e também os atributos chamados temáticos, cujos valores podem ou não depender do tempo/espaço. Muitas vezes a extração desse tipo de regra necessita de um pré-processamento da base dados. O pré-processamento visa expor as propriedades espaço-temporais e associa-las aos atributos temáticos. Regras temáticas, geralmente, apresentam o seguinte formato: $a_1(R_1, t_1) \rightarrow a_2(R_2, t_2) < sup, con >$, sendo a_1 e a_2 atributos temáticos, como, temperatura e pressão atmosférica no domínio climático, R_1 e R_2 regiões associadas a esses atributos e t_1 e t_2 períodos de tempo. E.g. *chuva(São Carlos, mes₂), chuva(Analândia, mes₂) → correnteza(Ribeirão do Feijão, mes₃) < sup, con >*, esse exemplo mostra que chovendo em São Carlos e Analândia no mês de fevereiro, haverá correnteza no Ribeirão do Feijão no mês de março, com suporte e confiança de *sup* e *con*, respectivamente. Um exemplo de trabalho é Landgrebe *et al.* (2013).

Com esses três tipos de RA espaço-temporais é possível extrair conhecimento envolvendo relacionamento de variações espaço-temporais entre valores de atributos. Esses tipos também demonstram como a mineração de RA espaço-temporal é uma tarefa flexível de MD (Mineração de Dados).

Um exemplo de algoritmo extrator de RA espaço-temporal é encontrado em Compieta *et al.* (2007). Este trabalho se baseia no Apriori para realizar a extração de RAs espaço-temporais. Em seu modelo de dados espaço-temporal, cada item ι é associado a um conjunto de pontos espaciais pe_ι no qual o item ocorre em um determinado tempo/período t . Um ponto virtual pv é definido com um ponto espacial que suporta um *itemset*. Um *itemset* espacial só é considerado frequente, se e somente se, for frequente no conjunto de pontos virtuais pvs . O ponto virtual pv tem a sua existência associada a um ou mais períodos temporais. Baseado nisso, a ideia do algoritmo é evitar processamentos desnecessários realizados pelo algoritmo Apriori Tradicional, como apresentado em Agrawal e Srikant (1994); assim, essa abordagem processa apenas os dados com maior relação espaço-temporal, usando os pontos virtuais. O resultado é um conjunto de *itemsets* frequentes e RAs. Outra alteração realizada ao Apriori Tradicional é que durante a junção de dois *itemsets*, ι_1 e ι_2 , feito para a geração dos candidatos, é verificado se a intersecção dos conjuntos de pontos virtuais associados (pv_{ι_1} e pv_{ι_2}) não é vazia, $pv_{\iota_1} \cap pv_{\iota_2} \neq \emptyset$.

2.4 Estado da Arte em Mineração de Dados Espaço-Temporal

Kawale *et al.* (2012) aplicam a extração de séries temporais em dados climáticos, dados de precipitação e de temperatura, para determinar anomalias que acontecem em uma região e que podem vir a acontecer em outra região após algum tempo. Nesta abordagem, os autores realizaram a extração de padrões positivos e negativos utilizando uma abordagem baseada em grafos, representando as restrições espaciais por meio de arestas. A ideia dessa abordagem é agrupar as localizações tal que membros de um agrupamento tenham características mais similares entre si do que em relação aos membros de agrupamentos distintos. Portanto, o objetivo do algoritmo é encontrar pares de agrupamentos com características diferentes. Como limitação, essa abordagem não leva em consideração a evolução temporal na identificação dos padrões.

Abirami-Kongu, Thangaraj e Priakanth-Kongu (2012) aplicam a mineração espaço-temporal a dados de tráfego de rede para a identificação de falhas em redes de sensores não-cabeados. Em seu estudo, a mineração é utilizada para a investigação das associações dentre os dados que chegam por meio da rede em busca de falhas. Em simulação, esse trabalho mostra que é possível utilizar essas associações para identificar falhas nos nós e nas ligações de uma rede. Esse estudo utilizou o algoritmo Predictive Apriori (SCHEFFER, 1995) sem alterações, além disso, não foi aplicado pré-processamento dos dados de tráfego da rede, assim, foi evitada a inserção de ruídos ou perda de dados pelo pré-processamento. O Predictive Apriori se difere do Apriori por balancear dinamicamente os valores de suporte e confiança mínimos; o Predictive Apriori foi proposto para a extração das N melhores regras (maior suporte e confiança) para um conjunto de dados, sendo N um parâmetro definido pelo usuário. Como limitação, o desempenho do Predictive Apriori tende a deteriorar com um conjunto de dados grande. Ademais, essa técnica teria de ser generalizada para poder ser aplicada em outros domínios como o de dados não convencionais.

Yoo e Bow (2012) propõem um *framework* para encontrar padrões correlacionados composto por dois algoritmos para a mineração dos N -vizinhos mais próximos. Dada a abordagem de filtrar as relações espaço-temporais e refinamento da forma dos objetos, esta abordagem reduz o número de candidatos em relação as abordagens tradicionais de mineração de dados. Foram também avaliadas maneiras distintas de se determinar os vizinhos mais próximos pelo uso de estimativas da distância em relação a cada borda do objeto. Como limitação, esse *framework* não possibilita evidenciar a evolução de um único elemento no correr do tempo.

O trabalho de Hana, Sami e Sami (2012) processa as relações de vizinhanças entre objetos durante um intervalo de tempo pelo uso de consultas espaciais com parâmetros temporais. Devido a utilização de bancos de dados espaciais, a MD tende a ter um melhor desempenho. O algoritmo desenvolvido é chamado de START. O START apresenta três fases: i) cálculo dos predicados espaço-temporais; ii) geração

dos *itemset* frequentes, e; iii) extração das regras espaço-temporais (baseado no Apriori). Nesse trabalho, os objetos espaço-temporais são caracterizados por uma quadrupla $\langle a_i, g_i, p_i, t_i \rangle$. O START mostra a evolução espaço-temporal de objetos geográficos, $(X, a_i t_i)(X, g_i, t_i)(X, p_i, t_i) \rightarrow (R, cr, t_{i+\epsilon})$, sendo X a referência a um objeto na base de dados, a_i o atributo que caracteriza X no tempo t_i , g_i a característica geométrica de X no tempo t_i , e p_i é uma relação topológica que X possuirá com um objeto caracterizado por cr , e R a probabilidade desse objeto ocorrer próximo a X em um tempo $t_{i+\epsilon}$. Por exemplo, $(Chuva, 0.15mm, outono), (Chuva, 20km^2, outono), (Chuva, vizinhança, outono) \rightarrow (0.8 \frac{m^3}{s}, aumento\ vazão, outono + 1mês)$, essa regra pode ser lida como: Chuva com as características de $0.15mm$, em uma região de $20Km^2$ durante o outono, implica em um aumento de $0.8 \frac{m^3}{s}$ no mês subsequente ao outono. Esta abordagem tem como limitação não considerar que mais de um evento pode influenciar outro e não levar em consideração informações que podem estar contidas em dados não-convencionais, como imagens. Além disso, é necessário um pré-processamento para definir os predicados espaço-temporais dos dados, definição de vizinhança sobreposição etc, o que faz a abordagem desconsiderar padrões que possam ser importantes por não ter um predicado espaço-temporal previamente definido para a combinação dos dados.

Huo, Zhang e Meng (2013) encontram regras espaço-temporais co-ocorrentes por meio da aplicação de um janelamento deslizante cujos incrementos de dados é dinâmico e o decaimento de importância também (em função da entrada dos dados). O algoritmo proposto, DIAD, faz a utilização de árvores *hash* para o armazenamento e acesso aos padrões, por meio dessa abordagem, o DIAD apresenta um ganho de desempenho em relação aos algoritmos que fazem as verificações por varredura. O DIAD distribui os eventos em partições e calcula a distância espacial entre os eventos. À medida que novos dados são adicionados, estes são distribuídos entre as partições e é feita a atualização dos eventos nas partições atingidas. O decaimento de importância é uma técnica adaptada que visa capturar dinamicamente mudanças no fluxo dos eventos; essa técnica atribui menor importância aos dados antigos e longe do ponto central da MD atual. O domínio utilizado é dados oriundos de redes sociais, que possuem referências espaciais. Essa abordagem não trabalha com dados complexos, por exemplo imagens. Para minerar imagens é necessária a extração de característica da imagem, porém, o domínio que DIAD se propõe a minerar um fluxo de dados intenso, desta forma, processar imagens causaria um atraso ao processamento e desempenho do algoritmo. Além de possíveis problemas com o janelamento deslizante.

Pillai *et al.* (2012) apresentam um novo algoritmo para encontrar regras espaço-temporais por meio da aplicação de filtros; esses filtros são usados para restringir os padrões que respeitam restrições espaço-temporais. O algoritmo também aplica o refinamento de formas geométricas para levar em consideração a topologia dos eventos. Esse trabalho utilizou uma base de dados de imagens solares; por meio da aplicação dessa técnica foi possível encontrar formas de eventos (explosões) que se moviam. O algoritmo, baseado

no Apriori, tem a capacidade de lidar com uma quantidade bastante grande de dados e é chamado FastSTCOPs-Miner. Pillai, Angryk e Aydin (2013) apresentam uma evolução ao trabalho anterior nos seguintes aspectos: um novo *framework* para a mineração de padrões co-ocorrentes é apresentado; os eventos espaciais são modelados como objetos 3D e a evolução de suas formas são capturadas; permite fazer a sobrescrita de formas diferentes tendo por base o volume, e; por fim, é apresentado um algoritmo para descoberta de RAs co-ocorrentes com base na evolução das relações espaciais. Aydin *et al.* (2015) apresentam uma evolução do *framework* adicionando índices que facilitam a estratégia de filtro-e-refina. Dessa forma, houve uma melhoria no desempenho do *framework*, porém não houve alterações nos padrões extraídos. Esses trabalhos têm como limitação o fato de não extraírem regras de associação, mas sim padrões sequenciais para a evolução de um evento, desta forma, as abordagens não consideram a implicação que um evento pode ter sobre outro. Além disso, nenhuma das abordagens consideram os atributos temáticos alfanuméricos do domínio, apenas as imagens e restrições espaço-temporais.

Cunjin *et al.* (2015) apresentam o algoritmo MIQarma para extrair associações de séries temporais de imagens de sensores. MIQarma se divide em duas etapas: 1) calcula a informação assimétrica entre os itens e os agrupa em pares respeitando as restrições do usuário, e; 2) gera as regras de associações com base no valor de suporte e confiança fornecidos pelo usuário. MIQarma espera que um passo de pré-processamento seja feito no qual as imagens são discretizadas para representações de objeto, em outras palavras é necessário rotular o conteúdo das imagens para realizar a mineração de dados. Desta forma, é bastante trabalhoso a reutilização do MIQarma em outros domínios e também a introdução de novos dados na base. Xue *et al.* (2015) apresentam a evolução do trabalho de Cunjin *et al.* (2015) acoplando uma etapa de pré-processamento ao algoritmo e transformando-o assim em um *framework*. Xue *et al.* (2016) apresenta uma nova evolução desse trabalho o que possibilita ao *framework* a extração de padrões sequenciais também. No entanto, as mesmas limitações são presentes em ambos os trabalhos.

Mouri, Ogata e Uosaki (2015) apresentam uma abordagem para a mineração espaço-temporal de dados não convencionais aplicada ao domínio de análise de logs de aprendizagem ambígua. A abordagem visa prever as próximas etapas da aprendizagem com base nos dados históricos. Como limitação, nessa abordagem, é levado em consideração apenas as relações espaço-temporais que o conjunto de logs possuem, desta forma, são extraídas as relações de vizinhanças desses dados; o conteúdo semântico dos dados (atributos temáticos) não é considerado para o processo de mineração.

Silvestri *et al.* (2015) apresentam um novo tipo de padrão espacial de vizinhança que considera uma representação comprimida com perdas dos padrões de vizinhança. Como ponto fraco, essa abordagem busca relações de vizinhança não explicitando a evolução temporal dos objetos. Ademais, a abordagem não considera atributos temáticos durante o

processo de extração de padrões.

Radhakrishna, Kumar e Janaki (2016) apresentam uma abordagem para a extração de padrões temporais do tipo *outlier*. Nessa abordagem são extraídas regras de associação temporais com alto nível de confiança e baixo nível de suporte, por meio da estratégia de filtro-e-refina de mineração de padrões. Essa abordagem pode ser generalizada para diversos domínios, como apresentado no artigo: médico, *internet of things*, dentre outros. Como ponto negativo, essa abordagem considera apenas os parâmetros temporais durante a extração de padrões sendo necessário aplicar as restrições espaciais a nível de pré-processamento. Devido à aplicação de restrições a nível de pré-processamento, esse tipo de abordagem pode sofrer perda de dados e introdução de ruído na base, pois dados frequentes podem ser prematuramente desconsiderados resultando em padrões espaço-temporais com baixa confiança.

Shi *et al.* (2018) apresentam uma abordagem para extrair anomalias de bases de dados espaço-temporais. Uma anomalia pode ser causada por ruídos ou por uma tendência de mudança no comportamento dos dados. Esta abordagem se assemelha a extração de regras de associação negativas em base de dados tradicionais, porém, aplicada a dados espaço-temporais. A abordagem visa encontrar as relações de vizinhança entre objetos que são não-frequentes. É utilizada uma matriz de similaridade entre os objetos para diferenciar se a relação não-frequente encontrada é causada por um ruído ou por uma tendência na mudança do comportamento dos dados. Essa abordagem somente considera objetos dissimilares na mesma relação. Como limitação, essa abordagem não considera atributos temáticos para o seu processamento, nem a evolução dos objetos no decorrer de um período.

2.5 Considerações Finais

A mineração de dados consiste na extração de padrões de interesse nos dados. Base de dados espaço-temporais apresentam um desafio a mineração de dados pela sua complexidade e a necessidade de tratar múltiplas dimensões de correlações entre eventos.

Este capítulo teve como foco os processos de mineração de dados dando ênfase na extração de Regras de Associação aplicadas a dados espaço-temporais. A proposta deste trabalho de pesquisa foca na extração de regras de associação espaço-temporais temáticas, definidas neste capítulo. Por fim, este capítulo apresentou o estado da arte levando em consideração a mineração de regras aplicada a dados espaço-temporais, salientando as principais características e limitações das abordagens correlatas ao trabalho proposto nesta pesquisa.

Dentre as diversas abordagem apresentadas na seção estado da arte, a extração de regras de associação espaço-temporais temáticas utilizando a estratégia de crescimento de

padrões é a mais promissora para o domínio de dados solares utilizado neste projeto de pesquisa. As abordagens para extração de padrões sequenciais não consideram a relação que eventos próximos (dentro de uma restrição espacial), assim, não é possível dizer o quanto a evolução de um evento impacta em seus vizinhos. As abordagens para a geração e teste de regras de associação geralmente apresentam um desempenho pior que as abordagens baseadas na estratégia de crescimento de padrões. Além disso, crescimento de padrões é também uma estratégia que facilita a paralelização, pois partes diferentes dos algoritmos sempre trabalham com partes diferentes dos dados (dividir e conquistar). Essa abordagem também é mais promissora que as abordagens baseadas em grafos, pois apresentam um melhor desempenho quando o conjunto de dados é grande; o desempenho de abordagens baseadas em grafos degrada exponencialmente com o aumento no espaço de busca de soluções.

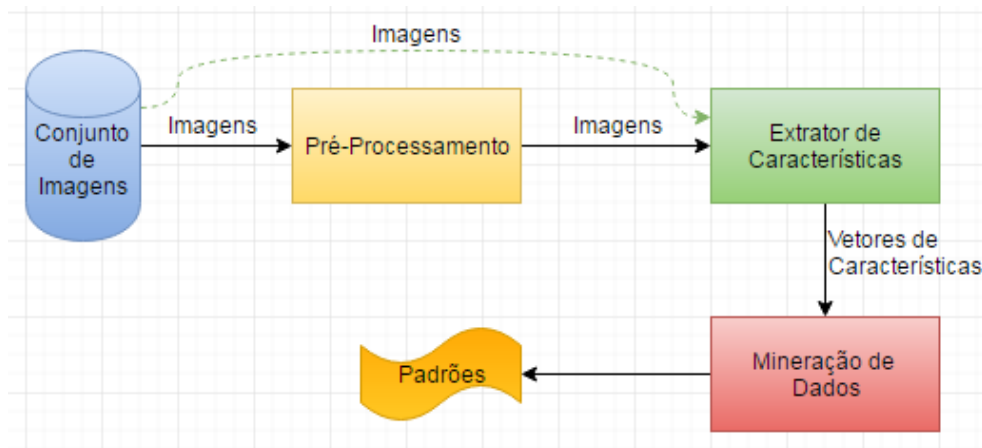
3 Mineração de Dados Aplicada a Imagens

Os processos de extração de conhecimento aplicados a dados complexos apresentam uma série de desafios. De modo geral, os dados complexos, tais como imagens, vídeos, e séries temporais, passam por etapas de pré-processamento cujo objetivo é prepara-los para o processo de mineração.

A mineração de imagens trata da extração de conhecimento implícito (FATMA; NASHIPUDIMATH, 2011; WANG *et al.*, 2018; CECI *et al.*, 2015; DESHMUKH; BHOSLE, 2016a) de bases de imagens. Esse conhecimento implícito pode ser o relacionamento entre as imagens de um conjunto, padrões de textura ou formas implícitas nas imagens, evolução de determinados objetos, entre outros. O processo de mineração de imagens pode ser auxiliado por técnicas que ajudam a revelar esse conhecimento implícito como visualização (ANDRIENKO *et al.*, 2011; ZURITA-MILLA *et al.*, 2013), classificação de imagens (JULEA *et al.*, 2011; JULEA; MEGER, 2013; JULEA *et al.*, 2014), etc. Ao longo da mineração de dados, as imagens são representadas por vetores de características que representam suas características visuais.

Em vias gerais, o processo da mineração de dados aplicado às imagens segue o seguinte procedimento: pré-processamento das imagens, extração de características, mineração de dados e interpretação (SU *et al.*, 2004; EL-HAJJ; ZAIANE, 2006; DESHPANDE; RAJURKAR; MANTHALKAR, 2013; SAMIA; ASSIA, 2015; SONAR; JADHAV; BHOSLE, 2016; SONAR; BHOSLE, 2016), como apresentado na Figura 4.

Figura 4 – Processamento de imagens para a mineração de dados.



Fonte: figura adaptada de Fatma e Nashipudimath (2011).

A etapa de pré-processamento das imagens não é obrigatória. Há algoritmos de mineração de dados que necessitam que os vetores de características sejam discretizados

para permitir a mineração. A discretização consiste em transformar o vetor de características composto por valores numéricos contínuos em um vetor de características composto por valores discretizados em intervalos. Essa etapa também é optativa.

Após o pré-processamento e extração de características, a informação contida na imagem é então processada por algoritmos de mineração de dados tradicionais. Se o processo não for realizado apropriadamente, o resultado do algoritmo de mineração pode apresentar informação dúbia e inapropriada. Os padrões extraídos ainda passam por uma etapa de interpretação - muitas vezes pelo uso de técnicas de visualização - para melhorar o entendimento dos padrões minerados.

Assim, por este trabalho aplicar mineração em imagens, este capítulo aborda a Mineração de Imagens e está dividido da seguinte maneira: na Seção 3.1, são apresentados os processos referentes ao pré-processamento de imagens; na seção 3.2, são apresentados os processos de extração de características; na Seção 3.3, é apresentado o estado da arte com os trabalhos relacionados que aplicam a mineração de imagens a dados oriundos de satélites, e por fim; na Seção 3.4, são apresentadas as considerações finais.

3.1 Pré-Processamento de Imagens

A etapa de pré-processamento de imagens visa realçar características interessantes presentes em imagens. Também é de comum objetivo destas técnicas a redução de ruídos contidos nas imagens (ALMODAIFER; HAFEZ; MATHKOUR, 2011; MINGYING, 2011; LIU *et al.*, 2016; RODRIGUES *et al.*, 2015; YU; DONGMING, 2016; YEH; LEE; LIN, 2015).

Algumas técnicas utilizadas e presentes na literatura são (ZHANG; WILLIAMS; WANG, 2017; TELIKANI; SHAHBAHRAMI, 2018): (i) remoção do fundo da imagem, ressaltando o conteúdo da imagem; (ii) suavização que consiste na aplicação de filtros para a redução de ruídos. Há também a aplicação de filtros para a remoção de detalhes não relevantes ou para a conexão de linhas descontínuas, e; (iii) máximo, mínimo e média dos pixels vizinhos, utilizados para normalização das intensidades dos pixels das imagens.

A utilização de tais técnicas visam, em geral, a melhoria da qualidade da imagem para a posterior aplicação do extrator de características na mesma.

3.2 Extração de Características de Imagens

Após o pré-processamento, o conjunto de imagens passa pela extração de características (SRINIVASULU; SAKTHIVEL, 2010; VATSAVAI, 2013). Essa etapa é de suma importância para a mineração de dados, pois obtém a representação das imagens

por meio de vetores de características, que possibilita a comparação entre as imagens (BANDYOPADPHYAY; MAULIK, 2013; FU, 2012; KENNEDY *et al.*, 2015).

Existem várias técnicas para a extração de características, entre elas destacamos:

Detecção de Formas : consiste em encontrar aproximações ou medidas para as formas encontradas nas imagens. De maneira simplificada, esse tipo de extração de características permite distinguir diferentes formas geométricas. Exemplos de detecção de formas são (i) Detecção de Arestas que consiste na extração dos contornos que são relevantes para a imagem, desta forma, informações como cores e textura são descartadas, e; (ii) Detecção de Cantos que consiste em encontrar mudanças bruscas de direção de uma aresta e seu contorno.

Detecção de Textura : consiste em detectar textura sobre superfícies contidas na imagem. É bastante utilizada quando o objetivo é determinar os objetos contidos em uma imagem, classificação, (BLANCHART; FERECATU; DATCU, 2011). Exemplos de detecção de texturas são (i) Matrizes de Co-Ocorrência que consiste em utilizar a combinação de diferentes valores que representam em níveis de cinza a intensidade pixel-a-pixel de uma vizinhança, desta forma, é possível representar e comparar texturas em imagens, e; (ii) Avaliação de Médias Texturas que consiste em utilizar janelas que subdividem a imagem para a detecção de diferentes tipos de texturas, desta forma, as janelas são utilizadas para calcular as texturas médias de cada sub-região, sendo que esses valores compõem o vetor de características da imagem.

Detecção de Cores : consiste em determinar a frequência da intensidade de cores das imagens. Um exemplo de detecção de cores é o histograma, que consiste em contar quantos píxeis a imagem possui para cada intensidade de cor. Essa frequência é utilizada para construir o vetor de características, que possui um valor de frequência para cada intervalo de intensidade contabilizado.

No domínio de imagens oriundas de satélites, é comum a aplicação de diversas técnicas de pré-processamento para a redução de ruídos (HONDA; KONISHI, 2001; RIEDEL *et al.*, 2015). A palavra ruídos é utilizada neste trabalho de pesquisa para se referir às nuvens, poluição, efeitos atmosféricos, etc. Muitas vezes são também utilizadas combinações de técnicas de pré-processamento de imagens, extração de características e pré-processamento de dados (normalização, discretização e seleção de características) para melhorar o resultado do processo de descoberta de conhecimento. Muitos trabalhos referentes a mineração de imagens de satélite, apresentados na literatura, aplicam classificação ou agrupamento nos vetores de características pré-processados das imagens.

Outro ponto relevante sobre imagens oriundas de satélite é sua abundância. Atualmente, uma grande gama de informação é oriunda dessas fontes remotas (DONG; LIU;

ZHAO, 2009; VATSAVAI *et al.*, 2012). Grande parte dos trabalhos que fazem a utilização desses grandes conjuntos de dados são direcionados a identificar mudanças de cores, texturas e/ou formas que ocorrem em uma determinada região, como apresentado por Tsai *et al.* (2013).

Devido a relevância para esse trabalho de pesquisa, na Subseção 3.2.1, é apresentado o descritor baseado em textura, Descritor de Haralick, e, na Subseção 3.2.2, é apresentado o vetor baseado em formas, SURF. Para a mineração de regras de associação neste trabalho, foi necessário que os vetores de características passassem por um processo de discretização. Para isso, foi utilizado o algoritmo Omega que é apresentado na Subseção 3.2.3.

3.2.1 Descritor de Haralick

O Descritor de Haralick foi proposto por Haralick, Shanmugam e Dinstein (1973) para a representação de imagens em escala de tons de cinza por meio das texturas que as imagens apresentam (MIYAMOTO; MERRYMAN, 2005). No trabalho de Attig e Perner (2011), é apresentada uma abordagem para levar em consideração as cores contidas nas imagens durante o processo de extração de textura. O Descritor de Haralick, apesar de ter sido proposto em 1973, ainda é largamente utilizado como sugere o trabalho de Zayed e Elnemr (2015).

A técnica de extração do descritor é dividida em duas etapas, a computação das Matrizes de Coocorrência de Tons de Cinza e o cálculo do vetor de características propriamente dito. Uma Matriz de Coocorrência mostra a frequência com que cada tom de cinza ocorre em um pixel localizado em uma posição geométrica fixa em relação aos outros pixels. Baseado na Matriz, nove medidas para os descritores de textura são calculadas: (i) energia derivada do segundo momento angular da imagem, que mede a uniformidade dos tons de cinza da imagem; (ii) a média dos valores dos pixels; (iii) o desvio padrão dos valores dos pixels; (iv) a medida do grau de assimetria na distribuição dos pixels; (v) a medida do pico da distribuição dos pixels; (vii) o cálculo da entropia para a Matriz de Coocorrência de Tons de Cinza; (viii) contraste da imagem, e; (ix) homogeneidade.

Para uma mesma imagem, a Matriz de Coocorrência pode ser computada para diversas posições geométricas (em diferentes angulações e espaçamento dos pixels). Para cada Matriz de Coocorrência computada são extraídas as nove medidas estatísticas que compõe o vetor de características da imagem. Assim, dependendo do número de Matrizes de Coocorrência geradas para a imagem, o vetor pode ter um número maior ou menor de características.

3.2.2 Extração de Características Baseado em Formas por meio do Algoritmo *Speeded Up Robust Features*

A extração de vetores de características *Speeded Up Robust Features* (SURF) foi proposta por Bay, Tuytelaars e Gool (2006). SURF foi projetado para a extração de características baseado em forma, tende a ser escalável e capaz de produzir resultados satisfatórios mesmo em imagens que apresentam ruídos.

No domínio de imagens solares, as manchas solares evoluem e com isso há alteração em suas formas. Assim, a extração de características baseada em forma foi escolhida para este trabalho por retratar a evolução que uma mancha solar sofre. Dentre os extratores de forma (YANG; KPALMA; RONSIN, 2008), a escolha do SURF ocorreu devido ao fato de que o SURF trabalha com a descontinuidade da coloração na imagem para encontrar áreas de interesse na imagem. Para cada área de interesse é gerado de um vetor de características (como posteriormente é detalhado nesta seção). Assim, pela utilização do SURF, são gerados um vetor de característica para cada mancha solar diferenciando-as dentro da imagem. Além disso, como mostrado na literatura (BAY; TUYTELAARS; GOOL, 2006), SURF tem desempenho superior aos extratores de forma clássicos.

Neste trabalho de pesquisa, o SURF é usado para a delimitação das manchas solares, tornando possível a associação das mesmas a seus dados semânticos (apresentado no Capítulo 5).

O algoritmo se divide em duas etapas:

Detecção dos Pontos de Interesse: filtros nos quais são utilizadas janelas quadradas fazem uma aproximação para realizar uma suavização Gaussiana em áreas de interesse da imagem. O Filtro é definido por:

$$S(x, y) = \sum_{i=0}^x \sum_{j=0}^y I(i, j)$$

Sendo i e j pontos adjacentes a x e y , e I a imagem.

A obtenção dos pontos de interesse é feita por meio da detecção de manchas na coloração da imagem por meio da matriz de Hessian. A matriz de Hessian é uma matriz determinada por meio da medida da mudança local em volta de cada ponto. SURF aplica o determinante de Hessian para obter a escala.

Assim sendo, seja o ponto $P = (x, y)$ de uma imagem I , a matriz de Hessian $H(P, \vartheta)$ em ponto P com a escala ϑ :

$$H(P, \vartheta) = \begin{pmatrix} L_{xx}(P, \vartheta) & L_{xy}(P, \vartheta) \\ L_{yx}(P, \vartheta) & L_{yy}(P, \vartheta) \end{pmatrix}$$

Sendo que, $L_{xx}(P, \vartheta) \dots$ são as derivadas de segunda ordem da imagem em tons de cinza. A aproximação Gaussiana ϑ representa o nível de maior resolução e é definida como: $\vartheta = \text{Tamanho da Janela do Filtro} \times \frac{\text{Escala Básica do Filtro}}{\text{Tamanho Básico do Filtro}}$. Para a implementação do SURF, o nível mais baixo da escala é calculado pela saída dos filtros cuja janela é 9×9 .

Descritor de Vizinhança Local: descritores de características visam proporcionar uma descrição única de um componente da imagem. Assim, como etapa inicial, é fixado um ponto de orientação para cada região circular cujo centro é um ponto de interesse. Em seguida, é feita a construção de uma região quadrada alinhada ao ponto de orientação, o descritor SURF é obtido a partir dessa região quadrada construída.

Atribuição de Orientação: este passo visa alcançar a invariância de rotação frente ao ponto de interesse. As respostas são obtidas a partir de Wavelets de Haar calculadas em função dos eixos x e y em uma região circular centrada em um ponto de interesse. A orientação dominante é calculada por meio da soma das respostas em uma janela deslizante. As respostas dentro da janela são somadas para obter o vetor de orientação local.

Descritor Baseado na Soma das Respostas das Wavelets de Haar: a descrição de uma região em torno de um ponto é feita por meio da extração de uma região quadrada cujo centro é um ponto de interesse. Cada região é subdividida em regiões menores, e para cada uma dessas regiões são extraídas as respostas das Wavelets de Haar.

3.2.3 Algoritmo Omega

Omega é um algoritmo supervisionado para a discretização dos valores contínuos dos vetores de características proposto por Ribeiro, Traina e Traina Jr. (2008). Foi desenvolvido para o pré-processamento de valores contínuos para a tarefa de associação, para a redução da entropia de cada intervalo gerado e a minimizar o número de intervalos gerados, de maneira a evitar o crescimento exponencial do número de padrões frequentes que podem ser gerados na tarefa de associação.

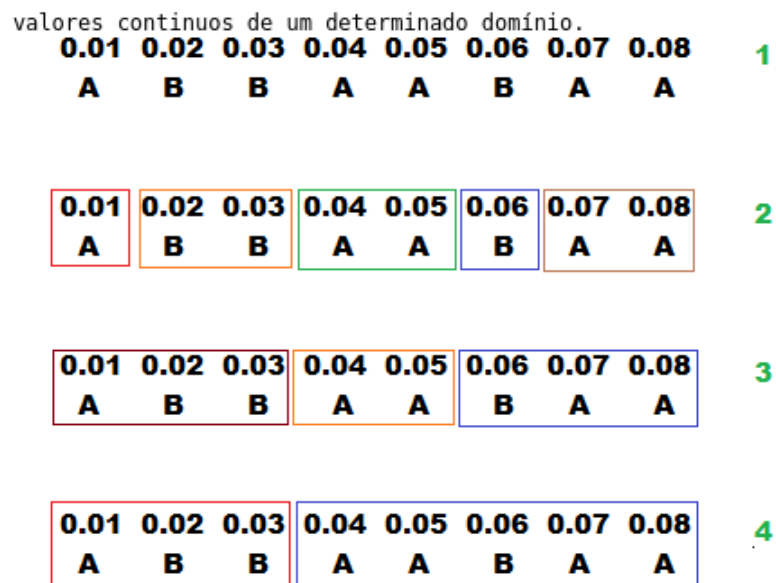
O algoritmo Omega recebe os vetores de características f tal que f_i é um valor do vetor de características f , c_i a instância de uma classe associada à f_i .

Seja U_k e U_{k+1} os limites do intervalo T_k , define-se que $I_i = (f_i, c_i)$ pertence ao intervalo $T_k = [U_k, U_{k+1}]$, se e somente se, $U_k < f_i < U_{k+1}$. Omega é dividido em quatro etapas como é apresentado pela Figura 5.

1. É feita a ordenação dos I_k com base no valor característica, f_k .

2. É feita a delimitação dos intervalos com base nas categorias. São gerados pontos de corte, sempre que houver mudança de categoria. Assim os intervalos gerados nesta etapa tem entropia zero. No exemplo, $T_1 = [0.01, 0.01]$, $T_2 = [0.02, 0.03]$...
3. É feita a fusão dos intervalos consecutivos que não possuem o número mínimo de I_k s. Essa etapa é essencial para reduzir o número de intervalos gerados na etapa anterior. No exemplo, T_1 e T_2 são fundidos formando $T_{1,2} = [0.01, 0.03]$.
4. Intervalos com mesma classe majoritária são fundidos, se isso não violar taxa de inconsistência permitida em cada novo intervalo gerado. A taxa de inconsistência é dada pela formula $\zeta_{T_k} = \frac{|T_k| - |M_{T_k}|}{|T_k|}$, sendo M_{T_k} a classe majoritária no intervalo T_k . No exemplo, é o responsável pela formação do intervalo $T_{3,4,5} = [0.04, 0.08]$ com base no T_3 e $T_{4,5}$ (fundido na etapa anterior).

Figura 5 – Exemplo dos quatro passos de execução do algoritmo Omega. Exemplo extraído de Ribeiro, Traina e Traina Jr. (2008).



Fonte: figura adaptada de Ribeiro, Traina e Traina Jr. (2008).

O algoritmo Omega foi projetado para ser usado no pré-processamento de valores contínuos para a tarefa de mineração de regras de associação. Essa característica é interessante para a extração de regras de associação, pois reduz o número de *itemsets*, facilitando o processamento e melhorando o desempenho do algoritmo de mineração. Por esse motivo, dentre outros algoritmos discretizadores como o descrito em (GARCÍA *et al.*, 2013), o Omega foi utilizado neste trabalho.

3.3 Estado da Arte em Mineração de Imagens

Romani *et al.* (2010) apresentam um novo algoritmo chamando *CLEARMiner* para a extração de regras em séries temporais oriundas de imagens de satélites. O algoritmo faz a extração de regras em dois passos: (i) transforma várias séries temporais em representações de padrões os quais podem ser picos, montanhas e vales, dessa forma é mantida a relação temporal presente nas séries, e; (ii) gera regras que associam os padrões de várias séries temporais. Por meio desse algoritmo é possível obter as regras em séries de imagens de satélites.

Em Mougel e Selmaoui-Folcher (2013), é apresentado um processo para a extração de regiões homogêneas em sequências de imagens de satélites. O algoritmo visa descobrir coleções de objetos que compartilham as mesmas propriedades (como cor) nos mesmos períodos de tempo.

Em Zhou e Zhang (2013), são usadas regras de associação para a classificação de imagens capturadas remotamente. Nessa abordagem, regras de associação associam combinações de características a classes de imagem.

ZuoCheng e LiXia (2014) propõem uma abordagem para encontrar estruturas locais frequentes em uma imagem por meio da extração de regras de associação. A frequência de uma estrutura é usada para caracterizar a textura. Assim, para minerar os padrões de textura frequente, cada imagem é considerada como um conjunto de transações. Os padrões frequentes de textura são chamados células, sendo que uma imagem possui muitas células de textura. Uma regra de associação tende a representar uma característica da textura da imagem. Além disso, a segmentação da imagem pode ser feita de acordo com as regras de associação mineradas.

Pitarch *et al.* (2015) fazem a extração de padrões sequenciais em dados multidimensionais e espaço-temporais. Essa classificação segue quatro passos: (i) tratamento dos dados: é feita a re-escrita das relações espaciais e hierarquia dos atributos; (ii) criação de uma projeção da base: é feita uma projeção para cada classe de valores; (iii) determinação dos atributos frequentes para tanto verifica-se as projeções da base de dados para remover os atributos não frequentes, e; (iv) mineração das sequências por meio da aplicação do algoritmo M2SP (PITARCH *et al.*, 2015, *apud* Plantevit *et al.*, 2005).

Petitjean *et al.* (2011) aplicam a extração de sequências temporais tendo como domínio imagens de satélite. Essa extração visa detectar a evolução de regiões nas imagens. A manipulação das imagens é feita pixel-a-pixel: para encontrar as sequências frequentes e para verificar como elas evoluem em função do tempo. O método apresentado é dividido em fases: segmentação das imagens; caracterização de regiões; construção de vetores de características de imagens; construção de séries temporais, e; classificação das séries temporais.

Petitjean *et al.* (2012) aplicam mineração à imagens de satélite com o objetivo de detectar padrões topológicos evolutivos pixel-a-pixel. Os padrões evolutivos extraídos por esse trabalho são regras de associação espaço-temporais topológicas.

Ambos os trabalhos (PETITJEAN *et al.*, 2011) e (PETITJEAN *et al.*, 2012) considerarem somente as informações visuais das imagens. Desta forma, são desconsideradas informações semânticas que acompanham esse tipo de imagens as quais podem ser importantes para trazer um maior valor semântico aos padrões extraídos.

McGuire, Gangopadhyay e Janeja (2012) aplicam pré-processamento ao conjunto de imagens e, durante a mineração, utilizam também a informação armazenada em arquivos descritivos que acompanham as imagens. Aplicam a mineração levando em consideração relações espaço-temporais através de uma técnica de janelamento deslizante, similar ao IncMSTS (Seção 2.2.1), e para as restrições espaciais é utilizada a distância de *Hamming*. Os padrões extraídos são *itemsets*, por exemplo, pressão relacionada à umidade do ar: $umidade(0.69) \text{ pressão}(0.63) \text{ } r = 0.59$, sendo r o raio. A abordagem não leva em consideração a evolução temporal do padrão, o que a diferencia deste trabalho de pesquisa.

O método apresentado em Lu *et al.* (2013) aplica um pré-processamento para extrair características das imagens. A partir das características é feita a mineração de regras de associação. Essas são utilizadas para recomendação de imagens com base em especificações do usuário. Um algoritmo chamado *Fuzzy rs-Image Recommender* cria um *ranking* para as imagens mais relevantes de acordo com as especificações do usuário e as apresentam com base nas regras extraídas. O método proposto não leva em consideração as restrições espaço-temporais nem os dados semânticos que ajudam na caracterização das imagens, em contraponto a este trabalho de pesquisa. Porém, é correlato por: aplicar pré-processamento para a extração de características das imagens, fazer a mineração de regras de associação levando em consideração as características visuais de cada imagem e pelo grande volume de dados processados.

Kouge *et al.* (2015) apresentam uma abordagem para recomendação de compras baseado no perfil do usuário. Essa abordagem combina uma série de dados de históricos de compras e imagens dos produtos para fazer uma melhor recomendação de produtos, apresentando a imagem do produto mais adequada para a sua venda. Essa abordagem não considera restrições espaço-temporais para a realização do processamento de mineração.

Li (2015) apresentam uma abordagem para a mineração de séries de imagens de satélite baseado na extração de regras de associação. Inicialmente as imagens são discretizadas transformando-as em uma série de dados discreta e então são extraídas as regras de associação temporais. Essa abordagem tem como ponto negativo não considerar restrições espaciais, desta forma, as regras extraídas são sempre relacionada a uma região restrita da imagem, sendo que para considerar outra região é necessário aplicar novamente o processamento sobre os dados da nova região.

Traore, Kamsu-Foguem e Tangara (2017) apresentam uma abordagem que utiliza mineração de dados sobre séries de imagens de satélites para determinar regiões que são passíveis de sofrerem com epidemias. Nos experimentos, foi realizada a mineração de áreas em países subdesenvolvido com a probabilidade de epidemia de cólera. O processo realiza a extração de regras de associação espaço-temporais sobre séries temporais de imagens discretizadas de satélite. Essa abordagem tem a limitação de não considerar dados semânticos que podem enriquecer as informações sobre o domínio analisado.

Ruiz e Casillas (2018) apresentam uma abordagem para descobrir relações que continuamente se adaptam para o domínio de objetos móveis. Essa abordagem utilizada mineração de regras *fuzzy* que minimiza os impactos da discretização para o domínio de imagens. Como ponto negativo, essa abordagem não leva em consideração as restrições espaço-temporais necessárias para o domínio de séries de imagens de satélite.

3.4 Considerações Finais

A mineração de dados aplicada a imagens é um processo bastante complexo devido à necessidade do pré-processamento das imagens e extração de características das mesmas. Além disso, é bastante comum imagens apresentarem descrições textuais com informações semânticas que devem ser consideradas no processo de mineração para enriquecer seus resultados. Por fim, neste capítulo, foi apresentado o estado da arte para mineração de imagens salientando semelhanças, diferenças e limitações dos trabalhos mais correlatos ao presente trabalho de pesquisa.

4 Classificador Associativo para Regras de Associação

Classificação associativa é uma área da mineração de dados que combina a tarefa de extração de regras de associação com a tarefa de classificação com o objetivo de criar um modelo de aprendizagem preditivo (THABTAH, 2007). A utilização de regras de associação para a geração de um modelo de aprendizagem, para a classificação associativa, surgiu em Ma e Liu (1998). Desde então, diversos trabalhos utilizam essa técnica em diferentes domínios de dados como Kobylinski e Walczak (2006), Mangalampalli e Pudi (2011) e Dua, Singh e Thompson (2009). Nesses trabalhos, a classificação associativa apresenta resultados competitivos quando comparados a algoritmos de árvore de decisão e abordagens probabilísticas.

Classificadores associativos são muitas vezes utilizados em tarefas de mineração de dados cujo conjunto é composto por imagens, pois, para o caso de mineração de imagens, a classificação associativa facilita a visualização dos resultados e permite a classificação preditiva de novas imagens: permite a classificação de imagens que não fizeram parte da extração das regras (SHEELA; SHANTHI, 2009). Exemplos desses trabalhos podem ser vistos na Seção 4.3.

No domínio de imagens, a classificação associativa é dividida em duas partes: geração do modelo de aprendizagem e classificação das imagens (GARCIA-FLORIANO *et al.*, 2017; NEDJAH *et al.*, 2017). Na geração de um modelo de aprendizagem, as imagens são pré-processadas e submetidas a um extrator de características. Seus vetores de características passam pela etapa de extração de regras de associação, como apresentado no Capítulo 2. As regras geradas, geralmente, em um grande volume, são então utilizadas para a geração do modelo de aprendizagem por meio da generalização das regras de associação encontradas.

Quando uma imagem é submetida ao modelo de aprendizagem, também chamado de classificador, é gerada uma classificação para essa imagem. Uma nova imagem é submetida ao classificador e, se essa imagem estiver de acordo com o antecedente de uma ou mais regras extraídas, o consequente das mesmas é retornado como classe da imagem.

Este capítulo se organiza da seguinte forma: Na Seção 4.1, a classificação associativa é formalizada; Na Seção 4.2, é apresentado um classificador associativo baseado em votos; Na Seção 4.3, é apresentado o estado da arte, e; Na Seção 4.4, são apresentadas as considerações finais.

4.1 Formalização da Classificação Associativa

Para a geração do modelo de aprendizagem são utilizadas regras de associação nas quais $r : A \rightarrow B$ na qual o conseqüente, *itemset* B , é uma classe (também chamada de *label*) (MANIKONDA; MANGALAMPALLI; PUDI, 2010). Por exemplo, $r : cor = vermelho, forma = espicular \rightarrow intensidade = alta$, então $intensidade = alta$ é uma classe associada a característica visual $cor = vermelho, forma = espicular$.

A definição do problema para a classificação associativa: seja T um conjunto treinamento composto por m atributos distintos: A_1, A_2, \dots, A_m e seja C uma lista de classes. Uma instância de treinamento em T é descrita como a combinação dos valores a_{ij} dos atributos A_i e uma classe c_k . Um item j é denotado pela tupla $\langle (A_i, a_{ij}) \rangle$ formada por um atributo e um valor. Um *itemset* ι é um conjunto disjuncto de itens. Uma regra é definida por um *itemset* ι e uma classe c . Uma regra do classificador associativo é formada por um conjunto de *itemsets* que são associados a mesma classe. Em suma, uma classificação associativa pode ser vista como uma função $h(\iota) \rightarrow c$ na qual um *itemset* ι sempre se relacionará com a classe c , porém, a classe c pode se relacionar com um ou mais *itemsets* distintos.

4.2 Classificador Associativo Baseado em Voto

O processo de classificação associativa baseada em voto aplicado ao domínio de imagens consiste em associar características visuais da imagem a uma classe (DESHMUKH; BHOSLE, 2016b).

Formalmente, seja R um conjunto de regras de associação tal que $r \in R$ seja uma regra de associação com o formato $A_r \rightarrow B_r$. O *itemset* A_r é composto por características visuais frequentes extraídas do conjunto de imagens e o *itemset* B_r é uma classificação que frequentemente é associada as características A .

Dada uma nova imagem i , é contabilizado um voto para cada classe B_r cujas características A_r estejam presentes na imagem. Assim, para cada imagem i , é contabilizado o número de votos de cada classe. A classe retornada para a imagem i é aquela que possuir maior número de votos.

De uma maneira geral, a classificação associativa baseada em votos para imagens visa mapear as características visuais da imagem e sua classificação mais frequente. Daí a imposição da restrição que A_r deve ser composto de apenas por características visuais.

Um exemplo de algoritmo para criação de um modelo de aprendizagem para a classificação associativa baseado em voto é apresentado no Algoritmo 4.

No Algoritmo 4, a entrada é um conjunto de regras do tipo *antecedente* \rightarrow

consequente no qual o *antecedente* sempre será um conjunto de características visuais e o *consequente* um conjunto de classes semânticas associadas as características visuais.

O objetivo desse algoritmo é a verificação de quantas vezes as características no *antecedente* aparece junto ao mesmo *consequente* (linhas 3 à 10). Cada aparição recebe um voto (linha 8), por isso o nome classificador associativo baseado em voto. Por fim, as associações menos votadas de cada uma das características são removidas, como mostrado nas linhas 11 à 13. As classes associadas à característica visual são armazenados em uma estrutura de dados em formato de dicionário chamada de *mapa*.

Algoritmo 4: Geração de um modelo de aprendizagem para a classificação associativa baseado em voto aplicada ao domínio de imagens.

Entrada: R : conjunto de regras de associação.

Saída: $mapa$: associação característica visual e classe.

```

1 início
2    $mapa \leftarrow \emptyset$  ;
3   para cada  $r \in R$  faça
4     para cada item  $b \in consequente$  faça
5       se  $r.antecedente, b \notin mapa$  então
6          $mapa_b^{r.antecedente()} \leftarrow 0$  ;
7       fim
8        $mapa_b^{r.antecedente()} \leftarrow mapa_b^{r.consequente()} + 1$  ;
9     fim
10  fim
11  para cada  $i \in mapa^i$  faça
12    remove( $b \in mapa_b^i \mid b \neq max(mapa^i)$ ) ;
13  fim
14 fim
```

Fonte: o próprio autor.

Para a classificação de uma nova imagem i , são retornados os consequentes associados às características presentes no *mapa*.

No domínio de imagem o retorno da classificação baseada em votos é bastante útil, pois é possível apresentar as classes que uma determinada imagem possui e atribuir maior semântica à análise dos dados extraídos.

4.3 Estado da Arte em Classificação Associativa para Regras de Associação

Alizadehsani *et al.* (2016) apresentam um método para a classificação de doenças cardiovasculares. O modelo classificativo é construído a partir de um conjunto de regras de associação que por sua vez é extraído de um conjunto de imagens. Esse modelo é capaz de

predizer as doenças coronárias apontando quais artérias possuem uma circulação irregular de sangue. Para tanto, o modelo considera características espaciais das coordenadas da imagem: ele é capaz de determinar o que é uma artéria e quais as artérias existentes na imagem que estão comprometidas. A classificação, no entanto, não considera as características temporais dos exames, por exemplo, se exames anteriores do mesmo paciente apresentaram algum tipo de anomalia.

Jiménez-Hernández *et al.* (2017) apresentam uma abordagem para a classificação de padrões em conjuntos de imagens baseado em memória associativa. A abordagem traça a relação entre os dados de entradas e os dados de saídas, modelando o conjunto de dados como um problema não linear e resultando em uma matriz multidimensional de relações. Tal abordagem não considera características espaço-temporais da imagem, nem dados semânticos que enriquecem as imagens.

Ribeiro *et al.* (2008b) apresentam um método chamado IDEA para dar assistência a diagnósticos médicos com base em conjuntos de imagens médicas. IDEA utiliza o algoritmo Omega para a discretização das características das imagens e o algoritmo Apriori para a extração das regras de associação. O diferencial do IDEA é que esse sugere múltiplos diagnósticos ordenados pela medida de qualidade. Ribeiro *et al.* (2008a) complementa o trabalho anterior apresentando uma etapa de retro-alimentação: os diagnósticos sugeridos e aceitos pelos médicos retro-alimentam o processo de sugestão de diagnóstico tornando-o assim, mais assertivo. Os algoritmos propostos são PreSAGE e HiCARE. PreSAGE combina em um único passo características selecionadas e discretização, enquanto HiCARE é o algoritmo minerado que combina múltiplas palavras-chaves de diagnósticos anteriores a imagem analisada. Ribeiro *et al.* (2009) utilizam as ideias anteriores para dar suporte a dois tipos de sistemas médicos: *Content-based Image Retrieval* (CBIR) e *Computer-Aided Diagnosis* (CAD). CBIR utiliza-se do Omega para aumentar a precisão na busca de imagens similares e o CAD utiliza o IDEA para a extração das regras de associação e o sistema de retro-alimentação de diagnósticos. Ambos métodos permitem a classificação da imagem analisada auxiliando no diagnóstico de doenças mamárias. Os métodos apresentados não levam em consideração características espaciais que as imagens podem apresentar, desta forma, sendo bastante específico para o domínio de imagens médicas. Ademais, o tipo da regra extraída é uma regra clássica, sendo que o *antecedente* limita-se somente a características visuais.

Com base no trabalho de Ribeiro *et al.* (2009) e Watanabe *et al.* (2010b), é apresentado um método para a classificação de mamografias com base em regras de associação, utilizando um classificador por voto. O processo de mineração extrai regras continuamente melhorando o processo com base na retro-alimentação dos diagnósticos. O classificador associativo é utilizado para determinar em qual caso a característica da imagem se enquadra. Watanabe *et al.* (2010a) apresentam o mesmo método para a classificação de

mamografias, porém, utilizando um extrator de regras diferente, o SACMiner. SACMiner é um minerador de regras de associação estatísticas: trabalha com atributos contínuos evitando inconsistências oriundas do processo de discretização dos vetores de características das imagens. Watanabe *et al.* (2012) complementam o método anterior com uma etapa de pré-processamento que seleciona as características mais relevantes para o processo de mineração, reduzindo o conjunto de regras de saída e evidenciando os resultados de maior relevância. Essas abordagens não levam em consideração as características espaciais dos dados, por exemplo, quando um vetor de características de uma região de anomalia influencia no desenvolvimento de uma região vizinha.

4.4 Considerações Finais

A classificação associativa é um processo que combina as tarefas de extração de regras de associação e a classificação. Essa combinação objetiva a geração de um modelo de aprendizagem preditivo com base nas regras de associação extraídas. No domínio de mineração de regras de associação de imagens, a classificação associativa é muitas vezes utilizada como etapa de pós-processamento das regras para a predição. Neste capítulo, foi atribuído foco na classificação associativa baseada em voto que tende associar padrões de antecedente aos consequentes mais votados. Esse tipo de classificador é bastante utilizado no domínio de imagens. Por fim, esse capítulo é finalizado apresentando o estado da arte para a classificação associativa.

A classificação de associativa de regras de associação é uma etapa importante deste projeto de pesquisa pois tem como principal objetivo facilitar a visualização em entendimento de padrões por meio da geração de um modelo de aprendizagem capaz de prever a classificação novas imagens de um domínio com um auto nível de precisão. No caso dos dados solares, novas imagens são geradas em um curto intervalo de tempo (muitas vezes menor que um dia), essa abordagem permite que seja feita a classificação preditiva de uma nova imagem de forma rápida facilitando assim o processo de predição do clima espacial.

Parte II

Desenvolvimento

5 O Método Proposto - *SolarMiner*

Neste trabalho, é proposto, desenvolvido e validado um método para realizar a extração de Regras de Associação Espaço-Temporais Temáticas em imagens e dados. Atualmente, as técnicas existentes na literatura são limitadas para a descoberta de regras espaço-temporais temáticas. Além disso, ao se trabalhar no domínio de Séries Temporais de Imagens de Satélite (SITS), os trabalhos existentes, geralmente, visam à descoberta de padrões visuais que representam a evolução de uma área solar (LI *et al.*, 2016, Seção 2). Tais abordagens possuem como limitação a impossibilidade de responder perguntas como: "Qual será a possível configuração magnética das manchas solares após três dias?" e "Qual é o relacionamento entre a ocorrência de manchas solares", devido ao fato de não considerarem tanto relação com eventos e a sua evolução no decorrer do tempo. No sentido de superar tais limitações, este trabalho visa encontrar esses tipos de padrões que levam em consideração as relações entre os eventos solares (manchas) e características espaço-temporais.

SITS são dados complexos, além da complexidade de processamento da imagem para a realização da mineração de dados, possuem características espaço-temporais. Associados às imagens há mais dados, que chamamos de dados semânticos, contendo informações como intensidade da explosão, quantidade de radiação liberada etc. Esses dados são levados em consideração pelo algoritmo de mineração.

Cada imagem solar pode apresentar nenhuma, uma, ou até mesmo muitas manchas solares que possivelmente podem evoluir para uma explosão solar. Assim, além dos dados semânticos, existem os dados visuais contidos na própria imagem, tais como intensidade das cores, contraste, etc. As imagens passam por um processo de extração de características. Com o processo de extração de características, são obtidos vetores de características, que descrevem as imagens intrinsecamente. Neste trabalho, estes vetores são de suma importância para a geração das regras de associação.

Este capítulo é dividido da seguinte maneira: na Seção 5.1, é apresentada a formalização da proposta; e, na Seção 5.2, são apresentadas as considerações finais.

5.1 Formalização da Proposta

Neste trabalho, é proposto o método *SolarMiner*, para a extração de conhecimento a partir de dados solares. O *SolarMiner* tem como principal componente o algoritmo *Miner of Thematic Spatio-temporal Associations for Images* (MiTSAI). O MiTSAI é um algoritmo minerador de Regras de Associação Espaço-Temporais Temáticas para Imagens.

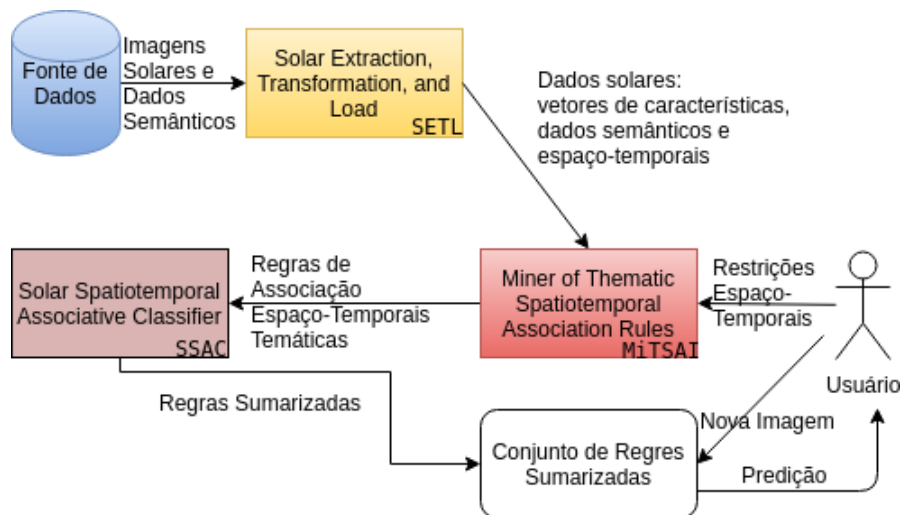
O resultado do processo de mineração do MiTSAI é um conjunto de regras de associação espaço-temporais temáticas que representam o comportamento comum dos dados contidos na base.

O método *SolarMiner* é projetado para lidar com características do domínio espaço-temporal: os dados se organizam pela posição na qual foram encontrados e pelo momento da coleta. Essas características são importantes para o domínio de SITS solar. As manchas solares podem ser ordenadas tanto pelo momento que foram capturadas quanto pela coordenada solar.

Além das informações espaço-temporais, as imagens estão associadas a dados semânticos que contém informações como intensidade da explosão solar, níveis de radiação, clima espacial etc. Essas informações também são de suma importância para a extração das regras de associação e são consideradas pelo método proposto.

Além desses dados, a própria imagem possui informações como contorno, geometria, coloração, brilho etc. Essas informações carregam grande relevância para o processo de extração de conhecimento significativo nesse domínio. Devido a isso, a maior parte do esforço de processamento dos dados é concentrada para a extração das características das imagens.

Figura 6 – Método proposto apresentado em alto nível.



Fonte: o próprio autor.

Com o objetivo de considerar as características do domínio solar, é proposto um método de extração de conhecimento (KDD) chamado *SolarMiner*. O fluxograma do SolarMiner é apresentado na Figura 6.

O *SolarMiner* tem como entrada conjuntos de imagens espaço-temporais: cada imagem pode ser organizada pelo momento de sua captura. Além disso, a imagem pode conter N manchas solares em coordenadas solares diferentes, sendo $N \geq 0$, e também

pode conter uma ou mais explosões solares.

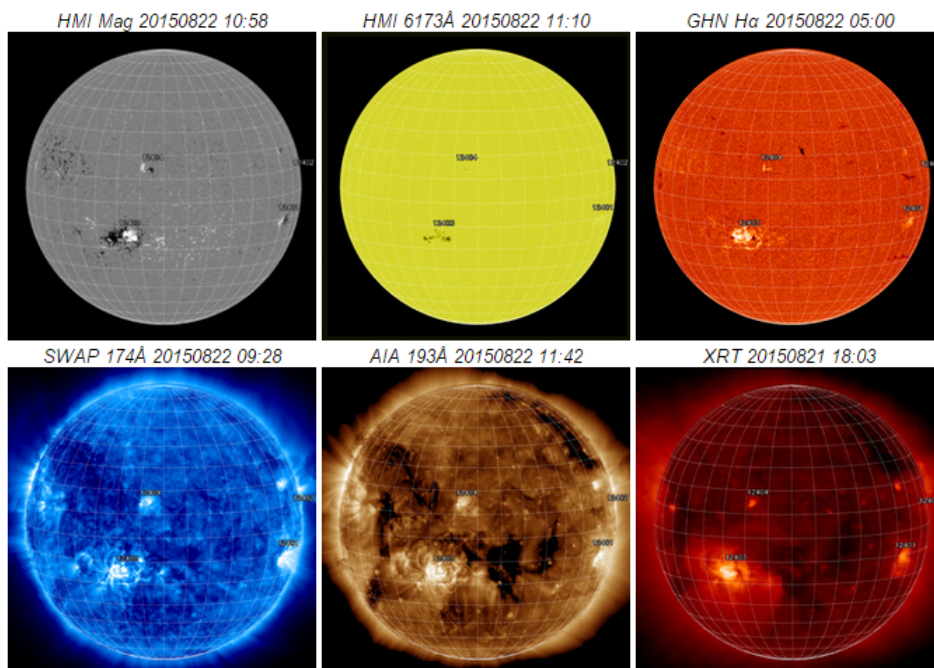
O *SolarMiner* tem como resultado um modelo preditivo utilizado na análise do comportamento solar.

Na Subseção 5.1.1, são apresentados os dados oriundos do Satélite NOAA (2015) e, como apresentado na Figura 6, o método *SolarMiner* proposto que se divide em três etapas: o *Processo SETL* para a extração e transformação dos dados solares, apresentado na Subseção 5.1.2; o *MiTSAI*, minerador para a extração de Regras Espaço-Temporais Temáticas em Imagens e seus dados semânticos, apresentado na Subseção 5.1.3, e; o SSAC, um classificador associativo que sumariza as regras mineradas pelo MiTSAI para apoiar o processo de análise, e usa as regras sumarizadas para a predição do comportamento futuro dos dados, apresentado na Subseção 5.1.4.

5.1.1 Caracterização da Fonte dos Dados

Para a validação deste trabalho e verificação das hipóteses são utilizados os dados do Satélite NOAA (2015). Na Figura 7, é apresentado um exemplo com seis imagens coletadas do sol no mesmo momento, cada imagem é coletada em um comprimento de onda diferente para destacar diferentes camadas da atmosfera solar (ZHAO *et al.*, 2010; ZHANG; LIU; WANG, 2011; BOBRA; COUVIDAT, 2015). As imagens estão disponíveis nessa base de dados.

Figura 7 – Essa figura é um *print screen* de NOAA (2015). Como pode ser visto, há 6 imagens disponíveis de 4 eventos (identificados por números na imagem). Essas imagens são dos tipos: HMI Mag, HMI, GHN H α , SWAP, AIA e XRT.



Fonte: figura extraída de (NOAA, 2015) em 22 de Agosto de 2015.

Os diferentes comprimentos de ondas e os respectivos aparelhos são:

- HMI Mag:** magnetograma do disco solar da região da fotosfera, obtido por meio do instrumento *Helioseismic and Magnetic Imager* (HMI) cujo comprimento de onda é de 617.3;
- HMI:** imagem visível da fotosfera do disco solar obtida por meio do instrumento HMI;
- GHN H α :** imagem em H- α do disco solar da região da cromosfera, obtida por meio de um aparelho cujo comprimento de onda é de 656.3 nanômetros;
- SWAP:** imagem em ultravioleta do disco solar obtida por meio do instrumento *Sun Watcher using Aps (Active pixel system) detectors and image Processing* (SWAP) cujo comprimento é de 17.4 nanômetros;
- AIA:** imagem do disco solar obtida usando raios-X por meio do instrumento *Atmospheric Imaging Assembly* (AIA) cujos comprimentos são de ondas visíveis e ultra-violeta, e;
- XRT:** imagem obtida com o instrumento *X-Ray Telescope* (XRT) cujo comprimento de onda varia entre 0.6 e 6.0 nanômetros.

As imagens são complementadas pela informação presente nos dados semânticos também disponíveis em NOAA (2015). Um exemplo dessa informação que complementa as imagens da Figura 7 está presente na Tabela 1.

Os atributos apresentados pela Tabela 1 (colunas) significam:

- *Número de Grupo:* é um identificador da mancha solar achado na imagem para a ocorrência de um evento solar;
- *Localização:* identifica as coordenadas do centro do grupo;
- *Classe de Robustez (Hale Class):* classifica as características magnéticas das manchas solares de acordo com as regras definidas por *Mount Wilson Observatory* na Califórnia, USA, em:

α : grupo unipolar de manchas solares, podendo ser positiva ou negativa;

β : grupo de manchas solares com polaridades magnéticas positiva e negativa e com divisão distinta entre elas;

γ : região complexa de atividade com polaridade positiva e negativa, irregularmente distribuída;

$\beta - \gamma$: grupo de manchas solares com polaridade magnética positiva e negativa, cuja superfície é completa a ponto de não ser possível desenhar uma linha entre polaridades opostas;

Tabela 1 – Exemplo de arquivo descritor que acompanha as imagens. Essas medições acompanham o conjunto de imagens na Figura 7.

Número de Grupo	Localização	Classe de Robustez	Classe McIntosh	Área	Número do lugar	Histórico Flares
12401	S11W65 (844",-227")	α/β	Hrx / Dao	0010/0060	02/06	-
12403	S14E16 (-254",-335")	$\beta\iota/\beta\iota$	Dkc/Dkc	0350/0330	43/29	C2.0(00:23) C5.1(06:27) C6.7(08:34) C5.1(09:09) C6.3(10:21) M1.2(06:39) /C1.2(08:19) C1.2(14:33) C1.3(23:34) M1.2(01:56) M1.4(09:34) M1.1(19:10)
12404	N14E06 (-96",117")	β/β	Cro/Cao	0020/0020	03/03	-
12402	N05W90 (944",82")					-

Fonte: extraído de NOAA (2015) em 22 de Agosto de 2015.

δ : indica que as regiões são separadas por menos de 2 graus com uma penumbra de polaridade oposta;

$\beta - \delta$: grupo de manchas solares cuja classificação magnética é β , porém, com uma ou mais regiões δ ;

$\beta - \gamma - \delta$: grupo de manchas solares cuja classificação magnética é $\beta - \gamma$, porém, com uma ou mais regiões δ , e;

$\gamma - \delta$: grupo de manchas solares cuja classificação magnética é γ , porém, com uma ou mais regiões δ .

- *Classe de McIntosh*: é baseada em uma classificação para manchas solares proposta por McIntosh (1990). Na Figura 8, a classificação McIntosh é visualmente explicada. Sua forma geral é "Zpc", sendo 'Z' a classe Zurich, 'p' descreve a penumbra da mancha principal e 'c' descreve a distribuição da mancha no interior de um grupo.

Z: A - Mancha pequena única e unipolar, tanto na formação quanto no término;

B - Grupo bipolar sem penumbra entre as manchas do grupo;

C - Grupo bipolar com penumbra;

D - Grupo bipolar com penumbra nos dois finais do grupo. Grupo cuja extensão na longitude não excede 10 graus;

E - Grupo bipolar com penumbra nos dois finais do grupo. Grupo cuja extensão na longitude está entre 10 e 15 graus;

F - Grupo bipolar alongado com penumbra nos dois finais do grupo. Grupo cuja extensão na longitude excede 15 graus, e;

H - Unipolar com penumbra.

p: x - sem penumbra;

r - penumbra rudimentar em volta da mancha maior;

s - penumbra pequena e simétrica. A mancha maior possui uma penumbra escura; o diâmetro norte-sul que cruza a penumbra é menor que 2,5 graus;

a - pequena e assimétrica. A maior mancha possui uma penumbra irregular delimitando-a;

h - grande e simétrica. A mancha maior possui uma penumbra escura; o diâmetro norte-sul que cruza a penumbra é maior que 2,5 graus, e;

k - grande e assimétrica. A mancha maior possui uma penumbra escura; o diâmetro norte-sul que cruza a penumbra é maior que 2,5 graus.

c: x - não definida para grupos unipolar;

o - aberto, interior da mancha bem pequeno;

i - intermediário, muitas manchas sem líder e seguidores, e;

c - compacto, área entre a líder e seguidores é populada por manchas fortes.

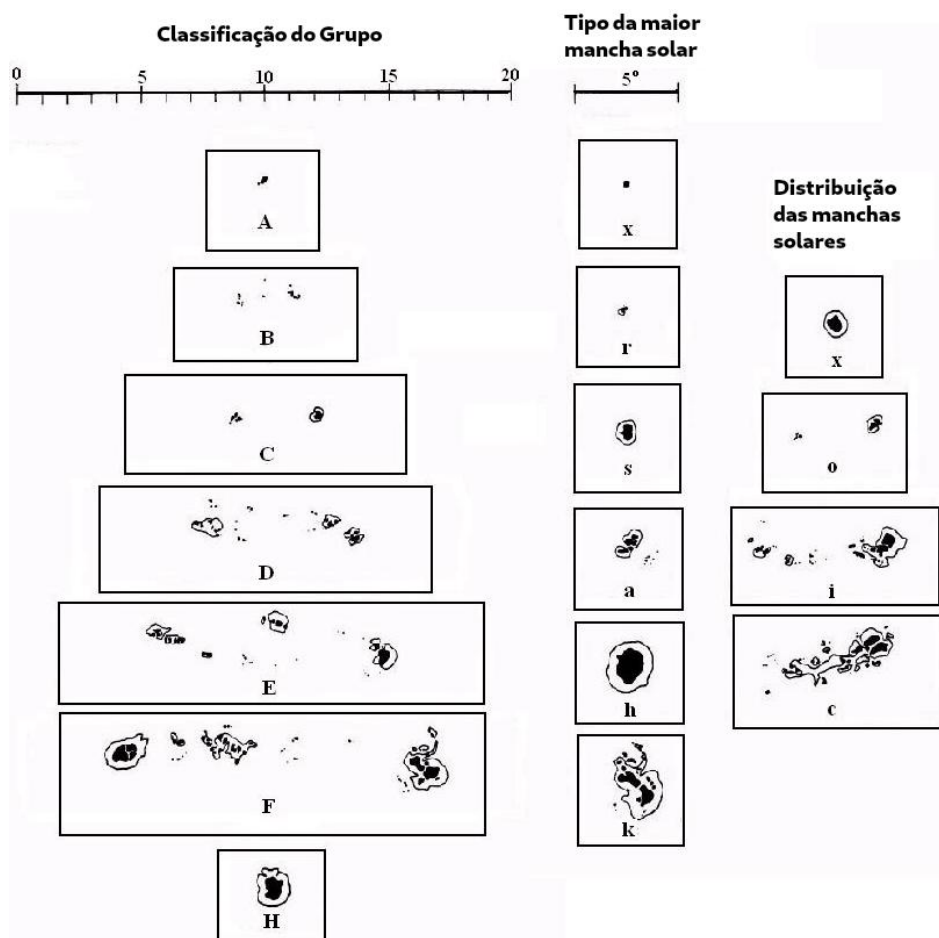
- *Área*: é a área do grupo;
- *Número do Lugar*: é um identificador para o local da ocorrência;
- *Histórico*: Eventos de uma mancha associados a mancha que está sendo analisada com possíveis ocorrências na própria mancha no passado. Sempre que uma mancha é detectada no início de um período solar, ela recebe um novo identificador, porém, devido ao movimento de rotação do sol, essa mancha pode já ter sido previamente detectada em outros períodos solares, assim é possível que uma mancha possua mais do que um histórico.

5.1.2 O processo SETL

Neste projeto, foi desenvolvido o processo de ETL (*Extraction, Transformation, and Load*) para dados solares, o qual denominamos SETL (*Solar ETL*).

As imagens são extraídas de NOAA (2015) junto com seus dados semânticos. Na Figura 9, é apresentado um exemplo dos dados coletados nesse processo, nos quais são destacados os dados espaciais, os dados temporais, as imagens do disco solar extraídas em diferentes comprimentos de ondas e os dados semânticos. As imagens (vetores de

Figura 8 – Classificação McIntosh visualmente explicada.



Fonte: figura adaptada de spaceweatherlive.com (2018).

características) junto com os dados semânticos compõem os valores dos atributos temáticos que são usados na mineração das regras de associação.

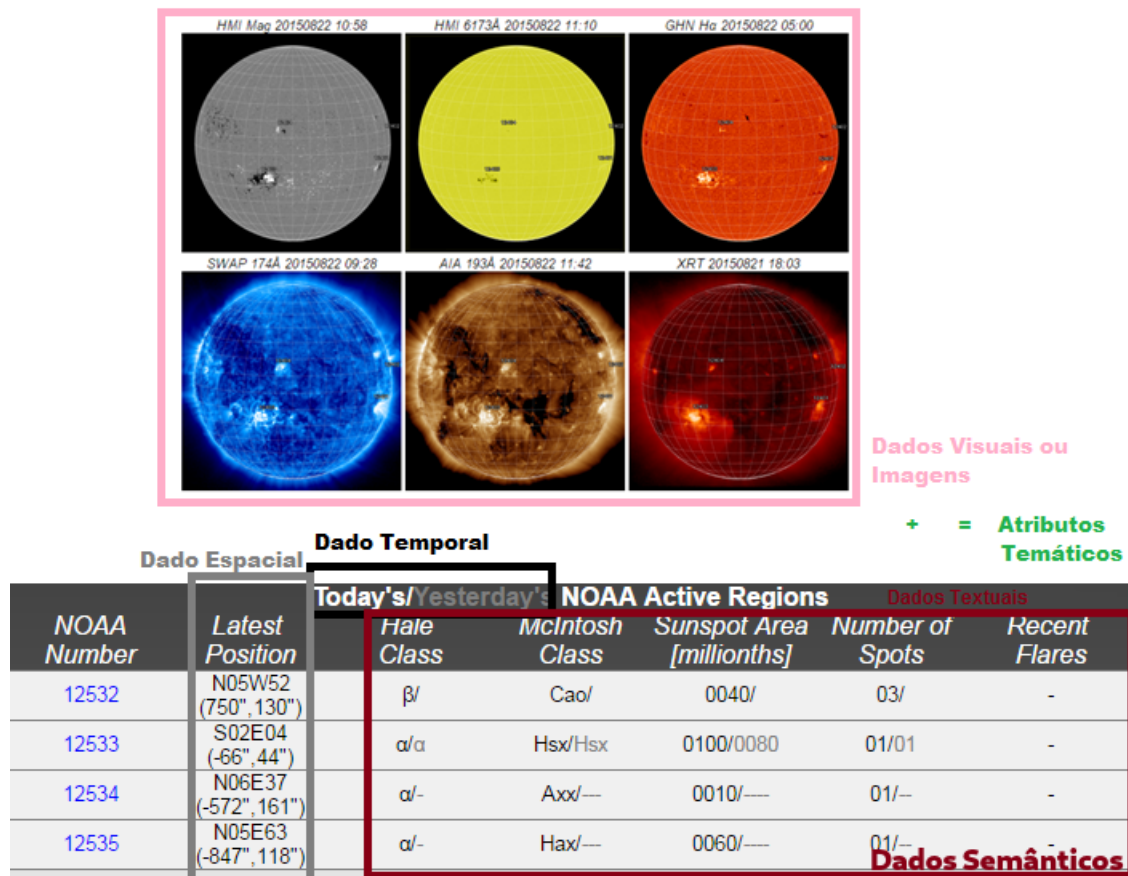
O processo SETL foi desenvolvido neste projeto para coletar, integrar e tratar os dados solares. Assim, os dados solares são preparados e ficam disponíveis para o processo de mineração de regras de associação.

O SETL desenvolvido é dividido em quatro processos: Extrator de Dados, Extrator de Características, Discretizador e Integrador. A interação entre esses processos é apresentada na Figura 10.

Extrator de Dados

Um Extrator de Dados pode também ser chamado de *Crawler* (Disponível em (SILVEIRA-JUNIOR; BERTONHA, 2015, último acesso 9 de fevereiro de 2019)), e é um *software* que atua como um "robô" cujas atividades são navegar entre as páginas e armazenar as informações coletadas relevantes das mesmas. O Extrator de Dados extrai os dados

Figura 9 – Um exemplo de registro dos dados que apresenta o que são os dados espaço-temporais, dados semânticos, as imagens e os atributos temáticos.



Fonte: figura adaptada de NOAA (2015).

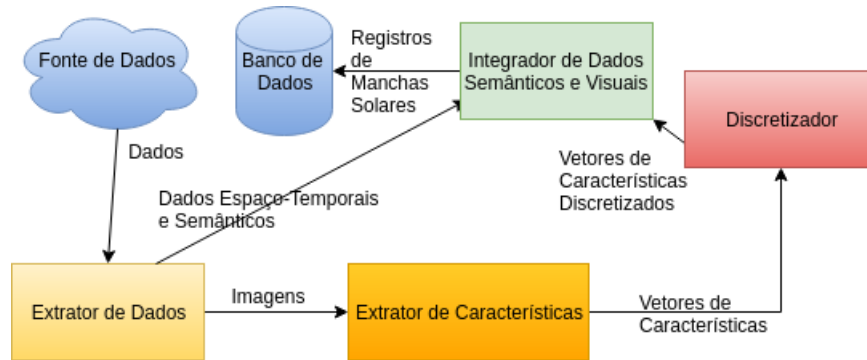
semânticos, espaço-temporais e imagens do site NOAA (2015). Os dados semânticos junto com os dados espaço-temporais são convertidos para o formato JSON, como exemplificado na Figura 11, e são armazenados em um Sistema de Gerenciamento de Banco de Dados (SGBD).

O formato JSON é utilizado, pois, ao fazer a extração dos dados, o Extrator de Dados armazena a informação extraída em um Dicionário, que é uma estrutura de dados do tipo chave-valor. Essa estrutura é naturalmente convertida para arquivos JSON, que são nativamente aceitos por SGBDs NoSQL, como o MongoDB, e SGBDs relacionais, como o Postgres. Conforme ilustrado na Figura 10, as imagens coletadas seguem para o processo Extrator de Características.

Extrator de Características

O processo Extrator de Características é aplicado para cada imagem coletada com a objetivo de realizar a extração de seus vetores de características. O Extrator de Características possibilita utilizar o SURF (BAY; TUYTELAARS; GOOL, 2006), Haralick

Figura 10 – O processo SETL aplicado para o domínio NOAA (2015).



Fonte: o próprio autor.

Figura 11 – Exemplo de Registro Solar em formato JSON.

```
[{
  "date": "20150825",
  "area": ["0930/0760", "0020/0030"],
  "number_of_spots": ["47/66", "04/07"],
  "hale_class": ["\u03b2\u03b3\u03b4/\u03b2\u03b3\u03b4", "\u03b2/\u03b2"],
  "mcintosh_class": ["Fkc/Ekc", "Cro/Cro"],
  "number": ["12403", "12404"],
  "image_urls": ["http://www.solarmonitor.org/data/2015/08/25/pngs/hxrt/hxrt_filter_fd_20150822_061940.png"],
  "flares": ["C1.1(04:14),C4.3(06:18),C2.3(07:58),C2.3(11:55),C1.3(12:33),C1.3(17:16),C1.3(19:18),C1.2(19:49), C3.0(22:40),C1.1(23:06)", ""],
  "location": ["S15W27 (417\u00b0,-345\u00b0)", "N14W40 (593\u00b0,141\u00b0)"],
  "images": [{"url": "http://www.solarmonitor.org/data/2015/08/25/pngs/hxrt/hxrt_filter_fd_20150822_061940.png", "path": "full/ede7ba3d08e92bbc62cd4e78cda875612054c450.jpg", "checksum": "2aee212047ec8914551702660ee08fb6"}],
  "type": "XRT and NOAA Active Regions"},
  ...]
```

Fonte: o próprio autor.

e Histograma. Como resultado do processo Extrator de Características, são obtidos os vetores de características que representam as características visuais da imagem durante o processo de mineração de regras de associação.

Pela aplicação do SURF, é encontrado um vetor de característica para cada mancha solar da imagem. Cada mancha solar é delimitada na primeira etapa do SURF, como apresentado na Seção 3.2.2. Para os outros extratores de características, Haralick e Histograma, é necessária a segmentação da imagem para determinar a área de interesse (área onde cada mancha solar está localizada), para tanto são utilizados os dados espaço-

Figura 12 – Vetor de características antes e depois de ser processado pelo Omega.

<0.15	0.49	0.15	2.423	1.98	...>
↓	↓	↓	↓	↓	
<[0.1-0.5)	[0.1-0.5)	[0.1-0.5)	[2.4-6.2)	[0.6-2.3)	...>

Fonte: o próprio autor.

temporais da mancha solar. Como as imagens possuem sempre o mesmo tamanho e qualidade, é possível traduzir as coordenadas solares para coordenadas da imagem. Desta forma, é feita a segmentação de cada imagem e ao segmento das mesmas são aplicadas a extração de característica na qual é utilizado o descritor de Haralick ou Histograma.

É comum encontrar, na literatura de domínio, processos que não fazem a segmentação da imagem solar, porém, como indicado pelo especialista de domínio, é importante poder analisar cada mancha individualmente e entender seu comportamento.

Após essa etapa, é feita a discretização dos vetores de características.

Discretizador

O processo Discretizador, utiliza o Algoritmo Omega (RIBEIRO; TRAINA; TRAINA JR., 2008) para realizar a discretização dos vetores de características. Isso se deve ao fato do Omega ser projetado para gerar um número reduzido de intervalos, o qual reduz a diversidade dos itens na base e aumenta a frequência de ocorrência dos mesmos. Um exemplo da aplicação do Omega pode ser observado na Figura 12.

O Omega recebe do Extrator de Característica vetores de característica da cada uma das imagens e os processa desta forma são geradas representações discretas para cada elemento do vetor, é respeitada a ordem dos elementos internos do vetor nesse processo. O resultado é um vetor composto por intervalos discretos.

Após essa etapa, é feita a integração dos dados semânticos, vetores de características discretos e dados espaço-temporais.

Integrador

O processo Integrador é o processo final do ETL. Os vetores de características discretos são fundidos aos seus dados semânticos e aos seus dados espaço-temporais e é feita a **inserção** desses dados no SGBD. Desta forma, os dados são disponibilizados para o algoritmo minerador.

Na Figura 13, é apresentado um exemplo do registro completo armazenado no banco. Nesse exemplo há duas manchas solares: mancha número 12403 e mancha número 12404. Cada mancha possui suas características espaciais: mancha 12403 – área 0930/7560 e localização – S15W27 (417,-345), e; mancha 120404 – área 0020/0030 e locação – N14W40

Figura 13 – Exemplo de Registro Solar em formato JSON com as características discretizadas das imagens e após a remoção de dados que não são utilizados na mineração.

```
[{
  "date": "20150825",
  "area": ["0930/0760", "0020/0030"],
  "number_of_spots": ["47/66", "04/07"],
  "hale_class": ["\u03b2\u03b3\u03b4/\u03b2\u03b3\u03b4", "\u03b2/\u03b2"],
  "mcintosh_class": ["Fkc/Ekc", "Cro/Cro"],
  "number": ["12403", "12404"],
  "location": ["S15W27 (417\",-345\"), "N14W40 (593\",141\" )"],
  "image_1": ["[0.1-0.5) [0.1-0.5) [0.1-0.5) ...",
              "[2.4-6.2) [0.6-2.3) [0.1-0.5) ..."]
  ...
  "image_6": ["[0.5-1.0) [2.4-6.2) [2.4-6.2) ...",
              "[2.4-6.2) [0.6-2.3) [0.6-2.3)"]
}, ...]
```

Fonte: o próprio autor.

(593,141). Cada mancha possui classificações McIntosh e Hale. E cada mancha possui 6 vetores de característica cada vetor é oriundo de uma imagem coletada por um aparelho calibrado para diferentes comprimentos de ondas. Desta forma, a mancha 120403 possui o primeiro vetor de image_1, o primeiro de image_2 até image_6 e a mancha 120403 o segundo vetor de cada um. Diariamente são geradas até seis imagens solares, sendo que existem dias que uma ou mais imagens não estão disponíveis. Quando uma das imagens não está disponível o seu vetor de característica fica vazio.

Discussão

O SETL tem por objetivo extrair os dados solares disponíveis e transformá-los de forma que seja possível a extração de regras de associação espaço-temporais temáticas em um tempo de processamento aceitável. Para isso, o SETL foi projetado para ser flexível a ponto de usar qualquer SGBD, relacional ou NoSQL, essa flexibilidade possibilitou experimentos os quais tem como objetivo a melhora no desempenho do algoritmo como apresentado no Capítulo 6.

Outro ponto importante é a possibilidade de escolher diferentes Extratores de Características, desta forma, foi possível realizar experimentos que determinaram qual dos extratores melhor se adapta aos dados solares e qual extrator implica em um maior precisão no resultado. Desta forma, possibilitou-se a flexibilidade na representação visual das imagens e a flexibilidade para o ganho de desempenho.

5.1.3 Miner of Thematic Spatio-temporal Associations for Images

O *Miner of Thematic Spatio-temporal Associations for Images* (MiTSAI) é um algoritmo de mineração de Regras de Associação Espaço-Temporais Temáticas em um domínio composto por: séries temporais de imagens e séries temporais com informações semânticas.

O MiTSAI combina a estratégia Filtro-e-Refina, baseada em Pillai, Angryk e Aydin (2013), para a extração de padrões espaço-temporais com as otimizações de desempenho do ARMADA (WINARKO; RODDICK, 2007), algoritmo utilizado para extração de regras de associação temporais. O MiTSAI é apresentado no Algoritmo 5.

Algoritmo 5: *Miner of Thematic Spatio-temporal Associations for Images* (MiTSAI).

Entrada: DB : base de dados; $minSup$: valor de suporte mínimo; $minConf$: valor de confiança mínima; d : restrição espacial; t : restrição temporal.

Saída: R : regras espaço-temporais temáticas.

```

1 início
2    $itemsets \leftarrow genItemset(\emptyset, DB, minSup, d)$  ;
3    $R \leftarrow genRegras(itemsets, minConf, t)$  ;
4 fim
```

Fonte: o próprio autor.

O algoritmo MiTSAI recebe como entrada a base de dados espaço-temporal que foi produzida pelo SETL; os valores de suporte e confiança mínimos, e; as restrições espaciais e temporais. O resultado é um conjunto de regras espaço-temporais temáticas que são frequentes na base de dados. MiTSAI é dividido em duas etapas: (i) geração dos *itemsets* espaciais frequentes, (Algoritmo 6) e; (ii) geração das regras de associação espaço-temporais temáticas (Algoritmo 7), linhas 2 e 3 respectivamente.

No Algoritmo 6, é apresentada a função para a geração dos *itemsets* espaciais frequentes. Essa função recebe um *itemset* semente que pode ser um conjunto vazio (como acontece na linha 2 do Algoritmo 5); a base de dados, que na primeira iteração é a projeção inicial da base (linha 2 do Algoritmo 5); o valor de suporte mínimo para considerar um *itemset* espacial frequente, e; a restrição espacial que precisa ser respeitada pelos *itemsets* encontrados. Como resultado, é retornado o conjunto de *itemsets* espaciais frequentes. Na linha 2 do Algoritmo 6, o conjunto de resultados R é inicializado. Na linha 3, fi recebe o conjunto de itens frequentes na projeção da base de dados DB (atributo de entrada da função).

Na linha 4, é iniciado um laço de repetição que para cada item i frequente é gerado um *itemset* candidato si , combinando todos os *itemsets* da base com o item i –linha 5.

O valor de suporte do *itemset* si é calculado –linha 6– e caso esse seja frequente (valor de suporte do *itemset* si maior ou igual a $minSup$ –linha 7) é gerada uma projeção

da base de dados db sobre si –linha 8– e si é adicionado ao conjunto de resultados R junto ao resultado da geração de $itemsets$ que contém $itemset$ si (chamada recursiva ao Algoritmo 6, passando si como base e a projeção da base p) –linha 9. A geração de um projeção da base de dados para um $itemset$ consiste na geração de um base de dados composta apenas pelos registros que contém o $itemset$.

Essa base é menor ou igual a original, desta forma, o espaço de busca por novos $itemsets$ na projeção considera menos dados melhorando o desempenho do algoritmo.

Algoritmo 6: MiTSAI – Função $genItemset$.

Entrada: $base$: Itemset; db : projeção da base de dados; $minSup$: valor mínimo de suporte, d : restrição espacial.

Saída: R : $itemsets$ espaciais frequentes.

```

1 início
2    $R \leftarrow \emptyset$  ;
3    $fi \leftarrow itens\ frequentes \in db$  ;
4   para cada item  $i \in fi$  faça
5      $si \leftarrow base \oplus i$  ;
6      $si.suporte \leftarrow \frac{|A \cup B, distancia(A,B) \leq d|}{|DB|}$  ;
7     se  $si.suporte \geq minSup$  então
8        $p \leftarrow projete\ db\ para\ si$ ;
9        $R \leftarrow R \cup \{si\} \cup genItemset(si, p)$  ;
10    fim
11  fim
12 fim

```

Fonte: o próprio autor.

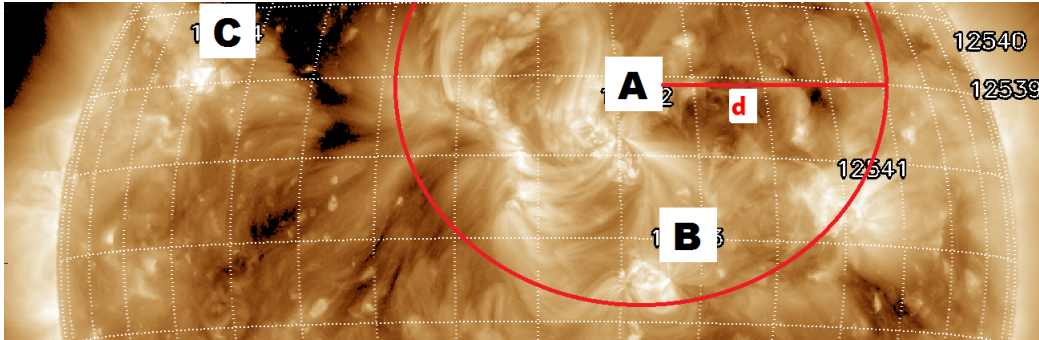
O cálculo de suporte, presente na linha 6, é flexibilizado para considerar a restrição espacial estabelecida pelo usuário. Essa restrição é fornecida em unidades do disco solar. Se o usuário estabelecer que durante o cálculo do suporte de $A \cup B$ o $itemset$ B deve estar a no máximo d unidades de distância do $itemset$ A , então, os que estiverem fora desta restrição não serão levados em consideração pela contagem.

Esse exemplo é ilustrado na Figura 14, a fórmula para o cálculo do suporte não sofre alteração, conforme apresentado na Equação 5.1. A alteração é na contagem de ocorrências dos $itemsets$ para considerar a restrição espacial.

$$suporte(A \cup B) = \frac{|A \cup B, distancia(A, B) \leq d|}{|DB|} \quad (5.1)$$

Pelo fato dos $itemsets$ respeitarem as restrições espaciais, é dado o nome de $itemsets$ espaciais. Esses, se frequentes, são chamados de $itemsets$ espaciais frequentes. O processo continua a iterar e a gerar $itemsets$ espaciais frequentes até não ser mais possível a geração de novos $itemsets$ espaciais frequentes, com um maior número de itens frequentes.

Figura 14 – Cálculo do valor de suporte levando em consideração a restrição espacial.



Fonte: o próprio autor.

Para o exemplo apresentado na Figura 14, o *itemset* C não obedece a restrição espacial imposta. Desta forma, durante o cálculo do suporte de $A \cup C$, esse registro não deve ser levado em consideração.

Algoritmo 7: MiTSAI – Função *genRegras*

Entrada: *itemsets*: conjunto de *itemsets* espaciais frequentes; *minConf*: valor mínimo de confiança; *t*: restrição temporal.

Result: R : Conjunto de regras espaço-temporais temáticas.

```

1 início
2    $R \leftarrow \emptyset$  ;
3   para cada itemset  $A$  e  $B \in \textit{itemsets} \mid A \neq B$  faça
4      $r \leftarrow \textit{regra} : \langle A \rightarrow B \rangle$  ;
5      $r.\textit{confiança} \leftarrow \frac{|A \cup B, \Delta\textit{periodo}(A,B) \leq t|}{|A|}$  ;
6     se  $r.\textit{confiança} \geq \textit{minConf}$  então
7        $R \leftarrow R \cup \{r\}$  ;
8     fim
9   fim
10 fim
```

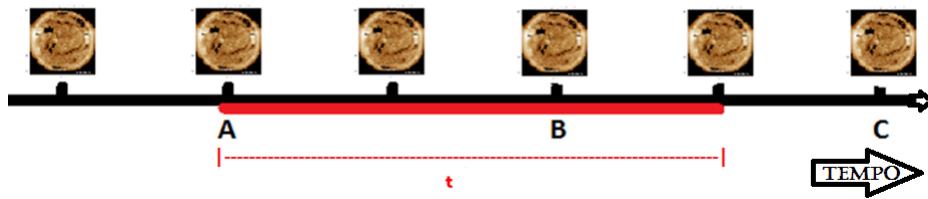
Fonte: o próprio autor.

O conjunto de *itemsets* espaciais frequentes são utilizados para a geração das regras de associação, como é apresentado no Algoritmo 7. Nesta etapa, a maneira como as regras são geradas é diferente da mineração tradicional; os *itemsets* não são subdivididos e recombinaos para calcular o valor de confiança de cada regra. No MiTSAI, é feita a combinação dos *itemsets* espaciais frequentes entre si (linhas 3 à 9), assim, são geradas regras do tipo $r : A \rightarrow B$ (linha 4) e o valor de confiança de cada uma é calculado por meio da consideração da restrição temporal estabelecida pelo usuário (linha 5). Assim, ao se calcular a confiança de uma regra r , conforme Equação 5.2, é afrouxada a contagem $|A \cup B|$ para que o *itemset* B possa estar presente em até d unidades de tempo do *itemset* A .

Na Figura 15, um exemplo é apresentado. A combinação dos *itemsets* é sempre dois a dois e como restrição do domínio os *itemsets* combinados estão relacionados as mesmas manchas solares. Após a combinação é determinado se ambos ocorrem com uma distância temporal aceitável (menor que t unidades de tempo). Se sim, as ocorrências são contadas para se estabelecer a confiança do valor da regra.

$$\text{confiança}(r : A \rightarrow B) = \frac{|A \cup B, \Delta_{\text{periodo}}(A, B) \leq t|}{|A|} \quad (5.2)$$

Figura 15 – Cálculo do valor de confiança levando em consideração a restrição temporal.



Fonte: o próprio autor.

Considere o exemplo da Figura 15, neste exemplo, caso o *itemset* B ocorra em um período que não está dentro do raio de abrangência da restrição temporal (maior que d unidades de tempo), a contagem da ocorrência da regra $A \rightarrow B$ não é considerada para o registro em questão.

O MiTSAI aplica somente as restrições espaciais para o cálculo do suporte do *itemset*, desta forma, é possível a extração de padrões que consideram as relações e interações entre as manchas próximas. As restrições temporais são aplicadas durante o cálculo da confiança da regra, assim, é possível fazer a extração do padrão, a partir da evolução do disco solar.

Um exemplo de uma regra a ser extraída é

$$(\langle [0.01 - 2.1) \dots \rangle, 010/080), \beta/\beta \rightarrow (\langle [2.2 - 2.8) \dots \rangle, 080/100)$$

$$\text{delta}_t(4, \text{dias}), \text{suporte} = 1.3\%, \text{confiança} = 89\%$$

Essa regra significa:

- A ocorrência em conjunto de:
 - O valor da Característica $\langle [0.01 - 2.1) \dots \rangle$ associada a uma mancha solar cuja área é 010/080; e,
 - Uma mancha classificada como β/β obedecendo à restrição de distância;
- Implica na ocorrência de:

- Uma mancha com a Característica $< [2.2 - 2.8) \dots >$ e área de 080/100;
- Sendo que o conseqüente ocorrerá em aproximadamente quatro dias após o antecedente, com suporte de 1.3% e confiança de 89%.

Por fim, todas as regras de associação que obedecem às restrições temporais e tem confiança maior que o mínimo definido pelo usuário formam o conjunto de Regras de Associação Espaço-Temporais Temáticas extraído pelo MiTSAI, i.e., a saída do algoritmo. Esse conjunto é disponibilizado para o Analisador.

Discussão

O MiTSAI preenche uma lacuna identificada na revisão bibliográfica para a extração de regras de associação espaço-temporais temáticas com foco em dados oriundos de satélites (séries de imagens de satélites). O objetivo do MiTSAI é permitir a extração de relações entre eventos que ocorrem ao mesmo tempo tanto quanto extrair relações a partir da evolução desses eventos no decorrer de um período de tempo. Para tanto, foi proposta a divisão da aplicação das restrições espaciais e temporais. A restrição espacial em nível de extração de *itemset* possibilita extrair os eventos frequentes e relaciona-los com eventos frequentes que ocorrem à sua volta (sendo o espaço de busca configurável). A restrição temporal em nível de geração de regras permite extrair a evolução desses eventos no decorrer do tempo.

Em versão anterior do MiTSAI, a geração das regras considerava a imagem como um todo, qualquer mancha em uma imagem poderia ser combinada com qualquer mancha no decorrer de um período. Embora gerasse regras mais flexíveis, tais regras foram consideradas não úteis frente à necessidade, explicitada pelo especialista de domínio, de conhecer o relacionamento entre o estado inicial de uma mancha e sua evolução.

Por esse motivo, a restrição espacial foi adicionada no MiTSAI, desta forma, é feito com que a regra seja mais semântica do ponto de vista do especialista.

A questão do desempenho do algoritmo também foi levada em consideração: por meio da estratégia de dividir e conquistar foi possível paralelizar a extração dos *itemsets* frequentes. Cada chamada à função de extração de *itemsets* frequentes pode ser feita por uma nova *thread* de processamento sem problemas com regiões críticas. O mesmo também é válido para a geração das regras, o processamento do conjunto de *itemsets* espaciais frequentes pode ser processado em paralelo por vários *threads*.

5.1.4 O Solar Spatiotemporal Associative Classifier

O *Solar Spatiotemporal Associative Classifier* (SSAC) é um classificador associativo baseado em voto que sumariza as regras de associação geradas pelo MiTSAI e usa as

regras sumarizadas para prever o comportamento futuro de novas imagens.

O SSAC recebe o conjunto de regras de associação espaço-temporais temáticas geradas pelo MiTSAI e gera uma regra para cada antecedente por meio do agrupamento das regras com o antecedente em comum. Dado uma nova imagem, o SSAC retorna o rótulo que mais frequentemente aparece nas regras sumarizadas associado às características visuais presentes na nova imagem analisada. Assim, dada uma nova imagem, para cada característica visual presente no antecedente das regras sumarizadas é computada a frequência de ocorrência de cada rótulo no consequente. O SSAC retorna o rótulo mais frequente como classe dessa nova imagem.

Algoritmo 8 apresenta o algoritmo SSAC. Na linha 2, *contador* é inicializado como vazio. *contador* é uma estrutura de dados *hashmap* na qual a chave é o vetor de características (representação de uma característica visual) e o valor é outro *hashmap*, chamado sub-*hashmap*. A chave sub-*hashmap* é um rótulo associado frequentemente ao vetor de características e seu valor é uma quintupla: contador, suporte, confiança, Δ tempo e Δ espaço, que são as informações provenientes da regra.

Na linha 3, é executado um laço sobre cada regra no conjunto de entrada (*regras*), cada regra é chamada *r* durante a interação do laço. Na linha 4, *característica* recebe o vetor de característica no antecedente de *r*. Nas linhas 5 e 6, é verificado se para cada *característica* $c \in$ *características* já está em *contador*. Se estiver, linhas 6 e 7, o algoritmo chama *atualiza* para o sub-*hashmap* (também chamado de nó) de *característica*. Caso contrário, linhas 8 e 9, um sub-*hashmap* é criado e inicializado como vazio e *atualiza* é chamado sobre este novo sub-*hashmap*. No final deste laço, todas as regras são processadas e *contador* tem a informação necessária sobre o conjunto *regras*.

Na linha 13, a saída *Map* é inicializada como vazia. Na linha 14, uma regra itera sobre o *contador* recebendo cada sub-*hashmap* (nó), chamado *n*. Na linha 15, *c* é criado por meio da obtenção do rótulo mais votado. Isto é, *c* obtenha a tupla que tenha o maior valor de *contador*, melhor explicado no detalhamento do Algoritmo 9. Esta operação é feita pela função *mostVoted*. Da linha 16 até a linha 19, os valores de suporte, confiança, Δ espaço e Δ tempo são atualizados. Até esse ponto, eles só foram acumulados (pela função *atualiza*) e essas linhas levam a média para isso. Na linha 20, *c* é adicionado à saída *Map*.

O Algoritmo 9 apresenta a função *atualiza*. Recebe nó (sub-*hashmap*) que será atualizado e a *regra* que está atualizando nó.

Na linha 2, faz-se um laço de repetição por cima de *regra* para obter todos os rótulos, chamado *l* durante as iterações. Um rótulo é um item tem que valor semântico, em outras palavras, é uma item que não é um vetor de característica. No domínio das manchas solares, um rótulo é uma classificação McIntosh. Se a regra não tiver classe no seu consequente, *atualiza* função atualiza nó. Caso contrário, linha 3, verificará se *l* já faz

Algoritmo 8: O Algoritmo *Solar Spatiotemporal Association Classifier -SSAC*.

Entrada: *Regras*: conjunto de regras de espaço-temporais temáticas cujo antecedente é composto por ao menos uma característica visual

Saída: *Map* : mapa da característica visual pela classificação

```

1 início
2   contador ← ∅ ;
3   para cada regra r ∈ Regras faça
4     características ← características visuais ∈ r.antecedente ;
5     para cada característica c ∈ características faça
6       se c ∈ contador então
7         | atualiza(contador.use(c), r) ;
8       senão
9         | atualiza(contador.novo(c), r) ;
10      fim
11     fim
12   fim
13   Map ← ∅ ;
14   para cada nó n ∈ contador faça
15     c ← mais votado(n) ;
16     c.suporte ←  $\frac{c.sup}{c.contador}$  ;
17     c.confiança ←  $\frac{c.confiança}{c.contador}$  ;
18     c.tempo ←  $\frac{c.tempo}{c.contador}$  ;
19     c.espaco ←  $\frac{c.espaco}{c.contador}$  ;
20     Map ← Map ∪ c ;
21   fim
22 fim

```

parte do nó. Se for, linha 4, *sn* recebe uma referência para o quintupla (valor sub-*hashmap*). Se não for, linha 6, um quintupla é criado e inicializado com zero em cada valor e está associado com *l*. Na linha 8, a votação é atualizada, adicionando um a *sn.contador*. Da linha 9 até 12, os valores de suporte, confiança, tempo delta e espaço delta são acumulados para *sn*. Esses valores vieram diretamente da regra de associação espaço-temporal temática chamada *regra*. Na linha 13, nó é atualizado.

A função *mais votado*, no Algoritmo 8 encontra o mais votado (maior número em *contador*) associado a um vetor de características. Isso mostra a associação mais comum de um vetor de características e uma classificação. Os valores de suporte, confiança e de variações espaciais e temporal é a média dos valores apresentados pela regras que compõem a classificação. Desta forma, ao se classificar uma nova imagem é possível calcular a probabilidade da predição de comportamento realmente ocorrer.

Como resultado é extraído um conjunto de associações entre as características visuais frequentes e futuras classificações para cada característica visual presente nos antecedentes das regras. Considere o seguinte exemplo de regra sumarizada:

Algoritmo 9: O Algoritmo da Função *atualiza*.

Entrada: *nó* : mapa de nós para atualizar. *regra* : regra a ser atualizada.

```

1 início
2   para cada rotulo  $l \in regra.consequência$  faça
3     se  $l \notin nó$  então
4       |  $sn \leftarrow$  crie sub – nó de nó para  $l$  ;
5     senão
6       |  $sn \leftarrow$  use sub – nó de nó para  $l$  ;
7     fim
8      $sn.contador \leftarrow sn.contador + 1$  ;
9      $sn.suporte \leftarrow sn.suporte + regra.suporte$  ;
10     $sn.confiança \leftarrow sn.confiança + regra.confiança$  ;
11     $sn.tempo \leftarrow sn.tempo + regra.tempo$  ;
12     $sn.espaço \leftarrow sn.espaço + regra.espaço$  ;
13     $nó \leftarrow nó \cup sn$  ;
14  fim
15 fim
```

$[2.4 - 6.2][0.6 - 2.3][0.1 - 0.5) \dots] : Fkc < 0.3, 0.8, 2, 15 >$,

significa que a mancha solar que apresenta a característica visual $[2.4 - 6.2][0.6 - 2.3][0.1 - 0.5) \dots]$ será classificada como *Fkc* com um suporte de 0.3. O suporte representa a frequência que esse padrão ocorreu na base de dados. A confiança para o exemplo é de 0.8, que representa uma probabilidade de 80% da mancha ser classificada como *Fkc* em um período de 2 dias, e; 15 representa a distância média (em partes do disco solar) que a mancha solar possui de outras manchas para isso ocorrer.

O protótipo implementando o SSAC permite que, dada uma nova imagem, as características visuais das manchas solares sejam extraídas e as mesmas sejam classificadas de acordo com as regras de associações sumarizadas.

Discussão

O objetivo principal do SSAC é facilitar o entendimento das regras de associação espaço-temporais temáticas extraídas pelo MiTSAI para o especialista de domínio. Após processadas pelo SSAC, as regras são sumarizadas por meio da utilização do mecanismo de votos. O conjunto resultante é menor e conseqüentemente mais fácil de ser analisado pelo especialista.

O SSAC também facilita a previsão de comportamento com base nas regras extraídas: sempre que uma nova imagem é processada, o modelo preditivo construído a partir da sumarização das regras é utilizado para classificar as manchas contidas na imagem. Se houver uma combinação (nova imagem e características visuais presentes nas regras sumarizadas) é possível prever o mais provável comportamento daquela mancha no

decorrer de um período.

O SSAC se difere dos demais classificadores associativos por receber como entrada regras espaço-temporais temáticas e por considerar na classificação componentes espaço-temporais. A variação temporal indica o quanto a mancha solar demora em média para ter uma evolução significativa (frequente). A variação espacial determina se esse é ou não um evento isolado (se uma mancha sofre influência de outras ao seu redor).

5.1.5 Comparação do Método proposto com os Trabalhos Correlatos

A avaliação desse trabalho está atrelada à sua comparação com trabalhos correlatos. Esses trabalhos foram apresentados nas seções (Seção 2.4 e Seção 3.3). Assim, esta seção visa compará-los com a proposta formalizada, na Seção 5.1. Na Tabela 2, são referenciados os trabalhos correlatos e comparados com este trabalho, nomeado de *Silveira Jr et. al.*.

Tabela 2 – Comparação dos trabalhos correlatos com a proposta desse trabalho.

	Regra de Associação	Dados Espaço-Temporais	Regras Temáticas
Hana, Sami e Sami (2012)	X	X	
Huo, Zhang e Meng (2013)	X	X	X
Pillai <i>et al.</i> (2012)	X	X	
Pillai, Angryk e Aydin (2013)	X	X	
McGuire, Gangopadhyay e Janeja (2012)	X	X	X
Petitjean <i>et al.</i> (2011)		X	
Petitjean <i>et al.</i> (2012)		X	
Lu <i>et al.</i> (2013)	X		
<i>Silveira Jr et. al.</i>	X	X	X
	Dados Complexos	Extração de Características	Dados Adicionais
Hana, Sami e Sami (2012)			X
Huo, Zhang e Meng (2013)			X
Pillai <i>et al.</i> (2012)	X		
Pillai, Angryk e Aydin (2013)	X		
McGuire, Gangopadhyay e Janeja (2012)	X		X
Petitjean <i>et al.</i> (2011)	X	X	
Petitjean <i>et al.</i> (2012)	X	X	
Lu <i>et al.</i> (2013)	X	X	
<i>Silveira Jr et. al.</i>	X	X	X

Fonte: o próprio autor.

O foco de *Silveira Jr et. al.* é a extração de regras de associação espaço-temporais temáticas. Apenas os trabalhos de Huo, Zhang e Meng (2013) e McGuire, Gangopadhyay e Janeja (2012), também realizam esses tipos de análises na mineração de dados. Trabalhos como Petitjean *et al.* (2011) e Petitjean *et al.* (2012) não fazem a extração de regras de associação, porém utilizam o domínio de séries de imagens de satélites. Trabalhos

como Pillai *et al.* (2012), Pillai, Angryk e Aydin (2013) e Hana, Sami e Sami (2012), apesar de extraírem regras de associação em domínios espaço-temporais, essas não são do tipo temáticas. Lu *et al.* (2013) aplica a extração de regras de associação porém não espaço-temporal, diferentemente deste trabalho.

Outra característica importante deste trabalho é a extração do conhecimento no domínio de séries temporais de imagens. A abordagem de McGuire, Gangopadhyay e Janeja (2012) não realiza uma etapa de extração de característica e, por essa razão, fica impossibilitado de utilizar atributos visuais extraídos das imagens em seu processo de mineração tal como é realizado no presente trabalho. Os trabalhos de Huo, Zhang e Meng (2013) e Hana, Sami e Sami (2012) utilizam apenas dados semânticos, mas não utilizam imagens, enquanto, Pillai *et al.* (2012), Pillai, Angryk e Aydin (2013), Petitjean *et al.* (2011), Petitjean *et al.* (2012) e Lu *et al.* (2013) utilizam imagens, mas não utilizam os dados semânticos referentes a elas.

5.2 Considerações Finais

Neste capítulo, foi apresentada a proposta do método *SolarMiner*, compostos pelos métodos SETL, MiTSAI e SSAC, que é uma nova solução para a extração de conhecimento em imagens que apresentam relações espaço-temporais e dados semânticos.

O MiTSAI é o principal componente do método, sendo um algoritmo para a extração de conhecimento oculto em imagens e dados solares. Sobre o domínio do dados: há outros trabalhos, não relacionados a mineração de dados, que utilizam a mesma fonte e também apresentam detalhadas explicações sobre cada tipo de imagens extraídas e suas informações adicionais, por exemplo: McAteer *et al.* (2005), Briand (2010), Yu *et al.* (2010), Petrie e Sudol (2010) e Kaufmann *et al.* (2011). Também, existem outros trabalhos que fazem a aplicação de tarefas de mineração de dados nesses mesmos dados como Higgins *et al.* (2011), Zhou e Zhang (2013), Bloomfield, Gallagher e McAteer (2011). Porém, ao contrário do trabalho aqui proposto, nenhum desses trabalhos considerou ao mesmo tempo os dados semânticos, dados visuais e as características espaço-temporais do domínio.

6 Experimentos

Neste capítulo, são apresentados os experimentos realizados para a validação do método *SolarMiner* desenvolvido neste projeto.

O método desenvolvido é dividido em três módulos: o SETL (processo de ETL), o MiTSAI (minerador) e o SSAC (classificador associativo). Os experimentos foram realizados para a validação de cada um desses módulos. Este capítulo está organizado da seguinte maneira: na Seção 6.1, é apresentada a base de dados utilizada nos experimentos; na Seção 6.2, é apresentado o especialista de domínio; na Seção 6.3, são apresentados os experimentos iniciais; na Seção 6.4 são apresentados os experimentos para validar o *Processo de ETL* proposto, o SETL; na Seção 6.5, são apresentados os experimentos com o minerador proposto, o *MiTSAI*; na Seção 6.6, são apresentados os experimentos com o classificador associativo proposto, o *SSAC*; na Seção 6.7, é apresentado o GQM para os experimentos realizados, e; na Seção 6.8, são apresentadas as considerações finais.

Os experimentos foram realizados em um notebook com a seguinte configuração: 8 GB de memória RAM DDR3, 500 GB de HD e processador Intel Dual Core 2,53 GHz. Sistema operacional Linux Arch sendo que os experimentos foram executados com prioridade de tempo real (*Fist In Fist Out* - FIFO). O método *SolarMiner* foi completamente implementado em Java 8 e Python 3.

6.1 A Base de Dados Usada

A base de dados utilizada nesses experimentos foi produzida pelo processo SETL e é composta por séries de imagens solares de satélites.

Como resultado do processo SETL, a base de dados usada para a mineração é composta por 10297 registros de manchas solares divididos por dia pelo período que começa em 25 de agosto de 2007 e termina em 24 de agosto de 2016, totaliza 70325 vetores de características foram processados.

Cada registro representa uma mancha solar em um determinado momento (com granularidade de um dia). Desta forma, um registro é composto pelos seguintes campos:

Id: identificador da mancha. Essa informação é oriunda da fonte de dados;

Data: data que as imagens solares foram coletadas;

Coordenadas solares: conjunto de latitude e longitude solares da mancha;

Área: área da mancha solar em unidades de área de disco solar ($6,09 \times 10^{12} km^2$);

Classe McIntosh: classificação do especialista para a mancha no momento da coleta dos dados;

Vetor de característica 1: vetor de característica extraído para a mancha solar em um determinado comprimento de onda, o tamanho desse vetor pode variar de acordo com tipo de extrator de característica que é utilizado. Por exemplo, ao se extrair por meio do extrator de características Histograma em tons de cinza, o vetor terá 255 posições; para Haralick e SURF os valores podem variar entre 125 à 512 posições.

Vetor de característica 2: vetor de característica extraído de uma imagem coletado com um aparelho em um comprimento de onda diferente do vetor de característica 1, e;

Demais vetores de características: são coletadas até 6 imagens em diferentes comprimentos de onda o que geram até 6 vetores de características por mancha solar por dia.

Esse exemplo se difere do apresentado na Figura 13 pelo fato dos vetores de características estarem normalizados: na Figura 13 havia duas manchas solares para o mesmo dia, em uma visão temporal do disco solar. Esse registro é como essa informação é lida pelo MiTSAI, previamente processada por mancha solar e tempo. Outra diferença com relação ao registro da Figura 13 é a não existência da classificação Hale, isso foi um requisito do especialista de domínio: a classificação McIntosh é mais completa que a classificação Hale. Desta forma, a classificação Hale é redundante e foi removida.

6.2 O Especialista de Domínio

Esse trabalho de pesquisa contou com a participação muito próxima de um especialista em clima espacial do Instituto Nacional de Pesquisas Espaciais (INPE). Suas principais atividades foram: orientação sobre os dados de domínio, como manipulá-los, como fazer as conversões para não adicionar ruídos, e quais as possíveis manipulações a serem feitas.

Durante o desenvolvimento do MiTSAI, o especialista apontou um dos principais requisitos que colaborou para o sucesso dos experimentos: que o antecedente de uma regra e o conseqüente devem estar relacionados sempre à mesma mancha solar, pois, para o domínio do problema, é necessário entender a evolução da mancha em si.

Durante os experimentos, o especialista ajudou na parametrização dos algoritmos, quantificou o tempo hábil de processamento e analisou as regras mineradas e os demais resultados obtidos.

6.3 Experimentos Iniciais

Os experimentos iniciais têm o objetivo de validar o processo SETL realizado nos dados oriundos de NOAA (2015). Assim, é demonstrado que os dados gerados pelo SETL estão preparados para serem entrada dos algoritmos de mineração.

Na Subseção 6.3.1, é apresentado o experimento inicial com a extração de Regras de Associação e na Subseção 6.3.2 é apresentado o experimento inicial com a extração de Padrões Sequenciais.

6.3.1 Experimento Inicial para a Extração de Regras de Associação

Para este experimento foram utilizados 1500 registros da base de dados descrita na Seção 6.1, nos quais são considerados apenas os dados semânticos. Esses registros foram coletados em datas sequenciais a qual teve início no dia 25 de agosto de 2015.

Na Figura 16, é apresentado o resultado da aplicação do Algoritmo Apriori nesse conjunto de dados. Para tanto, foi utilizada a implementação do Apriori presente no software Weka (HALL *et al.*, 2009).

Figura 16 – Resultado da pré-análise que utilizou apenas os atributos textuais de 1500 imagens. A geração desse resultado se deu pela aplicação do algoritmo Apriori implementado no Weka (HALL *et al.*, 2009).

```

13. haleClass2=β/β ncinClass0=Hsx/Hsx 80 ==> haleClass0=α/α 80 <conf:(1)> lift:(5.11) lev:(0.04) [64] conv:(64.33)
14. mcinClass0=Hsx/Hsx 211 ==> haleClass0=α/α 139 <conf:(0.66)> lift:(3.36) lev:(0.07) [97] conv:(2.32)
15. mcinClass1=Hsx/Hsx 179 ==> haleClass1=α/α 117 <conf:(0.65)> lift:(4.15) lev:(0.06) [88] conv:(2.39)
16. haleClass0=α/α haleClass2=β/β 132 ==> mcinClass0=Hsx/Hsx 80 <conf:(0.61)> lift:(4.25) lev:(0.04) [61] conv:(2.14)
17. haleClass0=α/α ncinClass0=Hsx/Hsx 139 ==> haleClass2=β/β 80 <conf:(0.58)> lift:(2.25) lev:(0.03) [44] conv:(1.72)
18. haleClass1=α/α 233 ==> ncinClass1=Hsx/Hsx 117 <conf:(0.5)> lift:(4.15) lev:(0.06) [88] conv:(1.75)
19. haleClass0=α/α 290 ==> ncinClass0=Hsx/Hsx 139 <conf:(0.48)> lift:(3.36) lev:(0.07) [97] conv:(1.64)
20. haleClass0=α/α 290 ==> haleClass2=β/β 132 <conf:(0.46)> lift:(1.78) lev:(0.04) [57] conv:(1.36)
21. haleClass0=β/β/β 209 ==> haleClass1=β/β/β 95 <conf:(0.45)> lift:(3.03) lev:(0.04) [63] conv:(1.54)
22. haleClass1=α/α 233 ==> haleClass2=β/β 105 <conf:(0.45)> lift:(1.76) lev:(0.03) [45] conv:(1.34)
23. haleClass0=β/β/β 209 ==> haleClass2=β/β/β 94 <conf:(0.45)> lift:(3.14) lev:(0.04) [64] conv:(1.54)
24. haleClass2=β/β/β 212 ==> haleClass0=β/β/β 94 <conf:(0.44)> lift:(3.14) lev:(0.04) [64] conv:(1.53)
25. haleClass0=α/α 290 ==> haleClass1=β/β 128 <conf:(0.44)> lift:(1.8) lev:(0.04) [56] conv:(1.34)
26. haleClass1=β/β/β 222 ==> haleClass0=β/β/β 95 <conf:(0.43)> lift:(3.03) lev:(0.04) [63] conv:(1.49)
27. haleClass1=β/β 363 ==> haleClass2=β/β 141 <conf:(0.39)> lift:(1.52) lev:(0.03) [48] conv:(1.21)
28. mcinClass0=Hsx/Hsx 211 ==> haleClass2=β/β 80 <conf:(0.38)> lift:(1.48) lev:(0.02) [26] conv:(1.19)
29. mcinClass0=Hsx/Hsx 211 ==> haleClass0=α/α haleClass2=β/β 80 <conf:(0.38)> lift:(4.25) lev:(0.04) [61] conv:(1.46)
30. haleClass2=β/β/β 212 ==> haleClass1=β/β/β 80 <conf:(0.38)> lift:(2.52) lev:(0.03) [48] conv:(1.36)
31. haleClass2=β/β 379 ==> haleClass1=β/β 141 <conf:(0.37)> lift:(1.52) lev:(0.03) [48] conv:(1.2)
32. haleClass1=β/β/β 222 ==> haleClass2=β/β/β 80 <conf:(0.36)> lift:(2.52) lev:(0.03) [48] conv:(1.33)
33. haleClass1=β/β 363 ==> haleClass0=α/α 128 <conf:(0.35)> lift:(1.8) lev:(0.04) [56] conv:(1.24)
34. haleClass0=β/β 332 ==> haleClass1=β/β 116 <conf:(0.35)> lift:(1.43) lev:(0.02) [34] conv:(1.15)
35. haleClass2=β/β 379 ==> haleClass0=α/α 132 <conf:(0.35)> lift:(1.78) lev:(0.04) [57] conv:(1.23)
36. haleClass0=β/β 332 ==> haleClass2=β/β 115 <conf:(0.35)> lift:(1.35) lev:(0.02) [30] conv:(1.13)
37. haleClass1=β/β 363 ==> haleClass0=β/β 116 <conf:(0.32)> lift:(1.43) lev:(0.02) [34] conv:(1.14)
38. haleClass2=β/β 379 ==> haleClass0=β/β 115 <conf:(0.3)> lift:(1.35) lev:(0.02) [30] conv:(1.11)
39. haleClass2=β/β 379 ==> haleClass1=α/α 105 <conf:(0.28)> lift:(1.76) lev:(0.03) [45] conv:(1.16)
40. haleClass0=α/α 290 ==> haleClass2=β/β ncinClass0=Hsx/Hsx 80 <conf:(0.28)> lift:(5.11) lev:(0.04) [64] conv:(1.3)
41. haleClass2=β/β 379 ==> ncinClass0=Hsx/Hsx 80 <conf:(0.21)> lift:(1.48) lev:(0.02) [26] conv:(1.08)
42. haleClass2=β/β 379 ==> haleClass0=α/α ncinClass0=Hsx/Hsx 80 <conf:(0.21)> lift:(2.25) lev:(0.03) [44] conv:(1.14)

```

Fonte: o próprio autor.

O Weka é um software de aprendizagem de máquina bastante conceituado e de fácil utilização. Por esses motivos, optou-se pela sua utilização nesse estudo. As regras apresentadas são numeradas de 13 a 42: as doze primeiras apresentavam atributos nulos, devido aos dados serem esparsos, e foram descartadas. No exemplo da Figura 16, as duas

primeiras regras são:

$$haleClass_2 = \beta / \beta \quad mcinClass_0 = Hsx / Hsx \rightarrow haleClass_0 = \alpha / \alpha \quad < sup \geq 0.05, conf : 1 >$$

Quando a Classe *Hale* é tipo β / β na segunda mancha solar da imagem e a classe *McIntosh* é Hsx / Hsx para a primeira mancha; a classe *Hale* da primeira mancha será α / α , com 100% de confiança e suporte $\geq 5\%$.

$$mcinClass_0 = Hsx / Hsx \rightarrow haleClass_0 = \alpha / \alpha \quad < sup \geq 0.05, conf : (0.66) >$$

Quando a Class *McIntosh* é Hsx / Hsx para a primeira mancha; a classe *Hale* da primeira mancha será α / α , com 66% de confiança e suporte $\geq 5\%$. As demais regras apresentadas na Figura 16 seguem o mesmo padrão de interpretação.

É importante ressaltar que neste experimento foram utilizadas somente 1500 registros da base de dados por causa da limitação de memória existente no uso do algoritmo Apriori do software Weka. Além da limitação de memória, processar mais de 1500 registros solares fez com que o tempo de resposta do sistema aumentasse muito a ponto de se tornar inviável para os experimentos iniciais: mais de 1 dia de processamento.

6.3.2 Experimento Inicial para a Extração de Padrões Sequenciais

Na Figura 17, são apresentados exemplos de sequências extraídas com o IncMSTS (SILVEIRA-JUNIOR; SANTOS; RIBEIRO, 2013; SILVEIRA-JUNIOR *et al.*, 2015). IncMSTS é um algoritmo utilizado para a extração incremental de sequências em dados espaço-temporais. O algoritmo aplica uma técnica chamada *Stretchy Time Windows* que permite configurar o espaçamento temporal máximo entre os dados, assim entre a ocorrência dos eventos em uma sequência extraída podem haver lacunas temporais (com ou sem eventos não frequentes). O IncMSTS também permite a extração de padrões semi-frequentes, que são padrões cujos valores de suporte são próximos do valor mínimo necessário para considerar os padrões como frequentes na base de dados, o quão próximo é também configurável. Essa última funcionalidade é utilizada para a extração incremental, quando os dados sofrem incrementos constantes, e não foi utilizada neste experimento.

Para esse experimento foi utilizada a base de dados completa descrita na Seção 6.1, mas apenas os dados semânticos e, diferentemente do experimento apresentado na Seção 6.3.1, o período analisado neste experimento foi o período completo de cinco anos. O suporte mínimo que esse estudo utilizou foi de 10% e o Janelamento (δ) foi de duas semanas. Assim, é possível uma lacuna temporal máxima de duas semanas entre a ocorrência de eventos para o mesmo ser considerado uma sequência.

O experimento extraiu 6885 sequências. Na Figura 17, são apresentadas três delas. A Sequência (1) mostra que $\beta / -$ ocorre após a ocorrência da classe Axx / Axx . Isso significa que a ocorrência de uma imagem que possui o atributo $B / -$ está associada a

Figura 17 – Resultado da pré-análise que utilizou apenas os dados semânticos. A geração desse resultado se deu pela aplicação do algoritmo IncMSTS implementado em (SILVEIRA-JUNIOR; SANTOS; RIBEIRO, 2013; SILVEIRA-JUNIOR *et al.*, 2015)

- $$(1) < Axx/ Axx \ \beta/ - > \text{support} : 0.1$$
- $$(2) < (Bxo/ Bxo \ \beta\gamma/ \beta\gamma) \ Hsx/ Hsx > \text{support} : 0.1$$
- $$(3) < \beta\gamma\delta/ \beta\gamma\delta \ Bxo/ Bxo \ Bxo/ Bxo > \text{support} : 0.1$$

Fonte: o próprio autor.

ocorrência de uma imagem com Axx/ Axx em até quatorze dias. A Sequência (2) mostra que a ocorrência de uma imagem com os atributos Bxo/ Bxo e $\beta\gamma/ \beta\gamma$ está associada a ocorrência de uma imagem com o atributo Hsx/ Hsx em até quatorze dias. Já a Sequência (3) mostra a ocorrência de três eventos. Nesse caso, a interpretação do janelamento é: a soma dos intervalos temporais entre os três eventos na sequência é no máximo quatorze dias.

Os experimentos iniciais, por comprovarem a aplicação de algoritmos de mineração sobre os dados produzidos pelo SETL, indicam que o mesmo é adequado para preparar os dados solares para serem usados como entrada de algoritmos de mineração.

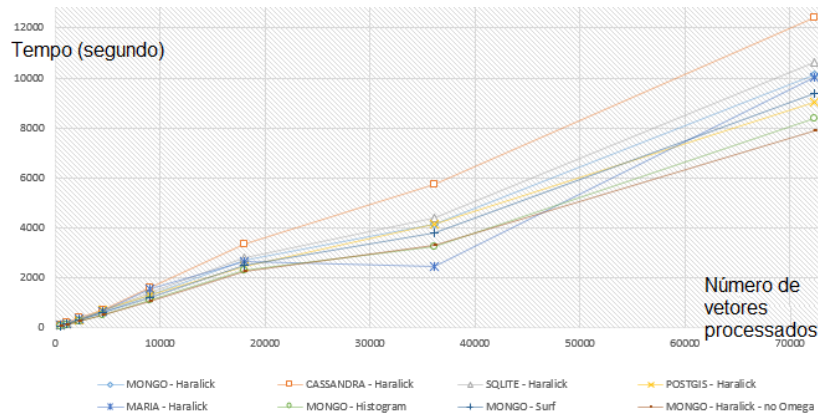
6.4 Experimentos com o Processo SETL

O processo SETL é composto por 4 etapas: extração de dados da fonte (NOAA, 2015); extração de características das imagens; discretização dos vetores de características, e; integração dos vetores de características com os dados semânticos e espaço-temporais da imagem. O objetivo desse experimento é mostrar a viabilidade em tempo computacional do processo SETL mediante a utilização de diversos SGBDs.

Na Figura 18 é apresentada uma comparação entre os SGBDs pelo SETL. Nos gráficos, é apresentado o número de regiões de manchas solares, nos eixos horizontais, pelo tempo (em segundos) necessário para o seu processamento, no eixo vertical. Para os experimentos foram utilizados os extratores de características Histograma, Haralick e SURF com e sem o processamento do Algoritmo Omega.

Por meio deste experimento é possível verificar que, para os dados solares, o PostGIS possui um melhor desempenho, 19% no valor da média dos outros sistemas de banco de dados para realização do *load* de informações que consiste na inserção dos registros solares na base de dados.

Figura 18 – Comparativo entre os SGBDs avaliados para o armazenamento de dados no SETL. O eixo horizontal é o número de imagens de manchas solares e o eixo vertical é o tempo de processamento em segundos.



Fonte: o próprio autor.

Por meio deste experimento é possível verificar que o algoritmo Omega utiliza em média 20% do tempo de processamento do processo de SETL. Além disso, o SURF é em média de 4,5% mais caro em termos de tempo do que o Haralick.

O SGBD Cassandra apresentou o pior desempenho no experimento. Isso ocorreu devido à característica de replicação de dados que foi usada no experimento. Foi usada a replicação padrão, que é a replicação de dados em três clusters, e isso pode ter causado a sobrecarga no processamento da inserção dos dados.

Apesar do PostGIS e MariaDB apresentarem bons resultados de desempenho, esses são SGBDs relacionais que possuem problemas relacionados a escalabilidade do processamento e, também, com a manipulação de dados distribuídos, conforme apresentado em Ivanova *et al.* (2013), Manu e Anandakumar (2015), Hu *et al.* (2017). Os problemas de dimensionamento dificultam a execução de algoritmos distribuídos/paralelos de mineração de dados, que são altamente recomendáveis para processar grandes quantidades de dados.

O MongoDB também apresentou bons resultados de desempenho, e em alguns casos, o mesmo desempenho do PostGIS e MariaDB.

Estes experimentos indicam que a arquitetura do processo SETL proposto é desacoplada, o qual permite o armazenamento dos dados coletados tanto em SGBDs relacionais, quanto em não-relacionais. Assim, é possível dizer que o SGBD pode ser escolhido com base nas necessidades dos algoritmos que irão usar os dados armazenados (os consumidores de dados).

Como SGBD a ser utilizado pode ser escolhido com base no algoritmo de mineração que usará os dados, neste projeto foi escolhido o MongoDB. Mesmo que em alguns casos isso signifique um custo computacional maior do processo SETL, ele traz o benefício de poder ser usado com algoritmos de mineração com implementação de rotinas paralelas,

que é o caso de como foi implementado o MiTSAI.

6.5 Experimentos com MiTSAI

Nesta seção, são apresentados três conjuntos de experimentos, para a base de dados obtida pelo SETL. No SETL, cada conjunto teve as imagens processadas por um dos extratores de características (Histograma, Haralick e SURF). Assim, os três experimentos foram realizados sobre os mesmos dados, no entanto, com as imagens representadas por diferentes vetores de características.

Os três experimentos tiveram a mesma configuração para a execução do MiTSAI: Suporte mínimo de 1%; mínimo de confiança de 75%; a variação máxima de espaço de 150 partes do disco solar e; variação máxima de tempo de 20 dias.

Na Figura 19, são apresentadas as regras extraídas da base cujas imagens foram processadas pelo extrator de características Histograma. Na Figura 21, são apresentadas as regras extraídas da base, cujas imagens foram processadas pelo extrator de características Haralick. Na Figura 22, são apresentadas as regras extraídas da base cujas imagens foram processadas pelo extrator de características SURF.

Figura 19 – Exemplo de Regras extraídas da base cujas imagens foram processadas pelo extrator Histograma

```
(R1)
<(3584760.000;3590618.000] , (3161025.000;3167250.000] , 0020> Bxo
      Cho ->
<(3584760.000;3590618.000] , (3154539.000;3161025.000] , 0040>
      sup=0.039 conf=0.958 space=116.171 time=2.869

(R2)
<(3584760.000;3590618.000] , (3161025.000;3167250.000] , 0020> Bxo
      Hsx Cso ->
<(3577235.000;3584760.000] , (3161025.000;3167250.000] , 0070>
      sup=0.054, conf=0.767, space=135.042, time=5.023

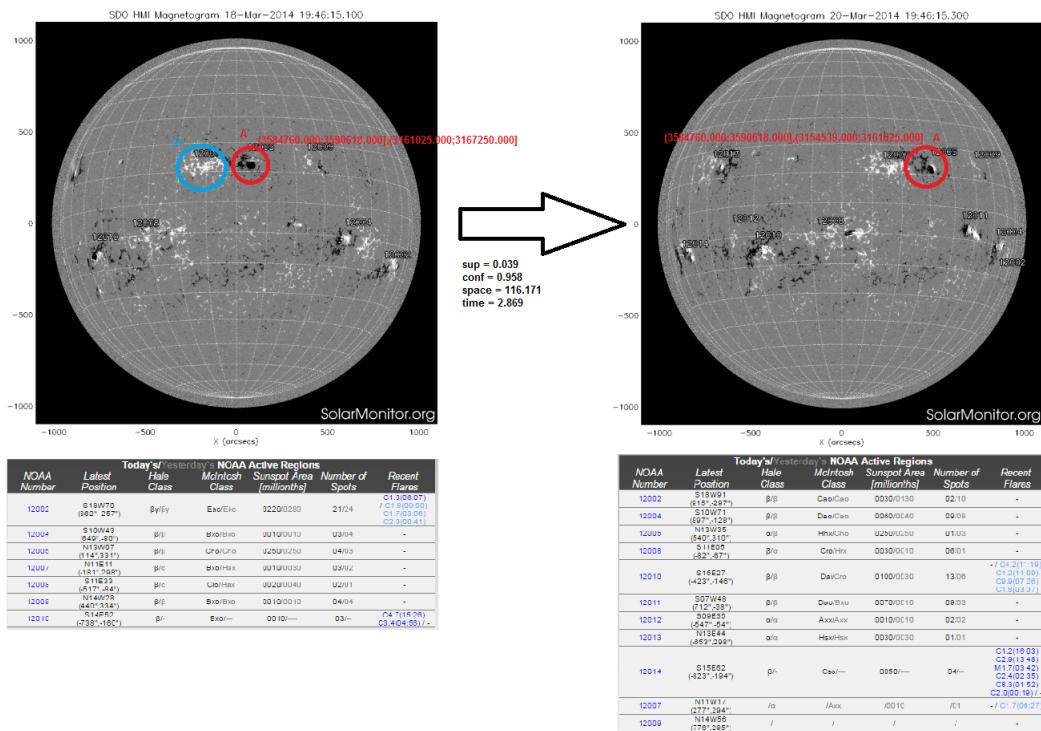
(R3)
<(3590618.000;3597197.000] , (3150223.000;3154539.000] , 0060> Cro
      ->
<(3590618.000;3597197.000] , (3150223.000;3154539.000] , 0060>
      sup=0.046, conf=0.875, space=25.703, time=2.714
```

Fonte: o próprio autor.

Na regra *R1*, é apresentada a característica visual de uma mancha solar cujo tamanho é de vinte partes do disco solar. A característica visual acontece ao mesmo tempo

de *Bxo*-McIntosh e *Cho*-McIntosh. Para esse antecedente, há dois cenários possíveis: (i) são duas manchas solares, uma tem a característica visual e a classificação *Bxo/Cho*-McIntosh e a outra mancha solar mais próxima tem a classificação *Cho/Bxo*-McIntosh; ou, (ii) são três manchas solares, uma representada pela característica visual descrita na regra, outra representada por *Bxo*-McIntosh e a outra representada por *Cho*-McIntosh. A distância entre as manchas solares é, em média, 116,171 partes do disco solar. O consequente da regra apresenta uma característica visual, sendo a evolução da característica visual apresentada no antecedente em uma média de variação de tempo de 2,869 dias. Esse padrão acontece em pelo menos *suporte* = 3,9% dos dados e a *confiança* = 95,8%: os quais indicam que essa regra ocorre em 3,9% da base de dados (frequência da ocorrência desse evento) e que em 95,8% das vezes que uma mancha solar apresentou uma característica solar como a apresentada no antecedente de *R1* a mancha evoluiu para apresentar a característica visual do consequente de *R1*. A Figura 20 apresenta um exemplo de cenário de ocorrência da regra *R1*.

Figura 20 – Exemplo de ocorrência da regra *R1*.



Fonte: o próprio autor, imagem adaptada de de NOAA (2015).

Na regra *R2*, é apresentado o mesmo vetor de característica do antecedente de *R1* associada com *Bxo*-McIntosh, *Hsx*-McIntosh e *Cso*-McIntosh. No caso, há dois cenários possíveis: (i) o vetor de características e a classificação McIntosh é da mesma mancha solar (três possibilidades) e; (ii) existem quatro manchas solares. A média da variação espacial entre as manchas solares é de 135,042 partes do disco solar.

No consequente da regra *R2*, é apresentada uma característica visual diferente se

comparado com o conseqüente da regra *R1*. A associação dos vetores de características que aparecem no antecedente de *R1* e *R2* com diferentes manchas solares, isto é, classificações diferentes, resulta em uma evolução diferente da mancha solar.

Na regra *R3*, são apresentadas duas manchas solares, uma representada pelo vetor de características e a segunda pelo *Cro*-McIntosh, a distância média entre elas é 25,703 partes do disco solar. No conseqüente é apresentada a evolução da mancha solar associada ao vetor de características. Esse evolui ainda assim apresenta o mesmo vetor de características no conseqüente.

Figura 21 – Regras extraídas da base de dados processada com o extrator de características Haralick.

```
(R4)
<(-0.350;0.341] , (9.764;23.692] , (32118.140;32751.662] ... 0120> Cso
      -> Hsx
sup=0.042, conf=0.758, space=1.249, time=9.227

(R5)
<(-0.350;0.341] , (9.764;23.692] , (45521.675;45809.851] ... 0110>
      ->
<(-0.350;0.341] , (9.764;23.692] , (45521.675;45809.851] ... 0110> Axx
sup=0.024, conf=0.875, space=40.564, time=2

(R6)
<(-0.350;0.341] , (9.764;23.692] , (29061.805;30339.404] ... 0300> Cao
      -> Dso
sup=0.177, conf=0.772, space=92.801, time=2.588
```

Fonte: o próprio autor.

Na Figura 21, são apresentadas as regras extraídas da base cujas imagens foram processadas pelo extrator de características Haralick. No antecedente da regra *R4*, são apresentadas duas manchas solares, a primeira é representada pelo vetor de características visuais e a segunda é classificada como *Cso*-McIntosh, a distância média entre elas é de 1.249 partes do disco solar. A mancha solar representada pelo vetor de características evolui para um *Hsx*-McIntosh em uma variação de tempo média de 9.227 dias. A regra é encontrada em 4,9% da base com uma confiança de 75,9%: o que indica que esse é um padrão que ocorreu em 4,9% dos registros da base de dados e que em 75,9% das vezes que uma mancha solar apresenta a característica visual do antecedente de *R4* essa mancha solar foi classificada como *Hsx* em um período médio de 9,227 dias.

Na regra *R5*, é apresentada, no seu antecedente, uma mancha solar representada por um vetor de características. Em seu conseqüente, uma característica visual semelhante

é apresentada pela mesma mancha solar associada a outra mancha solar classificada como *Axx*-McIntosh. A distância média entre essas manchas solares é de 40.564 partes do disco solar. A distância média entre as ocorrências do antecedente e consequente é de dois dias. O suporte é de 2,4% e a confiança é de 87,5%.

Na regra *R6*, são apresentadas duas manchas solares cuja distância média é de 92.801 partes do disco solar. No antecedente da regra, é apresentada uma característica visual associada a uma mancha solar classificada como *Cao*-McIntosh. A mancha solar representada pela característica visual evolui para a classificação *Dso*-McIntosh e frequentemente está associada a uma mancha solar *Cao*-McIntosh. A duração média desse processo é de 2.588 dias.

Figura 22 – Regras extraídas da base processada com o extrator de características SURF.

```
(R7)
Hsx <(665.847;666.469] , (14.534;15.198] , (-0.428;8.490] ... 0010>
      -> Dso
sup=0.164, conf=0.8, space=33.22, time=6.5

(R8)
<(406.643;407.087] , (14.534;15.198] , (-0.428;8.490] ... 0180> Bxo
      -> Dro
sup=0.107, conf=0.79, space=22.2, time=13$

(R9)
<(667.785;668.583] , (14.534;15.198] , (-0.428;8.490] ... 0040> Axx
      ->
Hsx <(406.643;407.087] , (14.534;15.198] , (-0.428;8.490] ... 0180>
      <(665.847;666.469] , (14.534;15.198] , (-0.428;8.490] ...
      0010> Axx
sup=0.028, conf=1, space=30.248, time=4
```

Fonte: o próprio autor.

Na Figura 22, são apresentadas as regras extraídas da base cujas imagens foram processadas pelo extrator de características SURF. Na regra *R7*, é apresentado uma mancha solar classificada como *Hsx*-McIntosh que evolui para *Dso*-McIntosh quando está associada a outra mancha solar com a característica visual apresentada. A distância entre as manchas solares é, em média, 33,22 partes do disco solar. O tempo médio para isso acontecer é de 6,5 dias. O padrão é recorrente em 16,4% dos dados e sua confiança é de 80%.

Na regra *R8*, também são apresentadas duas manchas solares em seu antecedente, a distância entre as 22,2 partes do disco solar. A mancha solar, que está relacionada à característica visual apresentada no antecedente da regra, evolui para um *Dro*-McIntosh

em uma média de tempo de treze dias. Isso ocorre quando a mancha está associada a uma outra mancha solar classificada como *Bxo*-McIntosh. A regra *R8* possui uma frequência de 10,7% e uma confiança de 79%.

Na regra *R9*, são apresentadas duas manchas solares em seu antecedente e a distância entre elas é 30,248 partes do disco solar. A mancha solar, que apresenta uma ou as duas características visuais que ocorrem no antecedente, evolui para a classificação *Hsx*-McIntosh, em um tempo médio de quatro dias. Essa mancha solar pode apresentar uma das duas características visuais presentes na regra, ou essas características podem ser de outras manchas solares próximas a ela em uma média de 30,248 partes do disco solar. Esta regra mostra que uma mancha solar pode estar diretamente conectada ao comportamento de outras manchas.

O comentário do especialista de domínio a respeito das regras extraídas foi positivo, pois as regras expõem comportamentos interessantes de determinadas manchas solares. Junto com os resultados do SSAC (Seção 6.6) as regras são utilizadas para saber o caminho que a mancha percorreu para chegar à classificação prevista. Essa funcionalidade foi apontada pelo especialista de domínio como necessária e fundamental para o sucesso do SolarMiner.

6.6 Experimentos com o SSAC

Nesta seção, são apresentados os experimentos realizados para a validação do Solar Spatiotemporal Associative Classifier - SSAC. O objetivo desse experimento é mostrar a viabilidade da geração de um modelo preditivo tendo como entrada um conjunto de regras de associação espaço-temporal temáticas e a precisão das previsões geradas por esse modelo.

SSAC recebe o conjunto de regras espaço-temporal temáticas oriundos do MiTSAI e tem como saída um conjunto de regras sumarizadas que são usadas para prever o comportamento de novas manchas solares.

Nas Figuras 23, 24, e 25, são apresentados os resultados do processamento do SSAC para as regras ilustradas nas Figuras 19, 21, e 22, respectivamente.

Na classificação *C1*, é apresentada uma associação da característica visual (3627361.000; 3632207.000], (3113326.000; 3119009.000] com uma mancha de tamanho 10 partes do disco solar. Essa característica visual ocorre frequentemente associada a *Cao*-McIntosh. O suporte médio das regras que fazem essa associação é de 19,4% e a confiança média dessas regras é de 78,9%. Essas regras apresentam uma variação no tempo de 5,067 dias em média, entre o antecedente e consequente, além disso, há uma variação na distância entre as manchas que estão associadas de 62,813 em média.

Figura 23 – Classificação das regras extraídas da base Histograma (exemplificadas nas Figura 19).

C1:
 (3627361.000;3632207.000], (3113326.000;3119009.000], 0010 : Cao
 sup = 0.194 conf = 0.789 time = 5.067 space = 62.813

C2:
 (3853748.000;3856799.000], (2894258.000;2897071.000], 0010 : Axx
 sup = 0.019 conf = 0.830 time = 3.781 space = 72.814

C3:
 (3590618.000;3597197.000], (3150223.000;3154539.000], 0060 : Cro
 sup = 0.023 conf = 0.778 time = 6.640 space = 3.598

Fonte: o próprio autor.

Na classificação *C2*, é apresentado uma associação da característica visual (3853748.000;3856799.000], (2894258.000;2897071.000] de uma mancha cujo tamanho é 10 partes do disco solar com *Axx*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 1,9% e a confiança é de 83%. A variação média no tempo é de 3,781 dias e a variação no espaço é de 72,814 partes do disco solar em média.

Na classificação *C3*, é apresentado uma associação da característica visual (3590618.000;3597197.000], (3150223.000;3154539.000] de uma mancha cujo tamanho é 60 partes do disco solar com *Cro*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 2,3% e a confiança é de 77,8%. A variação média no tempo é de 6,64 dias e a variação no espaço é de 3,598 partes do disco solar em média.

Na classificação *C4*, é apresentado uma associação da característica visual $(-0.350; 0.341]$, $(9.764; 23.692]$, $(30339.404; 31432.453]$, $(4949.939; 4952.800]$... de uma mancha cujo tamanho equivale a trinta partes do disco solar com *Dso*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 18,2% e a confiança é de 76,7%. A variação média no tempo é de 2,609 dias e a variação no espaço é de 44,885 partes do disco solar em média.

Na classificação *C5*, é apresentado uma associação da característica visual $(-0.350; 0.341]$, $(9.764; 23.692]$, $(45521.675; 45809.851]$, $(7083.643; 7088.084]$... de uma mancha cujo tamanho é 110 partes do disco solar com *Axx*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 2,2% e a confiança é de 93,8%. A variação média no tempo é de 2,583 dias e a variação no espaço é de 69,508 partes do disco solar em média.

Na classificação *C6*, é apresentado uma associação da característica visual

Figura 24 – Classificação das regras extraídas da base Haralick (exemplificadas nas Figura 21.)

C4:
 (-0.350;0.341] , (9.764;23.692] , (30339.404;31432.453] , (4949.939;4952.800] ,
 (0.341;4.240] , (9.764;23.692] , (508.144;519.027] , ... 0030 : Dso
 sup = 0.182 conf = 0.767 time = 2.609 space = 44.885

C5:
 (-0.350;0.341] , (9.764;23.692] , (45521.675;45809.851] , (7083.643;7088.084] ,
 (0.341;4.240] , (24.262;28.389] , (461.257;461.907] , ... 0110 : Axx
 sup = 0.022 conf = 0.938 time = 2.583 space = 69.508

C6:
 (-0.350;0.341] , (9.764;23.692] , (29061.805;30339.404] , (4964.073;4967.090] ,
 (0.341;4.240] , (9.764;23.692] , (519.027;657.910] , ... 0300 : Dso
 sup = 0.178 conf = 0.810 time = 2.477 space = 73.351

Fonte: o próprio autor.

(-0.350;0.341] , (9.764;23.692] , (29061.805;30339.404] , (4964.073;4967.090] ... de uma mancha cujo tamanho é 300 partes do disco solar com *Dso*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 17,8% e a confiança é de 81%. A variação média no tempo é de 2,477 dias e a variação no espaço é de 73,351 partes do disco solar em média.

Figura 25 – Classificação para o conjunto de regras extraídas da base SURF (exemplificadas nas Figura 22).

C7:
 (784.896;785.216] , (14.534;15.198] , ... 0020 : Axx
 sup = 0.227 conf = 1.000 time = 1.000 space = 0.000

C8:
 (776.995;777.279] , (14.534;15.198] , ... 0010 : Dro
 sup = 0.063 conf = 1.000 time = 2.000 space = 23.005

C9:
 (664.769;665.061] , (14.534;15.198] , ... 0030 : Axx
 sup = 0.053 conf = 1.000 time = 5.536 space = 42.423

Fonte: o próprio autor.

Na classificação *C7*, é apresentado a associação entre a característica visual de uma mancha solar (784.896;785.216] , (14.534;15.198] ... , cujo tamanho é de vinte partes do

Figura 26 – Visualização para manchas solares que não fizeram parte do conjunto de treinamento.

```
Sunspot 11147 will be classified as Bxo<sup=0.057,
  conf=0.769,time=3.791(120.144),space=103.589(39.003)>
```

```
Sunspot 11148 has no classification
```

Fonte: o próprio autor.

disco solar com *Axx*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 22,7% e a confiança é de 100%. A variação média no tempo é de um dia e não há variação no espaço, isso indica uma evolução da própria mancha sem a necessidade de estar associada a outra mancha para esse comportamento.

Na classificação *C8*, é apresentado uma associação da característica visual (776.995; 777.279], (14.534; 15.198]... de uma mancha cujo tamanho é dez partes do disco solar com *Dro*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 6,3% e a confiança é de 100%. A variação média no tempo é de dois dias e a variação no espaço é de 23,005 partes do disco solar em média.

Na classificação *C9*, é apresentado uma associação da característica visual (664.769; 665.061], (14.534; 15.198]... de uma mancha cujo tamanho é trinta partes do disco solar com *Axx*-McIntosh. O suporte médio das regras que contribuíram para essa associação é 5,3% e a confiança é de 100%. A variação média no tempo é de 5,536 dias e a variação no espaço é de 42,423 partes do disco solar em média.

Na Figura 26, são apresentados exemplos de duas manchas solares que não fizeram parte do conjunto de treinamento. Para esse exemplo a mancha solar é processada para a extração dos vetores de características. Os vetores de características são comparados com o modelo de aprendizado de classificação gerado pelo SSAC. A mancha solar 11148 não apresenta nenhuma classificação. A mancha solar 11147 será classificada como *Bxo* em 3,791 dias, com um desvio de até 120 dias. Essa classificação é mais provável se a mancha estiver próxima a outra mancha solar em até 103,589 partes do disco solar, com uma variação de 39,003 partes do disco solar. A frequência com que esse cenário ocorreu é de 5,7% nas regras do modelo de aprendizado gerado e a confiança nessa classificação é de 76,9%.

Para validar o resultado da classificação pelo SSAC foi usada uma base de teste de 566 manchas solares, aproximadamente três manchas por dia. Essas manchas não fazem parte da base de dados usada para gerar as regras.

Foi computado o valor de precisão obtido na classificação das imagens, no qual foram utilizados as três bases anteriores Haralick, SURF e Histograma. Os experimentos

mostraram que o valor de precisão variou de acordo com o tipo de extrator de característica utilizado:

- *Histogram*: precisão= 82,7%;
- *Haralick* : precisão = 84.1%,
- *SURF*: precisão= 87.3%.

Esses experimentos foram apresentados ao especialista de domínio que achou o resultado bastante interessante. O especialista achou bastante promissor poder verificar as regras de associação espaço-temporais temáticas que foram sumariadas para chegar na classificação.

Os trabalhos de Petitjean *et al.* (2011), Petitjean *et al.* (2012) conseguiram uma precisão maior para a classificação da imagens solares, maior que 90%. Porém, esses trabalhos não fazem a previsão do comportamento da mancha solar, pois informam apenas a sua classificação atual.

Desempenho

O processo SETL é responsável por 90% do tempo consumido na execução completa do *SolarMiner*, 9% é utilizado pelo algoritmo de mineração MiTSAI, e; 1% pelo algoritmo de SSAC. Como requisito do especialista de domínio, *SolarMiner* deve ser capaz de processar um ciclo solar cuja duração é de aproximadamente 11 anos. Foram utilizados dados dos últimos 10 anos, sendo, desta forma, possível utilizar 1 ano para a validação do modelo preditivo dentro do mesmo ciclo solar.

No pior cenário a geração de uma modelo preditivo leva 220 minutos de processamento para 10 anos de dados. A partir do modelo gerado é possível a classificação de uma séries de novas imagens e o processo de classificação leva apenas alguns segundos (entre 10 e 35 segundos).

Como requisito inicial do especialista de domínio, era necessário classificar as manchas solares em um tempo menor que a chegada de novos dados; para esse trabalho, a granularidade dos dados é de um dia. Desta forma, o *SolarMiner* atende ao requisito de desempenho do especialista de domínio.

6.7 Avaliação com o Método Goal, Question, and Metrics

Para melhorar a organização deste trabalho e apresentar de maneira clara sua avaliação, foi utilizado o método *Goal, Question, and Metrics* (GQM). Como previsto pelo método, o objetivo principal é a implementação e utilização do *SolarMiner* em imagens e

dados solares. Para verificar o cumprimento do objetivo, foi medido o cumprimento de cada objetivo específico.

Para cada objetivo específico, previamente apresentado na Seção 1.1, as seguintes perguntas foram utilizadas para verificar seu cumprimento com suas medidas:

- Projetar, desenvolver e validar um processo de ETL (Extração, Transformação e Carregamento) para a coleta dos dados considerados relevantes para a mineração de dados solares neste trabalho;
 - O processo de ETL desenvolvido (SETL) pode ser realizado em um tempo de processamento aceitável?
 - * Métrica: tempo de processamento. Essa métrica foi respondida por meio do experimento apresentado na Seção 6.4; o tempo utilizado para o processamento é aceitável para o domínio de manchas solares e para a granularidade de um dia para os dados.
 - Os dados extraídos podem ser lidos pelo algoritmo minerador?
 - * Métrica: a possibilidade de extração de regras de associação. Essa métrica foi respondida pelos experimentos apresentados na Seção 6.3 e na Seção 6.5. Os dados resultantes do processo SETL podem ser utilizados por algoritmos de mineração de dados.
- Oferecer um método para extrair regras de associação espaço-temporais temáticas de séries de imagens de satélite;
 - As regras extraídas representam a evolução temporal das manchas solares?
 - As regras extraídas representam as relações entre manchas que estão acontecendo ao mesmo momento?
 - * Métrica: ambas as questões podem ser respondidas pelas regras extraídas. Por meio das regras é possível verificar a evolução temporal das manchas e a relação espacial que elas possuem. Essa métrica foi respondida pelos experimentos apresentados na Seção 6.5 e na Seção 6.6.
- Projetar, desenvolver e validar um método de sumarização das regras espaço-temporais temáticas de séries de imagens de satélites.
 - As regras extraídas permitem a predição do comportamento das manchas solares de uma nova imagem?
 - * Métrica: precisão da predição do comportamento das manchas solares. Essa métrica é respondida pela Seção 6.6 no qual são apresentados os resultados dos experimentos com o SSAC.

Desta maneira, é possível medir o quão satisfatório o método desenvolvido *SolarMiner* é para a resolução do problema que se propõe.

6.8 Considerações Finais

Neste capítulo foram apresentados os experimentos para a validação do método proposto *SolarMiner*. Esses experimentos foram planejados para a validação dos três módulos do *SolarMiner*: o processo SETL, o minerador MiTSAI e o classificador associativo SSAC. Os experimentos foram realizados na base de imagens e dados solares por um período de dez anos. Nos experimentos para validar o SETL, o pré-processamento foi realizado por meio da utilização de três extratores de características: Histograma, Haralick e SURF. O extrator que possui maior precisão para a predição do comportamento de manchas solares foi o SURF com uma precisão de 87,3%.

O SGBD que apresentou melhor performance foi o SGBD Postgres, desta forma, a melhor escolha do ponto de vista computacional é o uso conjunto do Postgres com o SURF para a etapa SETL. Os experimentos subsequentes ao pré-processamento, com o minerador e com o classificador associativo, também foram realizados sobre essas três variações na representação dos dados.

Do ponto de vista do especialista, as regras extraídas permitem analisar tanto o comportamento esperado, como o caminho que a mancha passou para chegar a um determinado estado. Como conclusão dos experimentos, é possível dizer que os padrões extraídos são novos e interessantes para o domínio de acordo com o especialista. O tempo de execução dos algoritmos também é satisfatório de acordo com o especialista.

Parte III

Finalização

7 Conclusão

Séries Temporais de Imagens de Satélites (SITS - do inglês, *Satellite Image Time Series*) possuem uma grande quantidade de informação cuja análise acarreta um conhecimento importante para o entendimento das atividades solares. Apesar das diversas aplicabilidades deste conhecimento, a revisão sistemática da literatura indica que há poucos trabalhos cujo alvo é esse domínio. Por exemplo, a mineração de dados aplicada aos SITS é uma tarefa pouco explorada como indicado por Vatsavai *et al.* (2012). Assim, há também uma carência no desenvolvimento de técnicas de mineração para a análise dos dados solares.

A extração de conhecimento em SITS exige atenção multidisciplinar pois envolve processamento de imagens, pré-processamento de dados, mineração de dados e imagens. Além disso, a mineração de dados deve considerar as características espaço-temporais que o domínio apresenta.

Desta forma, objetivo geral deste trabalho foi auxiliar na análise e no entendimento dos dados do clima espacial. O objetivo é sustentado pela hipótese de que a extração de regras de associação espaço-temporais temáticas traz informações relevantes para o entendimento deste domínio. Isso porque as regras de associação espaço-temporais temáticas são padrões que retratam tanto a relação entre eventos que estão ocorrendo ao mesmo tempo quanto a evolução desses eventos e suas relações espaciais.

Por considerar as características relevantes do domínio de séries temporais de imagens de satélites, os padrões extraídos nesse trabalho tendem a melhorar o entendimento do clima espacial. Esse padrões consideram relação entre eventos solares próximos como a evolução desse evento no decorrer do tempo. Por meio da generalização, foi possível a extração de padrões que permitem a classificação preditiva de eventos solares. Por exemplo, dada uma terminada mancha solar com a característica visual A , essa mancha será classificada como α em um período médio de X dias. Esse tipo de padrão inédito na literatura da domínio da dados solares permite um melhor entendimento do mesmo.

Este capítulo encontra-se organizado da seguinte forma: na Seção 7.1, é apresentada uma discussão sobre os resultados obtidos nesse trabalho; na Seção 7.2, são apresentadas as contribuições deste trabalho; na Seção 7.3, são apresentados os trabalhos futuros, e; na Seção 7.4, este capítulo é finalizado com as considerações finais.

7.1 Avaliação dos Resultados

Por meio da análise dos resultados na qual foi utilizado o método GQM, apresentados na Seção 6.7, é possível conectar cada um dos experimentos realizados com um dos objetivos específicos deste trabalho de pesquisa. Com o cumprimento dos objetivos específicos é possível julgar que o objetivo geral "auxiliar a análise do clima espacial a partir da mineração de regras de associação espaço-temporais nas quais envolvem atributos temáticos" também foi completado.

Outro ponto a respeito do *SolarMiner* é que pode ser adaptado a outros domínios não necessariamente relacionados aos dados do clima espacial: o *SolarMiner* poder ser utilizado em qualquer domínio composto por séries de imagens e seus dados semânticos.

Os comentários do especialista de domínio são bastante positivos:

- A capacidade de prever classificações com base no estado atual da uma mancha solar e outras próximas aumenta bastante a compreensão dos fenômenos solares;
- Por meio da utilização das regras mineradas é possível verificar o caminho evolutivo mais comum entre as manchas solares.

Desta forma, é possível dizer que este trabalho de pesquisa teve êxito no cumprimento de seu objetivo.

7.2 Principais Contribuições

Nesta seção, são apresentadas as principais contribuições deste trabalho de pesquisa. Como o objetivo de melhorar a apresentação dos resultados, esta seção encontra-se subdividida em Seção 7.2.1 e Seção 7.2.2. Na Seção 7.2.1, são apresentados os principais pontos inovadores deste trabalho de doutorado em relação a literatura. Na Seção 7.2.2, é apresentada a lista com as publicações que visam divulgar esse trabalho à comunidade científica.

7.2.1 Principais Pontos Inovadores

A principal contribuição deste trabalho de pesquisa é o algoritmo MiTSAI. O MiTSAI se difere dos demais algoritmos de mineração de dados espaço-temporais por considerar a aplicação de restrições em fases diferentes de seu processamento: a restrição espacial durante a etapa de extração dos *itemsets* frequentes e a restrição temporal durante a geração das regras de associação. Essa característica faz com que o algoritmo extraia regras interessantes, pois consideram tanto a influência entre as manchas próximas entre

si, quando a evolução das mesmas no decorrer de um período: esse conhecimento foi um dos requisitos principais apontados pelo especialista de domínio.

Os resultados obtidos foram considerados relevantes e interessantes pelo especialista de domínio, pois, por meio deles, é possível prever o comportamento de determinadas manchas além de acompanhar sua evolução em um tempo de processamento aceitável.

O MiTSAI também pode ser facilmente aplicado a diferentes domínios desde que esses sejam compostos por séries temporais de imagens e dados espaço-temporais. Desta forma, o algoritmo MiTSAI é considerado a principal contribuição para esse trabalho de pesquisa.

Uma outra importante contribuição foi a ocorrência de uma melhora no entendimento do domínio do clima espacial por meio da extração das regras espaço-temporais. Esse foi o objetivo principal deste trabalho, comprovando, desta forma, a hipótese proposta para este trabalho.

As contribuições secundárias foram (i) a arquitetura de pré-processamento dos SITS e dados solares, com o desenvolvimento do processo SETL e (ii) o desenvolvimento do classificador associativo SSAC para auxiliar no entendimento dos padrões extraídos, resumindo as regras, e classificando novas imagens com base nas regras sumarizadas.

A arquitetura SETL é flexível podendo ser empregada para outros domínios compostos ou não por séries de imagens. O classificador associativo SSAC também é flexível e pode ser utilizado como pós-processamento de regras de associação espaço-temporais temáticas. Durante a análise da literatura, não foi encontrado outro classificador associativo aplicado a regras de associação espaço-temporais temáticas. Desta forma, o SSAC também é considerado um ponto inovador e de relevância.

A comparação de diferentes SGBDs para o armazenamento dos dados solares pré-processados também foi feita para a arquitetura de pré-processamento, Nessa contribuição, é comparado o desempenho de diferentes SGBDs na tarefa de inserção dos solares pré-processados.

7.2.2 Lista de Publicações

Os seguintes artigos foram propostos e publicados com o intuito de divulgação deste trabalho de pesquisa:

- SILVEIRA-JR, C.R.; SANTOS, M.T.P.; RIBEIRO, M.X. A Flexible Architecture to Integrate the Solar Satellite Image Time Series Data - The SETL Architecture. *Int. J. Data Mining, Modelling and Management*, Vol. 10, No. 1759-1163, p. 1-16 (2018).
- SILVEIRA-JR, C.R.; SANTOS, M.T.P.; RIBEIRO, M.X. Spatiotemporal Associative

Classifier for Satellite Image Time Series. In The 31th International Flairs Conference (2018).

- SILVEIRA-JR, C.R.; CECATTO, J.R.; SANTOS, M.T.P.; RIBEIRO, M.X. Thematic Spatiotemporal Association Rules to Track the Evolving of Visual Features and their Meaning in Satellite Image Time Series. In Information Technology-New Generations, p. 317-323. Springer, Cham (2018).

Artigos em etapa de revisão:

- SILVEIRA-JR, C.R.; CECATTO, J.R.; SANTOS, M.T.P.; RIBEIRO, M.X. MiTSAI: A Deeper Analysis of Satellite Image Time Series Using Thematic Spatiotemporal Association Rule. Em etapa de revisão, artigo de revista.
- CECATTO, J.R.; RIBEIRO, M.X.; ANTUNES DA SILVA, A.E.; GRADVOHL, A.L.S.; FERNANDES, M.M.; COELHO, G. P.; SANTOS, M.T.P; SILVEIRA-JR, C.R.; DISCOLA-JR., S.L.Solar Flare Forecasting - Origin, Evolution, Perspectives and Trends. In *Annales Geophysicae*, Special Issue: Space Weather Connections to Near-Earth Space and the Atmosphere (2018). Artigo de aceite em revista, porém em etapa de revisões finais.

As publicações anteriores visam disponibilizar o conhecimento desenvolvido para a comunidade enfatizando a mineração de regras de associação espaço-temporais no domínio do clima espacial.

7.3 Trabalhos Futuros

A primeira oportunidade de pesquisa futura para o MiTSAI é a utilização de tecnologia de distribuição de processamento, como o *Apache Spark Streaming*, para que o MiTSAI passe a trabalhar com dados de fluxo contínuo. Desta forma, é possível que as regras de associação sejam atualizadas online a medida que novos dados sejam produzidos.

Como segunda oportunidade de pesquisa futura, o MiTSAI pode ser adaptado para receber os vetores de características sem o processo de discretização, ou seja, o MiTSAI pode ser adaptado para trabalhar com dados contínuos, produzindo assim um algoritmo de mineração de regras de associação espaço-temporais temáticas contínuas. Essa melhoria ajudaria a reduzir o tempo de pré-processamento e também reduzia a possibilidade da inserção de ruídos pelo processo de discretização. Além disso, reduzira os parâmetros de configuração do pré-processamento tornando-o mais simples.

Outra oportunidade de pesquisa está no pós-processamento dos dados, após a etapa de classificação, quando as regras sumarizadas são geradas. É possível fazer enriquecimento

semântico pela aplicação de conceitos com ontologias. No pós-processamento, é também possível um estudo para verificar a melhor forma de visualizar os resultados obtidos e compará-los com o conjunto de dados processado.

Por fim, a aplicação do *SolarMiner* em domínios diferentes também é uma oportunidade de pesquisa. Para tanto é necessário adaptar o extrator de características e configurar o algoritmo MiTSAI de acordo com as recomendações do novo domínio a ser trabalhado.

7.4 Considerações Finais

A extração de conhecimento em SITS solares é uma tarefa de mineração de dados tão relevante quanto desafiadora. O conhecimento obtido é utilizado pelos especialistas de domínio para um melhor entendimento do clima espacial e de seus padrões. Essa é uma tarefa desafiadora pois envolve conhecimentos multi-disciplinares: pre-processamento de imagens, mineração de dados espaço-temporal, classificação entre outras. Para viabilizar a extração de conhecimento desse domínio, neste trabalho de pesquisa foi proposto o método *SolarMiner*.

O *SolarMiner* envolve o processamento dos dados através da arquitetura de *Extract, Transform, and Load* (ETL) proposta, chamada de SETL; o algoritmo extrator de regras de associação espaço-temporais temáticas, chamado de MiTSAI, e; o classificador associativo por voto para as regras extraídas, chamado de SSAC. Com base no experimentos realizados é possível dizer que o *SolarMiner* apresentou um bom desempenho para a extração de conhecimento dos SITS solares, no sentido de obter padrões novos e interessante segundo o especialista. Por fim, neste capítulo, foram apresentadas as contribuições para com o meio científico-acadêmico e os trabalhos futuros.

Referências

ABAR, O.; CHARNIGO, R. J.; RAYAPATI, A.; KAVULURU, R. On interestingness measures for mining statistically significant and novel clinical associations from emrs. In: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York, NY, USA: ACM, 2016. (BCB '16), p. 587–594. ISBN 978-1-4503-4225-4. Citado na página 34.

ABIRAMI-KONGU, T.; THANGARAJ, P.; PRIAKANTH-KONGU, P. Wireless sensor networks fault identification using data association. *Journal of Computer Science*, v. 8, n. 9, p. 1501–1505, 2012. Citado na página 45.

ABUL, O.; GAKASE, H. Knowledge hiding from tree and graph databases. *Data and Knowledge Engineering*, v. 72, n. 0, p. 148 – 171, 2012. ISSN 0169-023X. Citado na página 36.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. [S.l.: s.n.], 1994. v. 1215, p. 487–499. Citado 2 vezes nas páginas 36 e 44.

ALIZADEHSANI, R.; ZANGOUEI, M. H.; HOSSEINI, M. J.; HABIBI, J.; KHOSRAVI, A.; ROSHANZAMIR, M.; KHOZEIMEH, F.; SARRAFZADEGAN, N.; NAHAVANDI, S. Coronary artery disease detection using computational intelligence methods. *Knowledge-Based Systems*, v. 109, p. 187 – 197, 2016. ISSN 0950-7051. Citado na página 63.

ALJAWARNEH, S. A.; VANGIPURAM, R.; PULIGADDA, V. K.; VINJAMURI, J. G-spamine: An approach to discover temporal association patterns and trends in internet of things. *Future Generation Computer Systems*, v. 74, p. 430 – 443, 2017. ISSN 0167-739X. Citado na página 42.

ALMODAIFER, G.; HAFEZ, A.; MATHKOUR, H. Discovering medical association rules from medical datasets. In: *IT in Medicine and Education (ITME), 2011 International Symposium on*. [S.l.: s.n.], 2011. v. 2, p. 43–47. Citado na página 52.

ANDRIENKO, G.; ANDRIENKO, N.; KEIM, D.; MACEACHREN, A. M.; WROBEL, S. Challenging problems of geospatial visual analytics. *Journal of Visual Languages & Computing*, v. 22, n. 4, p. 251 – 256, 2011. ISSN 1045-926X. Part Special Issue on Challenging Problems in Geovisual Analytics. Citado na página 51.

ANTONIE, L.; ZAIANE, O. R.; HOLTE, R. C. Redundancy reduction: Does it help associative classifiers? In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2016. (SAC '16), p. 867–874. ISBN 978-1-4503-3739-7. Citado na página 34.

ATTIG, A.; PERNER, P. A comparison between haralick's texture descriptor and the texture descriptor based on random sets for biological images. In: _____. *Machine Learning and Data Mining in Pattern Recognition: 7th International Conference, MLDM 2011, New York, NY, USA, August 30 – September 3, 2011. Proceedings*. Berlin,

- Heidelberg: Springer Berlin Heidelberg, 2011. p. 524–538. ISBN 978-3-642-23199-5. Citado na página 54.
- AYDIN, B.; KEMPTON, D.; AKKINENI, V.; ANGRYK, R.; PILLAI, K. Mining spatiotemporal co-occurrence patterns in solar datasets. *Astronomy and Computing*, v. 13, p. 136 – 144, 2015. ISSN 2213-1337. Citado na página 47.
- BANDYOPADPHYAY, S.; MAULIK, U. *Data mining and knowledge discovery methods with case examples*. [S.l.: s.n.], 2013. 243 – 270 p. Citado na página 53.
- BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: SPRINGER. *European conference on computer vision*. [S.l.], 2006. p. 404–417. Citado 2 vezes nas páginas 55 e 76.
- BLANCHART, P.; FERECATU, M.; DATCU, M. Mining large satellite image repositories using semi-supervised methods. *2011 Ieee Int. Geoscience Remote Sensing Symposium (igarss)*, p. 1585–1588, 2011. Citado na página 53.
- BLOOMFIELD, S.; GALLAGHER, P. T.; MCATEER, R. T. J. Solar magnetic feature detection and tracking for space weather monitoring. 2011. Citado na página 89.
- BOBRA, M. G.; COUVIDAT, S. Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, v. 798, n. 2, p. 135, 2015. Citado na página 71.
- BOULILA, W.; FARAH, I. R.; SAHEB-ETTABA, K.; SOLAIMAN, B.; BEN-GHEZALA, H. A data mining based approach to predict spatiotemporal changes in satellite images. *Int. J. Appl. Earth Observation Geoinformation*, v. 13, n. 3, p. 386–395, 2011. Citado na página 25.
- BRIAND, C. *Soleil et microphysique des plasmas naturels*. Tese (Habilitation à diriger des recherches) — Université Pierre et Marie Curie - Paris VI, 2010. Citado na página 89.
- BURBEY, I.; MARTIN, T. A survey on predicting personal mobility. *International Journal of Pervasive Computing and Communications*, v. 8, n. 1, p. 5 – 22, 2012. Citado na página 44.
- CALDIERA, V.; ROMBACH, H. D. The goal question metric approach. *Encyclopedia of software engineering*, v. 2, n. 1994, p. 528–532, 1994. Citado na página 28.
- CECI, M.; LOGLISCI, C.; RUDD, L.; MALERBA, D. Discovering knowledge through multi-modal association rule mining for document image analysis. In: *ITALIA - AI* - IA*. [S.l.: s.n.], 2015. Citado na página 51.
- CHEN, M.; MAO, S.; LIU, Y. Big data: A survey. *Mobile Networks and Applications*, v. 19, n. 2, p. 171–209, 2014. Citado na página 42.
- CHENG, C.-W.; WANG, M. D. Improving personalized clinical risk prediction based on causality-based association rules. In: *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, NY, USA: ACM, 2015. (BCB '15), p. 386–392. ISBN 978-1-4503-3853-0. Citado na página 36.
- COFFI, J.; MARSALA, C.; MUSEUX, N. Adaptive complex event processing for harmful situation detection. *Evolving Systems*, v. 3, n. 3, p. 167–177, 2012. Citado na página 36.

COMPIETA, P.; MARTINO, S. D.; BERTOLOTTI, M.; FERRUCCI, F.; KECHADI, T. Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages & Computing*, v. 18, n. 3, p. 255–279, 2007. Citado na página 44.

CUNJIN, X.; WANJIAO, S.; LIJUAN, Q.; QING, D.; XIAOYANG, W. A mutual-information-based mining method for marine abnormal association rules. *Computers and Geosciences*, v. 76, p. 121 – 129, 2015. ISSN 0098-3004. Citado na página 47.

D'AMATO, C.; STAAB, S.; TETTAMANZI, A. G. B.; MINH, T. D.; GANDON, F. Ontology enrichment by discovering multi-relational association rules from ontological knowledge bases. In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2016. (SAC '16), p. 333–338. ISBN 978-1-4503-3739-7. Citado na página 34.

DESHMUKH, J.; BHOSLE, U. Image mining using association rule for medical image dataset. *Procedia Computer Science*, v. 85, p. 117 – 124, 2016. ISSN 1877-0509. International Conference on Computational Modelling and Security (CMS 2016). Citado na página 51.

DESHMUKH, J.; BHOSLE, U. Sift with associative classifier for mammogram classification. In: *2016 International Conference on Signal and Information Processing (ICONSIP)*. [S.l.: s.n.], 2016. p. 1–5. Citado na página 62.

DESHPANDE, D.; RAJURKAR, A.; MANTHALKAR, R. Medical image analysis an attempt for mammogram classification using texture based association rule mining. In: *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on*. [S.l.: s.n.], 2013. p. 1–5. Citado na página 51.

DONG, R. c.; LIU, M.; ZHAO, J. z. Study on variation of forestland in lugu lake scenery district based on technique of spatial data mining. In: *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*. [S.l.: s.n.], 2009. v. 5, p. 215–219. Citado na página 54.

DUA, S.; SINGH, H.; THOMPSON, H. Associative classification of mammograms using weighted rules. *Expert Systems with Applications*, v. 36, n. 5, p. 9250 – 9259, 2009. ISSN 0957-4174. Citado na página 61.

EL-HAJJ, M.; ZAIANE, O. R. Parallel leap: large-scale maximal pattern mining in a distributed environment. In: *12th International Conference on Parallel and Distributed Systems - (ICPADS'06)*. [S.l.: s.n.], 2006. v. 1, p. 8 pp. ISSN 1521-9097. Citado na página 51.

ELMASRI, R.; NAVATHE, S. B. *Sistemas de Banco De Dados*. 4. ed. [S.l.]: Pearson Addison Wesley, 2005. 744 p. Citado na página 34.

FANG, G.; WU, Y. Frequent spatiotemporal association patterns mining based on granular computing. *Informatica (Slovenia)*, v. 37, n. 4, p. 443–453, 2013. Citado na página 41.

FATMA, S.; NASHIPUDIMATH, M. Image mining using association rule. In: *Information and Communication Technologies (WICT), 2011 World Congress on*. [S.l.: s.n.], 2011. p. 587–593. Citado na página 51.

- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. In: SIMOUDIS, E.; HAN, J.; FAYYAD, U. (Ed.). *KDD-96 Conference Proceedings*. [S.l.]: AAAI Press, 1996. p. 82–88. Citado 2 vezes nas páginas 33 e 34.
- FISCH, D.; JANICKE, M.; KALKOWSKI, E.; SICK, B. Learning from others: Exchange of classification rules in intelligent distributed systems. *Artificial Intelligence*, v. 187–188, n. 0, p. 90 – 114, 2012. ISSN 0004-3702. Citado na página 36.
- FOURNIER-VIGER, P.; FAGHIHI, U.; NKAMBOU, R.; NGUIFO, E. M. Cmrules: Mining sequential rules common to several sequences. *Knowledge-Based Systems*, Elsevier, v. 25, n. 1, p. 63–76, 2012. Citado na página 37.
- FU, W. Multi-media data mining technology for the systematic framework. In: *Software Engineering and Service Science (ICSESS), 2012 IEEE 3rd International Conference on*. [S.l.: s.n.], 2012. p. 570–572. Citado na página 53.
- GALVÃO, N. D.; MARIN, H. de F. Data mining: a literature review. *Acta Paulista de Enfermagem*, v. 22, p. 686–690, 2009. SciELO Brasil. Citado na página 34.
- GARCIA-FLORIANO, A.; FERREIRA-SANTIAGO, A.; CAMACHO-NIETO, O.; YENEZ-MARQUEZ, C. A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images. *Computers and Electrical Engineering*, p. 14 – 21, 2017. ISSN 0045-7906. Citado na página 61.
- GARCÍA, S.; LUENGO, J.; SáEZ, J. A.; LÓPEZ, V.; HERRERA, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 25, n. 4, p. 734–750, 2013. ISSN 1041-4347. Citado na página 57.
- GUIL, F.; MARIN, R. A tree structure for event-based sequence mining. *Knowledge-Based Systems*, v. 35, n. 0, p. 186 – 200, 2012. ISSN 0950-7051. Citado na página 36.
- HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, 2009. ISSN 1931-0145. Citado na página 93.
- HAMMAMI, H.; TURKI, S.; FAIZ, S. Classification and projection of spatial association rules. In: . [S.l.: s.n.], 2012. p. 982–986. Citado na página 35.
- HAN, J.; KAMBER, M. *Data Mining Concepts and Techniques*. 2. ed. [S.l.]: Diane Cerra, 2006. 743 p. Citado na página 34.
- HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. [S.l.: s.n.], 2012. Citado na página 35.
- HAN, J.; PEI, J.; MORTAZAVI-ASL, B.; PINTO, H.; CHEN, Q.; DAYAL, U.; HSU, M. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: *proceedings of the 17th international conference on data engineering*. [S.l.: s.n.], 2001. p. 215–224. Citado na página 37.
- HAN, J.; PEI, J.; YAN, X. Sequential pattern mining by pattern-growth: Principles and extensions. In: . [S.l.]: Springer, 2005. v. 1, p. 183–220. Citado 2 vezes nas páginas 35 e 36.

- HAN, J.; PEI, J.; YIN, Y. Mining frequent patterns without candidate generation. In: ACM. *ACM SIGMOD Record*. [S.l.], 2000. v. 29, n. 2, p. 1–12. Citado na página 36.
- HANA, A.; SAMI, Y.; SAMI, F. Mining spatiotemporal associations using queries. In: . [S.l.: s.n.], 2012. Citado 3 vezes nas páginas 45, 88 e 89.
- HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, n. 6, p. 610–621, 1973. ISSN 0018-9472. Citado na página 54.
- HIGGINS, P. A.; GALLAGHER, P. T.; MCATEER, R. J.; BLOOMFIELD, D. S. Solar magnetic feature detection and tracking for space weather monitoring. *Advances in Space Research*, Elsevier, v. 47, n. 12, p. 2105–2117, 2011. Citado na página 89.
- HONDA, R.; KONISHI, O. Temporal rule discovery for time-series satellite images and integration with rdb. In: *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*. London, UK, UK: Springer-Verlag, 2001. (PKDD '01), p. 204–215. ISBN 3-540-42534-9. Citado na página 53.
- HONDA, R.; MORI, K. Extraction of highly correlated temporal event cluster recurrence from spatiotemporal data. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. [S.l.: s.n.], 2015. p. 1457–1461. Citado na página 43.
- HU, Y.; GUNAPATI, V. Y.; ZHAO, P.; GORDON, D.; WHEELER, N. R.; HOSSAIN, M. A.; PESHEK, T. J.; BRUCKMAN, L. S.; ZHANG, G. Q.; FRENCH, R. H. A nonrelational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites. *IEEE Journal of Photovoltaics*, v. 7, n. 1, p. 230–236, 2017. ISSN 2156-3381. Citado na página 96.
- HUO, J.; ZHANG, J.; MENG, X. On co-occurrence pattern discovery from spatio-temporal event stream. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 8181 LNCS, n. PART 2, p. 385–395, 2013. Citado 3 vezes nas páginas 46, 88 e 89.
- IVANOVA, M.; KERSTEN, M.; MANEGOLD, S.; KARGIN, Y. Data vaults: Database technology for scientific file repositories. *Computing in Science & Engineering*, Institute of Electrical and Electronics Engineers, Inc., USA United States, v. 15, n. 3, p. 32–42, 2013. Citado na página 96.
- JANSSOONE, T. Temporal association rules for modelling multimodal social signals. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 2015. (ICMI '15), p. 575–579. ISBN 978-1-4503-3912-4. Citado na página 41.
- JAYABABU, Y.; VARMA, G.; GOVARDHAN, A. Incremental topological spatial association rule mining and clustering from geographical datasets using probabilistic approach. *Journal of King Saud University - Computer and Information Sciences*, p. 1–24, 2016. ISSN 1319-1578. Citado na página 41.
- JIMÉNEZ-HERNÁNDEZ, H.; HERRERA-NAVARRO, A.-M.; BARRIGA-RODRÍGUEZ, L.; CÓRDOVA-ESPARZA, D.-M.; GONZÁLEZ-BARBOSA, J.-J. A framework for developing associative classifiers based on ica. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 58, p. 88–100, 2017. Citado na página 64.

- JULEA, A.; JULEA, T.; IONESCU, C.; TELEAGA, D.; PONCOS, V. Urban area monitoring by sequential patterns extracted from multi-temporal satellite data using connectivity measures. *2014 Ieee Int. Geoscience Remote Sensing Symposium (igarss)*, p. 3582–3585, 2014. Citado na página 51.
- JULEA, A.; MEGER, N. Connectivity constraint-based sequential pattern extraction from satellite image time series. *Image Signal Processing For Remote Sensing Xix*, v. 8892, p. UNSP 889213, 2013. Citado na página 51.
- JULEA, A.; MEGER, N.; BOLON, P.; RIGOTTI, C.; DOIN, M.-P.; LASSERRE, C. L.; TROUVE, E.; LAZARESCU, V. N. Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns. *Ieee Transactions On Geoscience Remote Sensing*, v. 49, n. 4, p. 1417–1430, 2011. Citado na página 51.
- KAUFMANN, P.; MARCON, R.; CASTRO, C. G. G. de; WHITE, S. M.; RAULIN, J.-P.; CORREIA, E.; FERNANDES, L. O.; SOUZA, R. V. de; GODOY, R.; MARUN, A. *et al.* Sub-thz and h α activity during the preflare and main phases of a goes class m2 event. *The Astrophysical Journal*, IOP Publishing, v. 742, n. 2, p. 106, 2011. Citado na página 89.
- KAWALE, J.; LIESS, S.; KUMAR, V.; LALL, U.; GANGULY, A. Mining time-lagged relationships in spatio-temporal climate data. In: . [S.l.: s.n.], 2012. p. 130–135. Citado na página 45.
- KENNEDY, B.; CARRAZZA, M.; RASIN, A.; FURST, J.; RAICU, D. S. Applying association rule mining to semantic data in the lung image database consortium. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. [S.l.: s.n.], 2015. p. 463–471. Citado na página 53.
- KIM, S.; LEE, J.; RYU, K.; KIM, U. A framework of spatial co-location pattern mining for ubiquitous gis. *Multimedia Tools and Applications*, v. 71, n. 1, p. 199–218, 2012. Citado na página 42.
- KOBYLINSKI, L.; WALCZAK, K. Image classification with customized associative classifiers. In: *Proceedings of the international multiconference on computer science and information technology*. [S.l.: s.n.], 2006. p. 85–91. Citado na página 61.
- KODESWARAN, P. A.; KOKKU, R.; SEN, S.; SRIVATSA, M. Idea: A system for efficient failure management in smart iot environments. In: *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. New York, NY, USA: ACM, 2016. (MobiSys '16), p. 43–56. ISBN 978-1-4503-4269-8. Citado na página 42.
- KOUGE, Y.; MURAKAMI, T.; KUROSAWA, Y.; MERA, K.; TAKEZAWA, T. Extraction of the combination rules of colors and derived fashion images using fashion styling data. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*. [S.l.: s.n.], 2015. v. 1, p. 1–4. Citado na página 59.
- LANDGREBE, T.; MERDITH, A.; DUTKIEWICZ, A.; MAFALER, R. Relationships between palaeogeography and opal occurrence in australia: A data-mining approach. *Computers and Geosciences*, v. 56, p. 76–82, 2013. Citado na página 44.
- LAPES. *StArt*. 2011. <<http://lapes.dc.ufscar.br/ferramentas/start-tool>> último acesso em 12 de setembro de 2011. Citado na página 133.

LAUSCH, A.; SCHMIDT, A.; TISCHENDORF, L. Data mining and linked open data: New perspectives for data analysis in environmental research. *Ecological Modelling*, v. 295, n. 0, p. 5 – 17, 2015. ISSN 0304-3800. Citado na página 41.

LEE, I.; CAI, G.; LEE, K. Exploration of geo-tagged photos through data mining approaches. *Expert Systems with Applications*, v. 41, n. 2, p. 397 – 405, 2014. ISSN 0957-4174. Citado na página 35.

LEMMERICH, F.; BECKER, M.; SINGER, P.; HELIC, D.; HOTHO, A.; STROHMAIER, M. Mining subgroups with exceptional transition behavior. In: ACM. *KDD '16: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.], 2016. Citado na página 43.

LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A. *et al.* Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 115, p. 119–133, 2016. Citado na página 36.

LI, W. Remote sensing information mining of soil moisture and surface temperature based on temporal sequence association. In: *2015 Sixth International Conference on Intelligent Systems Design and Engineering Applications (ISDEA)*. [S.l.: s.n.], 2015. p. 604–607. Citado na página 59.

LI, Z.; LI, L.; YAN, K.; ZHANG, C. Automatic image annotation using fuzzy association rules and decision tree. *Multimedia Systems*, p. 1–12, 2016. ISSN 1432-1882. Citado na página 69.

LIN, M.-Y.; LEE, S.-Y. Fast discovery of sequential patterns by memory indexing. In: _____. *Data Warehousing and Knowledge Discovery: 4th International Conference, DaWaK 2002 Aix-en-Provence, France, September 4–6, 2002 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 150–160. ISBN 978-3-540-46145-6. Citado na página 37.

LIU, H.; HE, G.; JIAO, W.; WANG, G.; PENG, Y.; CHENG, B. Sequential pattern mining of land cover dynamics based on time-series remote sensing images. *Multimedia Tools and Applications*, p. 1–24, 2016. ISSN 1573-7721. Citado na página 52.

LO, H.; DING, W.; NAZERI, Z. Temporality and context for detecting adverse drug reactions from longitudinal data. *Applied Intelligence*, v. 41, n. 4, p. 1069–1080, 2014. Citado na página 35.

LU, E.-C.; HONG, J.-H.; SU, Z. L.-T.; CHEN, C. hao. A fuzzy data mining approach for remote sensing image recommendation. In: *Granular Computing (GrC), 2013 IEEE International Conference on*. [S.l.: s.n.], 2013. p. 213–218. Citado 3 vezes nas páginas 59, 88 e 89.

LU, H.; SETIONO, R.; LIU, H. Neurorule: A connectionist approach to data mining. In: DAYAL, U.; GRAY, P. M. D.; NISHIO, S. (Ed.). *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*. [S.l.]: Morgan Kaufmann, 1995. p. 478–489. Isbn 1-55860-379-4. Citado na página 33.

- MA, B. L. W. H. Y.; LIU, B. Integrating classification and association rule mining. In: *Proceedings of the fourth international conference on knowledge discovery and data mining*. [S.l.: s.n.], 1998. Citado na página 61.
- MADRAKY, A.; OTHMAN, Z.; HAMDAN, A. Analytic methods for spatio-temporal data in a nature-inspired data model. *International Review on Computers and Software*, v. 9, n. 3, p. 547–556, 2014. Citado na página 41.
- MAHMOOD, A.; SHI, K.; KHATOON, S.; XIAO, M. Data mining techniques for wireless sensor networks: A survey. *International Journal of Distributed Sensor Networks*, v. 2013, 2013. Citado na página 41.
- MANGALAMPALLI, A.; PUDI, V. Fuzzy associative rule-based approach for pattern mining and identification and pattern-based classification. In: ACM. *Proceedings of the 20th international conference companion on World wide web*. [S.l.], 2011. p. 379–384. Citado na página 61.
- MANIKONDA, L.; MANGALAMPALLI, A.; PUDI, V. Uaci: Uncertain associative classifier for object class identification in images. In: *2010 25th International Conference of Image and Vision Computing New Zealand*. [S.l.: s.n.], 2010. p. 1–8. ISSN 2151-2191. Citado na página 62.
- MANU, M. N.; ANANDAKUMAR, K. R. A current trends in big data landscape. In: IEEE. *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICIC)*. [S.l.], 2015. p. 1–6. Citado na página 96.
- MARTIN, R. C. *Clean code: a handbook of agile software craftsmanship*. [S.l.]: Pearson Education, 2009. Citado na página 27.
- MARTÍNEZ-BALLESTEROS, M.; MARTÍNEZ-ÁLVAREZ, F.; TRONCOSO, A.; RIQUELME, J. C. An evolutionary algorithm to discover quantitative association rules in multidimensional time series. *Soft Computing*, Springer, v. 15, n. 10, p. 2065–2084, 2011. Citado na página 38.
- MCATEER, R. J.; GALLAGHER, P. T.; IRELAND, J.; YOUNG, C. A. Automated boundary-extraction and region-growing techniques applied to solar magnetograms. *Solar Physics*, Springer, v. 228, n. 1-2, p. 55–66, 2005. Citado na página 89.
- MCGUIRE, M. P.; GANGOPADHYAY, A.; JANEJA, V. P. Exploring multivariate spatio-temporal change in climate data using image analysis techniques. In: *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*. New York, NY, USA: ACM, 2012. (COM.Geo '12), p. 13:1–13:10. ISBN 978-1-4503-1113-7. Citado 3 vezes nas páginas 59, 88 e 89.
- MCINTOSH, P. S. The classification of sunspot groups. *Solar Physics*, Springer, v. 125, n. 2, p. 251–267, 1990. Citado na página 73.
- MEHTA, P.; SACHARIDIS, D.; SKOUTAS, D.; VOISARD, A. Keyword-based retrieval of frequent location sets in geotagged photo trails. In: *Proceedings of the 8th ACM Conference on Web Science*. New York, NY, USA: ACM, 2016. (WebSci '16), p. 348–349. ISBN 978-1-4503-4208-7. Citado na página 36.

- MENG, T.; SHYU, M.-L. Automatic annotation of drosophila developmental stages using association classification and information integration. In: *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*. [S.l.: s.n.], 2011. p. 142–147. Citado na página 35.
- MENNIS, J.; GUO, D. Spatial data mining and geographic knowledge discovery - an introduction. *Computers, Environment and Urban Systems*, v. 33, n. 6, p. 403 – 408, 2009. ISSN 0198-9715. Spatial Data Mining - Methods and Applications. Citado na página 41.
- MINGYING, L. Study on association patterns in images based on domain knowledge-driven. *Procedia Engineering*, v. 15, p. 5553 – 5557, 2011. ISSN 1877-7058. Citado na página 52.
- MIYAMOTO, E.; MERRYMAN, T. Fast calculation of haralick texture features. *Human computer interaction institute, Carnegie Mellon University, Pittsburgh, USA. Japanese restaurant office*, 2005. Citado na página 54.
- MOHAN, A.; REVESZ, P. Applications of spatio-temporal data mining to north platter river reservoirs. In: . [S.l.: s.n.], 2014. p. 306–309. Citado 2 vezes nas páginas 35 e 43.
- MOSKOVITCH, R.; SHAHAR, Y. Medical temporal-knowledge discovery via temporal abstraction. In: *AMIA*. [S.l.: s.n.], 2009. p. 452–456. Citado na página 38.
- MOUGEL, P.-N.; SELMAOUI-FOLCHER, N. A data mining approach to discover collections of homogeneous regions in satellite image time series. *2013 Ieee Int. Geoscience Remote Sensing Symposium (igarss)*, p. 4360–4363, 2013. Citado na página 58.
- MOURI, K.; OGATA, H.; UOSAKI, N. Analysis of ubiquitous-learning logs using spatio-temporal data mining. In: *2015 IEEE 15th International Conference on Advanced Learning Technologies*. [S.l.: s.n.], 2015. p. 96–98. ISSN 2161-3761. Citado na página 47.
- NEDJAH, N.; MOURELLE, L. de M.; BUARQUE, F.; WANG, C. New trends for pattern recognition: Theory and applications. *Neurocomputing*, v. 265, p. 1 – 3, 2017. ISSN 0925-2312. New Trends for Pattern Recognition: Theory and Applications. Citado na página 61.
- NGUYEN, T. V.; VO, T. N. C. An efficient tree-based frequent temporal inter-object pattern mining approach in time series databases. H.: ĐHQGHN, 2015. Citado na página 38.
- NOAA. *SolarMonitor.org*. 2015. <<http://www.solarmonitor.org/>>. Citado 9 vezes nas páginas 71, 72, 73, 74, 76, 77, 93, 95 e 98.
- OBULESU, O.; REDDY, A. R. M. Fast and efficient mining of frequent and maximal periodic patterns in spatiotemporal databases for shifted instances. In: *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. [S.l.: s.n.], 2016. p. 35–40. Citado na página 42.
- PAI, M.; MCCULLOCH, M.; GORMAN, J.; PAI, N.; ENANORIA, W.; KENNEDY, G.; THARYAN, P.; JR, J. C. Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National Medical Journal of India*, v. 17, n. 2, p. 86–95, 2004. Citado na página 133.

- PATHANIA, S.; SINGH, H. A new associative classifier based on cfp-growth++ algorithm. In: *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015*. New York, NY, USA: ACM, 2015. (ICCCT '15), p. 20–25. ISBN 978-1-4503-3552-2. Citado na página 36.
- PETITJEAN, F.; KURTZ, C.; PASSAT, N.; GANCARSKI, P. Spatio-temporal reasoning for the classification of satellite image time series. *Pattern Recognition Lett.*, v. 33, n. 13, p. 1805–1815, 2012. Citado 4 vezes nas páginas 59, 88, 89 e 105.
- PETITJEAN, F.; MASSEGLIA, F.; GANCARSKI, P.; FORESTIER, G. Discovering significant evolution patterns from satellite image time series. *Int. J. Neural Systems*, v. 21, n. 6, p. 475–489, 2011. Citado 5 vezes nas páginas 58, 59, 88, 89 e 105.
- PETRIE, G.; SUDOL, J. J. Abrupt longitudinal magnetic field changes in flaring active regions. *The Astrophysical Journal*, IOP Publishing, v. 724, n. 2, p. 1218, 2010. Citado na página 89.
- PILLAI, K.; ANGRYK, R.; AYDIN, B. A filter-and-refine approach to mine spatiotemporal co-occurrences. In: . [S.l.: s.n.], 2013. p. 104–113. Citado 4 vezes nas páginas 47, 80, 88 e 89.
- PILLAI, K.; ANGRYK, R.; BANDA, J.; SCHUH, M.; WYLIE, T. Spatio-temporal co-occurrence pattern mining in data sets with evolving regions. In: . [S.l.: s.n.], 2012. p. 805–812. Citado 5 vezes nas páginas 42, 43, 46, 88 e 89.
- PISÓN, F. J. Martínez-de; SANZ, A.; PISÓN, E. Martínez-de; JIMÉNEZ, E.; CONTI, D. Mining association rules from time series to explain failures in a hot-dip galvanizing steel line. *Computers & Industrial Engineering*, Elsevier, v. 63, n. 1, p. 22–36, 2012. Citado na página 38.
- PITARCH, Y.; IENCO, D.; VINTROU, E.; BAGUA, A.; LAURENT, A.; PONCELET, P.; SALA, M.; TEISSEIRE, M. Spatio-temporal data classification through multidimensional sequential patterns: Application to crop mapping in complex landscape. *Engineering Applications of Artificial Intelligence*, v. 37, p. 91 – 102, 2015. ISSN 0952-1976. Citado na página 58.
- RADHAKRISHNA, V.; KUMAR, P. V.; JANAKI, V. A survey on temporal databases and data mining. In: *Proceedings of the The International Conference on Engineering and MIS 2015*. New York, NY, USA: ACM, 2015. (ICEMIS '15), p. 52:1–52:6. ISBN 978-1-4503-3418-1. Citado na página 37.
- RADHAKRISHNA, V.; KUMAR, P. V.; JANAKI, V. Mining outlier temporal association patterns. In: *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. New York, NY, USA: ACM, 2016. (ICTCS '16), p. 105:1–105:6. ISBN 978-1-4503-3962-9. Citado na página 48.
- RAHEJA, V.; RAJAN, K. Comparative study of association rule mining and mistic in extracting spatio-temporal disease occurrences patterns. In: . [S.l.: s.n.], 2012. p. 813–820. Citado na página 35.
- RAO, K. V.; GOVARDHAN, A.; RAO, K. C. Spatiotemporal data mining: Issues, tasks and applications. *International Journal of Computer Science & Engineering Survey (IJCES)* Vol, v. 3, p. 39–52, 2012. Citado na página 43.

RASHID, M.; GONDAL, I.; KAMRUZZAMAN, J. Mining associated sensor patterns for data stream of wireless sensor networks. In: . [S.l.: s.n.], 2013. p. 91–98. Citado na página 42.

RASHID, M. M.; GONDAL, I.; KAMRUZZAMAN, J. A technique for parallel share-frequent sensor pattern mining from wireless sensor networks. *Procedia Computer Science*, v. 29, n. 0, p. 124 – 133, 2014. ISSN 1877-0509. 2014 International Conference on Computational Science. Citado na página 42.

REFONAA, J.; LAKSHMI, M.; VIVEK, V. Analysis and prediction of natural disaster using spatial data mining technique. In: *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*. [S.l.: s.n.], 2015. p. 1–6. Citado na página 37.

RIBEIRO, M. X.; BUGATTI, P. H.; TRAINA, C.; MARQUES, P. M.; ROSA, N. A.; TRAINA, A. J. Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques. *Data and Knowledge Engineering*, v. 68, n. 12, p. 1370 – 1382, 2009. ISSN 0169-023X. Citado na página 64.

RIBEIRO, M. X.; TRAINA, A. J. M.; JR., C. T.; AZEVEDO-MARQUES, P. M. An association rule-based method to support medical image diagnosis with efficiency. *IEEE Transactions on Multimedia*, v. 10, n. 2, p. 277–285, 2008. ISSN 1520-9210. Citado na página 64.

RIBEIRO, M. X.; TRAINA, A. J. M.; JR, C. T.; ROSA, N. A.; MARQUES, P. M. d. A. How to improve medical image diagnosis through association rules: The idea method. In: *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. [S.l.: s.n.], 2008. p. 266–271. ISSN 1063-7125. Citado na página 64.

RIBEIRO, M. X.; TRAINA, A. J. M.; TRAINA JR., C. A new algorithm for data discretization and feature selection. In: *Proceedings of the 2008 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2008. (SAC '08), p. 953–954. ISBN 978-1-59593-753-7. Citado 3 vezes nas páginas 56, 57 e 78.

RIEDEL, I.; GUAGUEN, P.; MURA, M. D.; PATHIER, E.; LEDUC, T.; CHANUSSOT, J. Seismic vulnerability assessment of urban environments in moderate-to-low seismic hazard regions using association rule learning and support vector machine methods. *Natural Hazards*, v. 76, n. 2, p. 1111–1141, 2015. Citado na página 53.

RODRIGUES, E. O.; VITERBO, J.; CONCI, A.; MACHENRY, T. A context-aware middleware for medical image based reports. In: *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*. [S.l.: s.n.], 2015. p. 1–4. Citado na página 52.

ROMANI, L. d.; AVILA, A. D.; JR., J. Z.; CHBEIR, R.; JR., C. T.; TRAINA, A. Clearminer: A new algorithm for mining association patterns on heterogeneous time series from climate data. In: . [S.l.: s.n.], 2010. p. 900–905. Citado na página 58.

RUIZ, E.; CASILLAS, J. Adaptive fuzzy partitions for evolving association rules in big data stream. *International Journal of Approximate Reasoning*, v. 93, p. 463 – 486, 2018. ISSN 0888-613X. Citado na página 60.

- SAMIA, B.; ASSIA, K. Retrieval of high resolution satellite images based on steerable pyramids. In: *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication*. New York, NY, USA: ACM, 2015. (IPAC '15), p. 32:1–32:5. ISBN 978-1-4503-3458-7. Citado na página 51.
- SAMMOURI, W.; CAFAME, E.; OUKHELLOU, L.; AKNIN, P. Mining floating train data sequences for temporal association rules within a predictive maintenance framework. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 7987 LNAI, p. 112–126, 2013. Citado na página 41.
- SAMMOURI, W.; CAFAME E.ME, E. b.; OUKHELLOU, L. b.; AKNIN, P. b.; FONLLADOSA, C.-E. b.; PRENDERGAST, K. b. Temporal association rule mining for the preventive diagnosis of onboard subsystems within floating train data framework. In: . [S.l.: s.n.], 2012. p. 1351–1356. Citado na página 41.
- SCHEFFER, T. Finding association rules that trade support optimally against confidence. In: *Intelligent Data Anal.* [S.l.: s.n.], 1995. p. 9: 381–395. Citado na página 45.
- SCHLUTER, T.; CONRAD, S. Mining several kinds of temporal association rules enhanced by tree structures. In: IEEE. *Information, Process, and Knowledge Management, 2010. eKNOW'10. Second International Conference on.* [S.l.], 2010. p. 86–93. Citado na página 38.
- SHAHEEN, M.; SHAHBAZ, M.; GUERGACHI, A. Context based positive and negative spatio-temporal association rule mining. *Knowledge-Based Systems*, v. 37, p. 261–273, 2013. Citado na página 42.
- SHARMA, S.; BHATIA, S. Analysis of association rule in data mining. In: *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. New York, NY, USA: ACM, 2016. (ICTCS '16), p. 32:1–32:4. ISBN 978-1-4503-3962-9. Citado na página 37.
- SHEELA, L. J.; SHANTHI, V. Dimar - discovering interesting medical association rules form mri scans. In: *2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*. [S.l.: s.n.], 2009. v. 02, p. 654–658. Citado na página 61.
- SHEKHAR, S.; LI, Y.; ALI, R. Y.; EFTELIOGLU, E.; TANG, X.; JIANG, Z. Spatial and spatiotemporal data mining. In: HUANG, B. (Ed.). *Comprehensive Geographic Information Systems*. Oxford: Elsevier, 2018. p. 264 – 286. ISBN 978-0-12-804793-4. Citado na página 37.
- SHI, Y.; DENG, M.; YANG, X.; GONG, J. Detecting anomalies in spatio-temporal flow data by constructing dynamic neighbourhoods. *Computers, Environment and Urban Systems*, v. 67, p. 80 – 96, 2018. ISSN 0198-9715. Citado na página 48.
- SHU, H.; GUO, K.; ZHANG, H. Multiscale analysis of climate data in changchun, china. *Proc. SPIE*, v. 7498, p. 74982J–74982J–6, 2009. Citado na página 35.
- SILVEIRA-JUNIOR, C. R.; BERTONHA, C. *Crawler Solar Monitor*. 2015. Open source. <https://github.com/carlossilveirajr/crawler-solarmonitor>. Citado na página 75.

- SILVEIRA-JUNIOR, C. R.; CARVALHO, D. C.; SANTOS, M. T. P.; RIBEIRO, M. X. Incremental mining of frequent sequences in environmental sensor data. In: *The Twenty-Sixth International FLAIRS Conference*. [S.l.: s.n.], 2015. Citado 5 vezes nas páginas 37, 39, 42, 94 e 95.
- SILVEIRA-JUNIOR, C. R.; SANTOS, M. T. P.; RIBEIRO, M. X. Stretchy time pattern mining: A deeper analysis of environment sensor data. In: *The Twenty-Sixth International FLAIRS Conference*. [S.l.: s.n.], 2013. Citado 5 vezes nas páginas 37, 38, 42, 94 e 95.
- SILVESTRI, C.; CAGNIN, F.; LETTICH, F.; ORLANDO, S.; WACHOWICZ, M. Mining condensed spatial co-location patterns. In: *Proceedings of the Fourth ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*. New York, NY, USA: ACM, 2015. (MobiGIS '15), p. 84–87. ISBN 978-1-4503-3977-3. Citado na página 47.
- SOLINGEN, R. van; BASILI, V.; CALDIERA, G.; ROMBACH, H. D. Goal question metric (gqm) approach. In: _____. *Encyclopedia of Software Engineering*. [S.l.]: John Wiley & Sons, Inc., 2002. ISBN 9780471028956. Citado na página 28.
- SONAR, P.; BHOSLE, U. Optimized association rules for mri brain tumor classification. In: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. [S.l.: s.n.], 2016. p. 2644–2649. Citado na página 51.
- SONAR, P.; JADHAV, D.; BHOSLE, U. Optimized association rules using objective function for mammography image classification. In: *2016 International Conference on Communication and Signal Processing (ICCSP)*. [S.l.: s.n.], 2016. p. 1115–1119. Citado na página 51.
- SPACEWEATHERLIVE.COM. *Table of the characteristics of sunspot regions per Zürich/McIntosh system*. 2018. www.spaceweatherlive.com/en/help/the-classification-of-sunspots-after-malde. Citado na página 75.
- SPATENKOVAA, O.; VIRRANTAUSB, K. Discovering spatio-temporal relationships in the distribution of building fires. *Fire Safety Journal*, v. 62, Part A, p. 49 – 63, 2013. ISSN 0379-7112. Special Issue on Spatial Analytical Approaches in Urban Fire Management. Citado na página 41.
- SRIKANT, R.; AGRAWAL, R. Mining generalized association rules. In: *VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. p. 407–419. ISBN 1-55860-379-4. Citado 2 vezes nas páginas 36 e 37.
- SRIKANT, R.; AGRAWAL, R. Mining sequential patterns: Generalizations and performance improvements. In: SPRINGER. *International Conference on Extending Database Technology*. [S.l.], 1996. p. 1–17. Citado 2 vezes nas páginas 36 e 37.
- SRINIVASAN, A.; BHATIA, D.; CHAKRAVARTHY, S. Discovery of interesting episodes in sequence data. In: *Proceedings of the 2006 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2006. (SAC '06), p. 598–602. ISBN 1-59593-108-2. Citado na página 35.
- SRINIVASULU, S.; SAKTHIVEL, P. Extracting spatial semantics in association rules for weather forecasting image. In: *Trendz in Information Sciences Computing (TISC2010)*. [S.l.: s.n.], 2010. p. 54–57. ISSN 2325-5919. Citado na página 52.

- SU, F.; ZHOU, C.; LYNE, V.; DU, Y.; SHI, W. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecological Modelling*, v. 174, n. 4, p. 421 – 431, 2004. ISSN 0304-3800. Citado na página 51.
- SU, F.; ZHOU, C.; SHI, W. Geoevent association rule discovery model based on rough set with marine fishery application. In: *Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004 IEEE International*. [S.l.: s.n.], 2004. v. 2, p. 1455–1458 vol.2. Citado na página 42.
- TAYAL, K.; RAVI, V. Particle swarm optimization trained class association rule mining: Application to phishing detection. In: *Proceedings of the International Conference on Informatics and Analytics*. New York, NY, USA: ACM, 2016. (ICIA-16), p. 13:1–13:8. ISBN 978-1-4503-4756-3. Citado na página 36.
- TELIKANI, A.; SHAHBAHRAMI, A. Data sanitization in association rule mining: An analytical review. *Expert Systems with Applications*, v. 96, p. 406 – 426, 2018. ISSN 0957-4174. Citado na página 52.
- THABTAH, F. A review of associative classification mining. *The Knowledge Engineering Review*, Cambridge University Press, v. 22, n. 1, p. 37–65, 2007. Citado na página 61.
- TRAORE, B. B.; KAMSU-FOGUEM, B.; TANGARA, F. Data mining techniques on satellite images for discovery of risk areas. *Expert Systems with Applications*, v. 72, p. 443 – 456, 2017. ISSN 0957-4174. Citado na página 60.
- TSAI, F.; LAI, J.-S.; CHEN, W. W. .; LIN, T.-H. Analysis of topographic and vegetative factors with data mining for landslide verification. *Ecological Engineering*, v. 61, p. 669–677, 2013. Citado na página 54.
- VATSAVAI, R. R. Object based image classification: State of the art and computational challenges. In: *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. New York, NY, USA: ACM, 2013. (BigSpatial '13), p. 73–80. ISBN 978-1-4503-2534-9. Citado na página 52.
- VATSAVAI, R. R.; GANGULY, A.; CHANDOLA, V.; STEFANIDIS, A.; KLASKY, S.; SHEKHAR, S. Spatiotemporal data mining in the era of big spatial data: Algorithms and applications. In: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. New York, NY, USA: ACM, 2012. (BigSpatial '12), p. 1–10. ISBN 978-1-4503-1692-7. Citado 3 vezes nas páginas 25, 54 e 111.
- WANG, C.; NG, W.; CHEN, H. From data to knowledge to action: A taxi business intelligence system. In: . [S.l.: s.n.], 2012. p. 1623–1628. Citado na página 43.
- WANG, D.; DING, W.; LO, H.; MORABITO, M.; CHEN, P.; SALAZAR, J.; STEPINSKI, T. Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns. *Computers, Environment and Urban Systems*, v. 39, n. 0, p. 93 – 106, 2013. ISSN 0198-9715. Citado na página 43.
- WANG, L.; MENG, J.; XU, P.; PENG, K. Mining temporal association rules with frequent itemsets tree. *Applied Soft Computing*, v. 62, p. 817 – 829, 2018. Citado na página 51.

- WATANABE, C. Y.; RIBEIRO, M. X.; JR, C. T.; TRAINA, A. J. Sacminer: A new classification method based on statistical association rules to mine medical images. In: SPRINGER. *ICEIS*. [S.l.], 2010. p. 249–263. Citado na página 64.
- WATANABE, C. Y.; RIBEIRO, M. X.; JR, C. T.; TRAINA, A. J. Statistical associative classification of mammograms-the sacminer method. In: *ICEIS*. [S.l.: s.n.], 2010. p. 121–128. Citado na página 64.
- WATANABE, C. Y. V.; RIBEIRO, M. X.; TRAINA, A. J. M.; JR, C. T. A statistical associative classifier with automatic estimation of parameters on computer aided diagnosis. In: *2012 11th International Conference on Machine Learning and Applications*. [S.l.: s.n.], 2012. v. 1, p. 564–567. Citado na página 65.
- WAZLAWICK, R. *Metodologia de Pesquisa para Ciência da Computação, 2ª Edição*. [S.l.]: Elsevier Brasil, 2014. v. 2. Citado na página 27.
- WINARKO, E.; RODDICK, J. F. {ARMADA}: An algorithm for discovering richer relative temporal association rules from interval-based data. *Data & Knowledge Engineering*, v. 63, n. 1, p. 76 – 90, 2007. ISSN 0169-023X. Data Warehouse and Knowledge Discovery, 7th International Congress on Data Warehouse and Knowledge Discovery. Citado 4 vezes nas páginas 37, 40, 41 e 80.
- XINDONG, W.; KUMAR, V.; QUINLAN, J.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G.; NG, A.; LIU, B.; YU, P. *et al.* Top 10 algorithms in data mining. In: . [S.l.]: Springer, 2010. v. 14, n. 1, p. 1–37. ISSN 0219-1377. Citado na página 34.
- XUE, C.; LIU, J.; LI, X.; DONG, Q. Normalized-mutual-information-based mining method for cascading patterns. *ISPRS International Journal of Geo-Information*, v. 5, n. 10, 2016. ISSN 2220-9964. Citado na página 47.
- XUE, C.; SONG, W.; QIN, L.; DONG, Q.; WEN, X. A spatiotemporal mining framework for abnormal association patterns in marine environments with a time series of remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, v. 38, p. 105 – 114, 2015. ISSN 0303-2434. Citado na página 47.
- YANG, M.; KPALMA, K.; RONSIN, J. A survey of shape feature extraction techniques. *Pattern recognition*, IN-TECH, p. 43–90, 2008. Citado na página 55.
- YANG, W.; LIAO, Q.; ZHANG, C. An association rules mining algorithm on context-factors and users' preference. In: . [S.l.: s.n.], 2013. v. 1, p. 190–195. Citado na página 36.
- YAZTI, D. Z.; KRISHNASWAMY, S. Mobile big data analytics: Research, practice, and opportunities. In: *Proceedings of the 2014 IEEE 15th International Conference on Mobile Data Management - Volume 01*. Washington, DC, USA: IEEE Computer Society, 2014. (MDM '14), p. 1–2. ISBN 978-1-4799-5705-7. Citado na página 42.
- YEH, C. H.; LEE, G.; LIN, C. Y. Robust laser speckle authentication system through data mining techniques. *IEEE Transactions on Industrial Informatics*, v. 11, n. 2, p. 505–512, 2015. ISSN 1551-3203. Citado na página 52.
- YIN, G. c. An algorithm of spatial association rule mining based on concept lattice. In: *Management and Service Science, 2009. MASS '09. International Conference on*. [S.l.: s.n.], 2009. p. 1–4. Citado na página 42.

- YOO, J.; BOW, M. Mining spatial colocation patterns: A different framework. *Data Mining and Knowledge Discovery*, v. 24, n. 1, p. 159–194, 2012. Citado na página 45.
- YU, D.; HUANG, X.; HU, Q.; ZHOU, R.; WANG, H.; CUI, Y. Short-term solar flare prediction using multiresolution predictors. *The Astrophysical Journal*, IOP Publishing, v. 709, n. 1, p. 321, 2010. Citado na página 89.
- YU, W.; DONGMING, S. Premature brain damage image classification algorithm based on association rules. In: *2016 11th International Conference on Computer Science Education (ICCSE)*. [S.l.: s.n.], 2016. p. 788–791. Citado na página 52.
- ZARAGOZA, B.; RABASA, A.; RODRIGUEZ-SALA, J.; NAVARRO, J.; BELDA, A.; RAMON, A. Modelling farmland abandonment: A study combining {GIS} and data mining techniques. In: . [S.l.: s.n.], 2012. v. 155, n. 0, p. 124 – 132. ISSN 0167-8809. Citado na página 42.
- ZAYED, N.; ELNEMR, H. A. Statistical analysis of haralick texture features to discriminate lung abnormalities. *Journal of Biomedical Imaging*, Hindawi Publishing Corp., New York, NY, United States, v. 2015, p. 12:12–12:12, 2015. ISSN 1687-4188. Citado na página 54.
- ZHANG, J.; WILLIAMS, S. O.; WANG, H. Intelligent computing system based on pattern recognition and data mining algorithms. *Sustainable Computing: Informatics and Systems*, p. 123 – 131, 2017. ISSN 2210-5379. Citado na página 52.
- ZHANG, X.; LIU, J.; WANG, Q. Image feature extraction for solar flare prediction. In: *Image and Signal Processing (CISP), 2011 4th International Congress on*. [S.l.: s.n.], 2011. v. 2, p. 910–914. Citado na página 71.
- ZHAO, X.; FENG, X.; XIANG, C.; LIU, Y.; LI, Z.; ZHANG, Y.; WU, S. Multi-spacecraft observations of the 2008 january 2 cme in the inner heliosphere. *The Astrophysical Journal*, IOP Publishing, v. 714, n. 2, p. 1133, 2010. Citado na página 71.
- ZHOU, Z.; ZHANG, Y. Integration of association-rule and decision tree for high resolution image classification. In: *Geoinformatics (GEOINFORMATICS), 2013 21st International Conference on*. [S.l.: s.n.], 2013. p. 1–4. ISSN 2161-024X. Citado 2 vezes nas páginas 58 e 89.
- ZUOCHENG, W.; LIXIA, X. A fast algorithm for mining association rules in image. In: *Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on*. [S.l.: s.n.], 2014. p. 513–516. ISSN 2327-0586. Citado na página 58.
- ZURITA-MILLA, R.; GIJSEL, J. . A. . E. . van; HAMM, N. . A. . S. .; AUGUSTIJN, P. . W. . M. .; VRIELING, A. . Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps. *Ieee Transactions On Geoscience Remote Sensing*, v. 51, n. 4, p. 1914–1921, 2013. Citado na página 51.

A Revisão Sistemática

A utilização do método de Revisão Sistemática visa melhorar a qualidade do referencial teórico por meio da aplicação de métodos científicos com passos bem definidos. Este capítulo é dedicado a explicar como a revisão sistemática foi realizada no contexto desse trabalho. O processo é apresentado na Seção A.1 e os trabalhos encontrados através compõem os capítulos de referencial teórico, Capítulos 2 e 3. Neste capítulo, são apresentados os processos de revisão sistemática da seguinte forma: na Seção A.2, é apresentado o processo de revisão sobre o tema Regras de Associação Espaço-Temporais; na Seção A.3, é apresentado o processo de revisão sobre o tema Mineração de Dados em Imagens, e; na Seção A.4, é apresentado o processo de revisão sobre o tema Classificador Associativo.

A.1 O Método de Revisão Sistemática

Revisão sistemática é um método para realizar levantamento bibliográfico em trabalhos científicos que possui três etapas bem definidas: Planejamento, Seleção e Extração (PAI *et al.*, 2004).

No Planejamento determina-se o objetivo da revisão, principais questões a serem respondidas, população abordada, escopo da revisão, resultados esperados, o tipo de estudo (qualitativo ou quantitativo; observação, caracterização ou viabilização), as máquinas de buscas utilizadas, os critérios para a exclusão ou aceitação de um documento para leitura completa e o conteúdo que se pretende extrair dos documentos aceitos.

A segunda etapa, Seleção, consiste em separar os documentos para a leitura completa baseando-se no título, resumo e palavras-chave. Essa fase resulta em três classes de documentos: aceitos - que seguirão para a próxima fase; rejeitados - que infringiram critérios de aceitação e não seguirão, e; duplicados - quando o mesmo documento é retornado por buscadores distintos gerando cópias que devem ser desconsideradas. Determinam-se, também, as prioridades de leitura distribuindo-os nas classes: baixíssima, baixa, alta e altíssima.

Em seguida, os documentos aceitos passam pela etapa de Extração que consiste em extrair as informações planejadas de cada documento lido. Nessa fase, o documento ainda sofre outra classificação: aceitos, rejeitados e duplicados. Os aceitos compõem a base teórica para o trabalho, os rejeitados não. São considerados duplicados os documentos de mesma autoria com diferenças não significativas, que na prática acabam sendo rejeitados.

Para a realização deste processo foi utilizada a ferramenta de apoio StArt (LAPES, 2011) produzida no Laboratório de Pesquisa em Engenharia de *Software* (LaPES) no

Departamento de Computação da Universidade Federal de São Carlos.

A.2 Regras de Associação Espaço-Temporais

No Planejamento, as Máquinas de Busca definidas foram: *ACM*, *IEEE*, *Scopus*, *Science Direct* e *Web of Science*. As línguas definidas foram: português, inglês, francês, espanhol e italiano; assim, as *strings* de busca (aqui apresentada genericamente e em português) utilizou tanto os sinônimos dos termos como sua tradução para as outras línguas. A *string* de busca genérica foi:

```
(Mineração de Dados) E (Regras de Associação) E
  ((Dados Espaciais) OU (Dados Temporais) OU
   (Dados Espaço-Temporais))
```

Nesse etapa decidiu-se os seguintes critérios de exclusão de artigos:

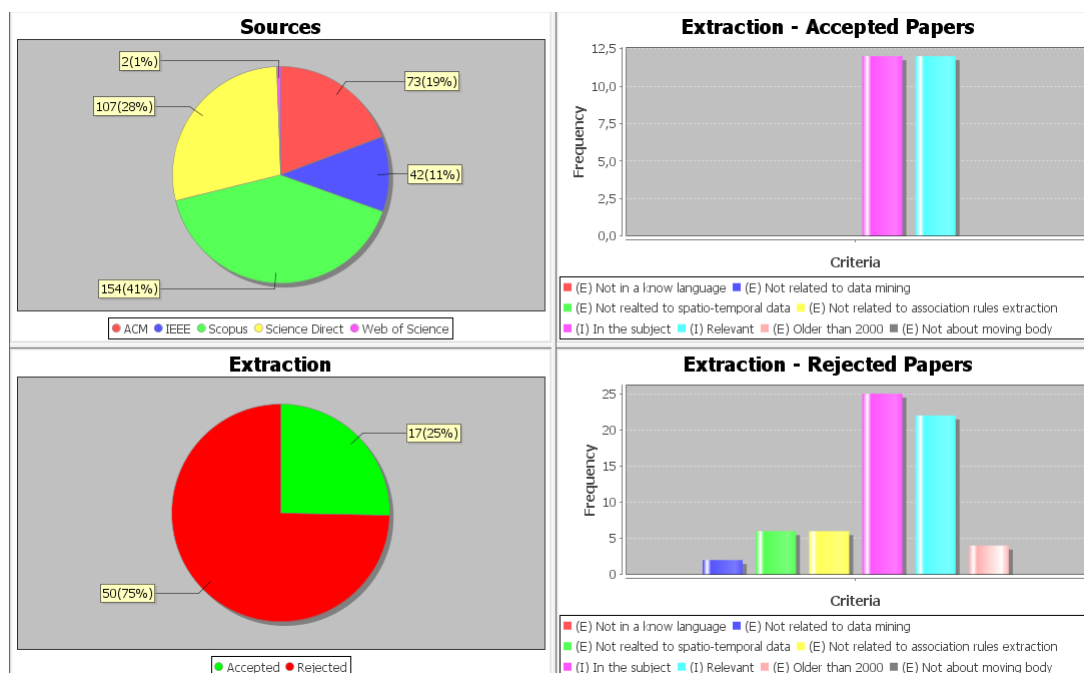
- Artigo em uma língua desconhecida;
- Artigo não relacionado a mineração de dados;
- Artigo não está relacionado a dados espaço-temporais;
- Artigo não está relacionado a extração de Regras de Associação;
- Artigo anterior a 2000, e;
- Artigo referente a dados com movimento (filmes, sons etc).

O objetivo dessa Revisão Sistemática é encontrar trabalhos correlatos: que façam a extração de regras espaço-temporais temáticas. Dando maior relevância a dados multi-dimensionais. Assim, foram definidos os seguinte critérios de aceitação:

- Referente ao assuntos, e;
- Relevante (trabalhos sobre dados multi-dimensionais).

Foram analisados 378 *papers* durante a etapa de Seleção. Foram selecionados 67 *papers* para a fase de Execução. A fase de Execução selecionou 17 *papers* como trabalhos correlatos. O outros, rejeitados pela Execução, foram usado nas definições apresentadas no Capítulo 2. Esses resultados pode ser analisados nos gráficos apresentados na Figura 27 extraídos com a ferramenta StArt.

Figura 27 – Resultado da aplicação do processo de Revisão Sistemática para Regras Espaço-Temporais.



Fonte: o próprio autor por meio da ferramenta StArt.

A.3 Mineração de Dados em Imagens

No Planejamento, as Máquinas de Busca definidas foram: *ACM*, *IEEE*, *Scopus*, *Science Direct* e *Web of Science*. As línguas definidas foram: português, inglês, francês, espanhol e italiano; assim, as *strings* de busca (aqui apresentada genericamente e em português) utilizou tanto os sinônimos dos termos como sua tradução para as outras línguas. A *string* de busca genérica foi:

(Mineração de Dados E Imagens E

(Regras de Associação OU Padrões Associativos OU Padrões Sequencias))

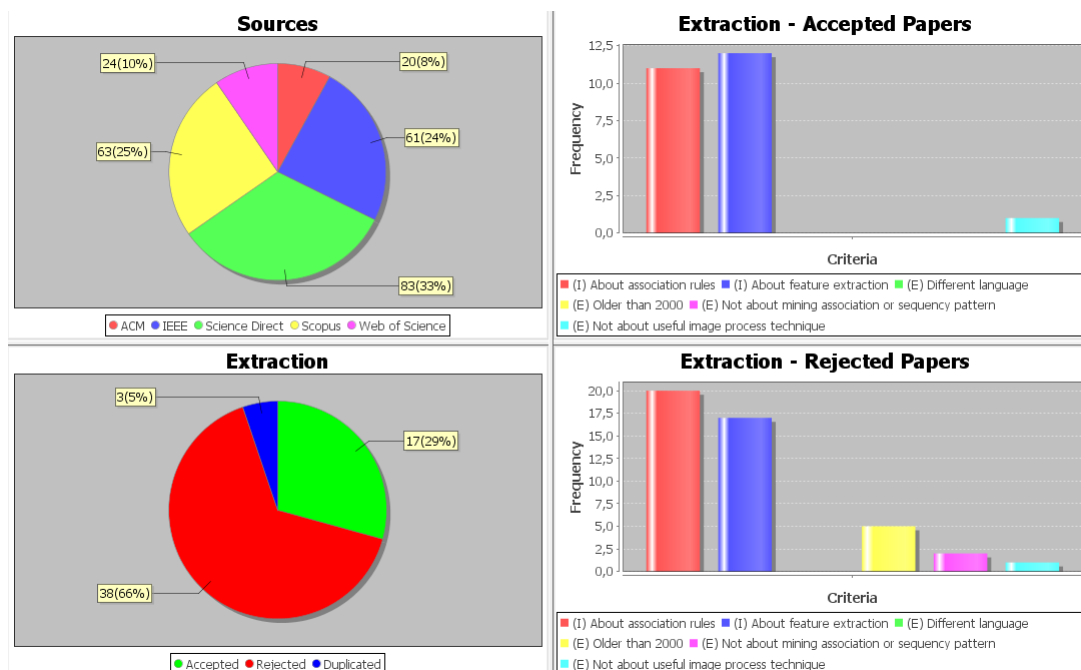
Nesse etapa decidiu-se os seguintes critérios de exclusão de artigos:

- Artigo em uma língua desconhecida;
- Trabalho não realizava a extração de regras de associação ou padrões sequenciais;
- Trabalho não utilizava imagem como fonte de informação para o algoritmo;
- Artigo anterior a 2000, e;
- Dado com movimento (filme, som etc).

O objetivo dessa Revisão Sistemática é encontrar trabalhos correlatos: que aplicam a mineração de dados a complexo. Dando maior relevância a imagens. Assim, foram definidos os seguinte critérios de aceitação:

- Aplica a extração de regras de associação, e;
- Aplica alguma técnica de extração de características às imagens.

Figura 28 – Resultado da aplicação do processo de Revisão Sistemática para Mineração de Dados aplicada a Imagens.



Fonte: o próprio autor por meio da ferramenta StArt.

Foram analisados 251 *papers* durante a etapa de Seleção. Foram selecionados 58 *papers* para a fase de Execução. A fase de Execução selecionou 17 *papers* como trabalhos correlatos. O outros, rejeitados pela Execução, foram usado nas definições apresentadas no Capítulo 3. Esses resultados pode ser analisados nos gráficos apresentados na Figura 28 extraídos com a ferramenta StArt.

A.4 Classificador Associativo

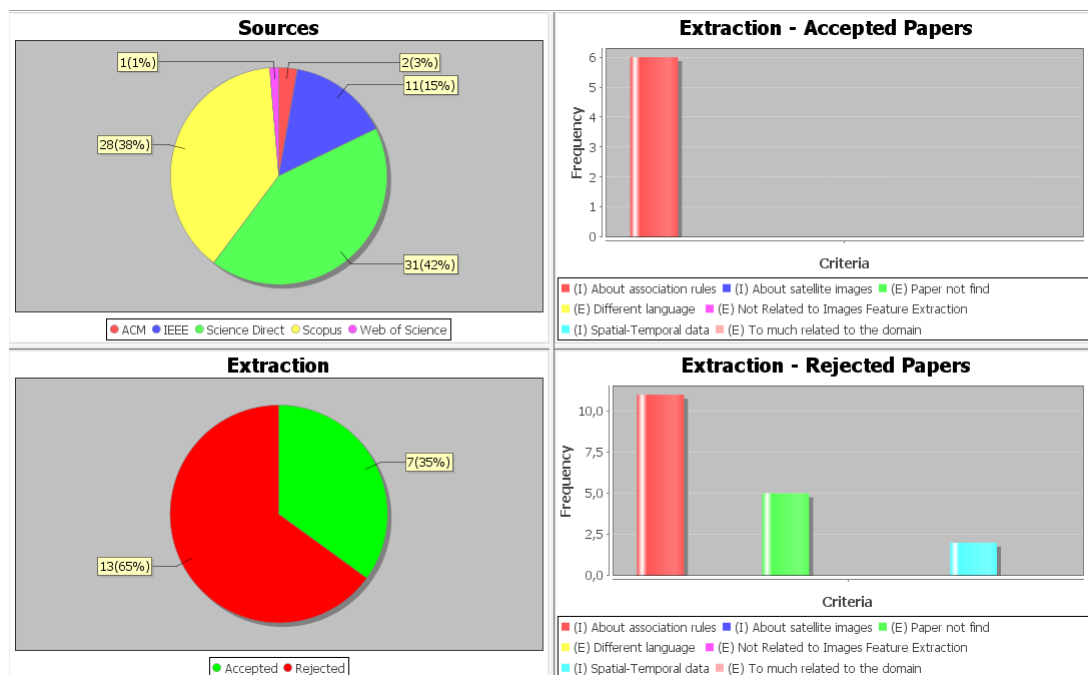
No Planejamento, as Máquinas de Busca definidas foram: *ACM*, *IEEE*, *Scopus*, *Science Direct* e *Web of Science*. As línguas definidas foram: português, inglês, francês, espanhol e italiano; assim, as *strings* de busca (aqui apresentada genericamente e em português) utilizou tanto os sinônimos dos termos como sua tradução para as outras línguas. A *string* de busca genérica foi:

(Classificador Associativo E Imagens E
(Series Temporais de Imagens OU Dados Espaço-Temporais))

Nesse etapa decidiu-se os seguintes critérios de exclusão de artigos:

- Artigo em uma língua desconhecida;
- Trabalho não relacionado a extração da características de imagens;
- Trabalho específico para um domínio;
- Artigo anterior a 2000, e;
- Artigo não encontrado.

Figura 29 – Resultado da aplicação do processo de Revisão Sistemática para Classificação de Imagens via Classificador Associativo.



Fonte: o próprio autor por meio da ferramenta StArt.

O objetivo dessa Revisão Sistemática é encontrar trabalhos correlatos: que aplicam a mineração de dados a complexo. Dando maior relevância a imagens. Assim, foram definidos os seguinte critérios de aceitação:

- Aplica a extração de regras de associação;
- No domínio de dados espaço-temporais, e;
- No domínio de Séries Temporais de Imagens de Satélite.

Foram analisados 73 *papers* durante a etapa de Seleção. Foram selecionados 20 *papers* para a fase de Execução. A fase de Execução selecionou 7 *papers* como trabalhos correlatos. O outros, rejeitados pela Execução, foram usado nas definições apresentadas no Capítulo 4. Esses resultados pode ser analisados nos gráficos apresentados na Figura 29 extraídos com a ferramenta StArt.