Jane Piantoni

Análise comparativa de técnicas avançadas de agrupamento

Sorocaba, SP

29 de Janeiro de 2016

Jane Piantoni

Análise comparativa de técnicas avançadas de agrupamento

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCCS) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Área de concentração: Inteligência Artificial.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCCS

Orientador: Katti Faceli

Sorocaba, SP

29 de Janeiro de 2016

Piantoni, Jane

Análise comparativa de técnicas avançadas de agrupamento / Jane Piantoni. -- 2016.
62 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus Sorocaba, Sorocaba
Orientador: Katti Faceli
Banca examinadora: Prof. Dr. André Carlos Ponce de Leon F. de Carvalho, Prof. Dr. Tiago Agostinho de Almeida
Bibliografia

1. Análise de Agrupamento. 2. Combinação de agrupamento. 3. Agrupamento multi-objetivo. I. Orientador. II. Sorocaba-Universidade Federal de São Carlos. III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Biblioteca campus Sorocaba (B-So).
DADOS FORNECIDOS PELO(A) AUTOR(A)

Jane Piantoni

Análise comparativa de técnicas avançadas de agrupamento

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCCS) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Área de concentração: Inteligência Artificial.

Trabalho aprovado. Sorocaba, SP, 29 de Janeiro de 2016:

Katti Faceli
Orientadora

Prof. Dr. André Carlos Ponce de Leon F. de Carvalho

Prof. Dr. Tiago Agostinho de Almeida

Sorocaba, SP
29 de Janeiro de 2016

Em memória do meu pai, Antonio Piantoni, a quem dedico este trabalho.

Agradecimentos

Agradeço,

A minha orientadora Prof. Dra. Katti Faceli, pela generosidade e perseverança ao me conduzir no desenvolvimento deste trabalho, e pelo exemplo de profissional apaixonada pelo trabalho que realiza.

Aos professores do PPGCCS que colaboraram com meu aprendizado ao longo das atividades do mestrado.

Aos colegas Angelina Melaré, Ricardo Leme, Joaquim Machado e Talita Reis Lopes, pela amizade, apoio e por todos os momentos de descontração.

A toda equipe da BSo e secretaria do PPGCCS, pelo suporte e acolhimento ao longo do mestrado.

A minha família, que sempre está ao meu lado e é fonte inesgotável de amor.

E acima de tudo a Deus, que faz caminhos e torna tudo possível.

Resumo

Este trabalho tem como objetivo investigar as características das novas abordagens de agrupamento de dados, realizando um estudo comparativo das técnicas de agrupamento que combinam ou selecionam múltiplas soluções, analisando essas técnicas mais recentes em relação a variedade e completude do conhecimento que pode ser extraído com sua aplicação. Foram realizados estudos relacionados a influência das partições base nos *ensembles* tradicionais e *ensemble* multi-objetivo. O desempenho dos métodos foi avaliado, aplicando-os em diferentes conjuntos de partições base, com o objetivo de avaliá-los com respeito a sua capacidade de identificar partições de qualidade a partir de diferentes cenários iniciais. O outro estudo realizado teve como objetivo avaliar a capacidade das técnicas em relação a recuperar as informações existentes nos dados. Para isto, foram realizadas investigações nos dois contextos: partições, que é a forma tradicional de análise e *clusters* para verificar internamente se as partições recuperadas contém mais informações relevantes do que a análise de partições demonstra. Para realizar tais análises, foram observadas a qualidade das partições e dos *clusters*, a porcentagem de informações reais (partições e *clusters*) realmente recuperadas, nos dois contextos, e o volume de informações irrelevantes que cada técnica produz. Dentre as análises realizadas, estão a busca por partições inéditas e mais robustas que o conjunto de partições base utilizados nos experimentos, a análise da influência das partições base nos *ensembles*, a análise da capacidade das técnicas na obtenção de múltiplas partições e a análise dos *clusters* extraídos.

Palavras-chaves: Análise de agrupamento. Avaliação de agrupamento. Combinação de agrupamento. Agrupamento multi-objetivo.

Abstract

The goal of this study is to investigate the characteristics of the new data clustering approaches, carrying out a comparative study of clustering techniques that combine or select multiple solutions, analyzing these latest techniques in relation to variety and completeness of knowledge that can be extracted with your application. Studies have been conducted related to the influence of partitions based on traditional ensembles and multi-objective ensemble. The performance of the methods was evaluated by applying them to different sets of base partitions, in order to evaluate them with respect to their ability to identify quality partitions from different initial scenarios. The other study, was conducted to evaluate the ability of the techniques in relation to recover the information available in the data. And for this, investigations were carried out in two contexts: partitions, which is the traditional form of analysis and clusters to internally verify that the recovered partitions contains more relevant information than the partition analysis shows. And to undertake such analyzes were observed the quality of partitions and clusters, the percentage of actual information (partitions and clusters) really recovered, in both contexts, and the volume of irrelevant information that each technique produces. Among the analyzes are the search for novel partitions and more robust than the sets of base partitions assembly used in the experiments, analysis of the influence of the partitions based on ensembles, the capacity analysis techniques in obtaining multiple partitions, and the analysis of the clusters extracted.

Key-words: Cluster analysis. Cluster evaluation. Cluster ensemble. Multi-objective clustering.

Lista de ilustrações

Lista de tabelas

Lista de símbolos

X	é um conjunto de dados com n objetos
c_k	é um *cluster* de X
Π_{TP}	é um conjunto de partições reais
Π_S	é um conjunto de partições produzidas por uma técnica de agrupamento
Π_I	é um conjunto de partições iniciais
π_I	é uma partição inicial
π_S	é uma partição obtida através de uma técnica de agrupamento
C_{TP}	é um conjunto de *clusters* das partições reais
C_S	é um conjunto de *clusters* extraído do conjunto de soluções
C_I	é um conjunto de *clusters* extraídos das partições iniciais
c_{TP}	é um *cluster* das partições reais
c_S	é um *cluster* do conjunto de *clusters* obtidos por uma técnica de agrupamento

Sumário

Introdução

Agrupamento de dados faz parte do paradigma de aprendizado de máquina não supervisionado e tem como objetivo identificar grupos nos dados (*clusters*) que apresentam semelhança de padrões e que reflitam a forma como esses dados são estruturados. Esse tipo de técnica realiza o aprendizado a partir dos dados sem que haja a necessidade de conhecimento prévio dos rótulos dos dados (JAIN; DUBES, 1988).

Diferente da classificação (supervisionada), que analisa objetos previamente rotulados, o agrupamento (*clustering*) utiliza dados cujas classes não são previamente conhecidas, explorando o conjunto de dados em busca de propriedades que o descrevam. Na literatura, pode-se encontrar técnicas de *clustering* aplicadas em diversas áreas, dentre as quais podemos destacar: a pesquisa de mercado e educação (EVERITT et al., 2009), reconhecimento de padrões (DUDA; HART, 1973), segmentação de imagens (JAIN; FLYNN, 1996), mineração de textos (STEINBACH; KARYPIS; KUMAR, 2000), análise de dados de expressão gênica (XU; WUNSCH D., 2005), descoberta de subtipos de câncer (D'HAESELEER, 2005), aplicações *Web* (SRIVASTAVA et al., 2000), entre outras.

Existem diversas algoritmos de agrupamento com características distintas entre si, porém, nenhum algoritmo de agrupamento consegue identificar todos os tipos de estruturas que possam estar presente nos dados (FACELI et al., 2011). Diante das limitações que os algoritmos de agrupamento apresentam e por não existir um conhecimento prévio dos dados que favoreça a escolha de um algoritmo e seus parâmetros ideais, novas abordagens como a combinação de agrupamentos (*clustering ensemble*) e abordagens multi-objetivo, têm sido propostas e se mostrado robustas para a descoberta das estruturas presentes nos dados. Algumas dessas novas abordagens permitem a descoberta de múltiplas soluções, onde cada uma das alternativas representa conhecimento adicional a respeito dos dados.

A combinação de agrupamentos consiste em técnicas que permitem obter uma solução que seja o consenso entre um conjunto de partições (partições base) obtidas por meio de algoritmos distintos de agrupamento, ou de um mesmo algoritmo com diferentes variações de parâmetros, ou mesmo nos dados empregados, de modo a fornecer soluções robustas e de melhor qualidade do que os algoritmos tradicionais de agrupamento (FACELI et al., 2011). A abordagem de combinação de agrupamento apresenta limitações relacionadas à influência negativa de partições base de baixa qualidade, uma vez que todas as partições base são consideradas no processo de geração da partição consenso, comprometendo o resultado final.

O agrupamento multi-objetivo permite otimizar simultaneamente mais de um critério de agrupamento (HANDL; KNOWLES, 2007), e diferente da combinação de

agrupamentos que produz apenas uma partição consenso final, o resultado do agrupamento multiobjetivo é um conjunto de soluções (partições), com variados números de *clusters*, representando diferentes compromissos entre os critérios otimizados.

O método de seleção de partições produz um conjunto de partições relevantes a partir de um conjunto de partições iniciais obtidas com critérios de agrupamento distintos. Este método descarta as partições mais similares e reduz em um conjunto das partições mais evidentes obtidas por meio de diferentes critérios de agrupamento. (SAKATA et al., 2010).

O objetivo deste trabalho consiste em investigar as características dessas novas abordagens, realizando um estudo comparativo das técnicas de agrupamento que combinam ou selecionam múltiplas soluções. Objetiva analisar essas técnicas mais recentes em relação a variedade e completude do conhecimento que pode ser extraído com sua aplicação.

Foram realizados estudos relacionados a influência das partições base nos *ensembles* tradicionais e *ensemble* multi-objetivo (FACELI et al., 2015). Esta análise foi motivada por trabalhos como Handl e Knowles (2007) e Faceli, Carvalho e Souto (2007) que indica que os *ensembles* tradicionais levam a perda de informação. O desempenho dos métodos foi avaliada, aplicando-os em diferentes conjuntos de partições base, com o objetivo de avaliá-los com respeito a sua capacidade de identificar partições de qualidade a partir de diferentes cenários iniciais.

O outro estudo realizado neste trabalho, teve como objetivo avaliar a capacidade das técnicas em relação a recuperar as informações existentes nos dados. Este estudo foi motivado pelo trabalho realizado por Faceli e Sakata (2015). Para isto, foram realizadas investigações nos dois contextos: partições, que é a forma tradicional de análise e *clusters* para verificar internamente se as partições recuperadas contém mais informações relevantes do que a análise de partições demonstra. Para realizar tais análises foram observadas a qualidade das partições e dos *clusters* através dos índices ARI e InD, a porcentagem de informações reais (partições e *clusters*) realmente recuperadas, nos dois contextos, e o volume de informações irrelevantes que cada técnica produz.

Dentre as análises realizadas, estão a busca por partições inéditas e mais robustas que o conjunto de partições base utilizados nos experimentos, a análise da influência das partições base nos *ensembles*, a análise da capacidade das técnicas na obtenção de múltiplas partições e a análise dos *clusters* extraídos.

Este trabalho está organizado da seguinte forma: o Capítulo 2 introduz as técnicas de agrupamento, o Capítulo 3 apresenta o trabalho referente ao estudo da influência das partições base nos *ensembles* e *ensemble* multi-objetivo, o Capítulo 4 apresenta os métodos de agrupamento e descreve o protocolo dos experimentos realizados e o Capítulo 5 apresenta os resultados obtidos nos experimentos e a Conclusão resume as análises e

investigações realizadas.

1 Agrupamento de Dados

As etapas que fazem parte do processo de agrupamento de dados consistem na identificação, seleção e extração de características, a definição de uma medida de similaridade entre os objetos do grupo, a identificação dos grupos e a avaliação dos resultados obtidos (JAIN; MURTY; FLYNN, 1999).

Os algoritmos de agrupamentos podem ser categorizados de diversas formas. Uma delas é em relação ao método utilizado para definir os *clusters* (JAIN; MURTY; FLYNN, 1999), em que os algoritmos podem ser divididos em hierárquicos, particionais, baseados em densidade e baseados em *grid*. Porém, sabe-se que existem alguns algoritmos que se enquadram em mais de uma categoria.

Dentre os algoritmos tradicionais, podemos citar o *Single-link* e *Average-link* que estão dentre os principais algoritmos da categoria de métodos hierárquicos e que produzem como resultado, um conjunto de partições aninhadas formando uma hierarquia de partições (BERKHIN, 2006). Já o algoritmo *K-means* representa um dos métodos particionais mais amplamente utilizados na tarefa de agrupamento (MACQUEEN, 1966). Esse algoritmo seleciona aleatoriamente **K** objetos ou pontos iniciais como centróides, e a cada iteração do algoritmo, realoca os objetos nos *clusters* de acordo com o cálculo das distâncias entre os objetos e o centro do *cluster*. Outro algoritmo clássico da literatura é a rede neural artificial SOM – (do inglês, *Self Organizing Map*) (KOHONEN, 1995), em que se associa pesos aos neurônios (objetos) em um mapa topológico de acordo com a similaridade de cada neurônio, formando regiões de neurônios próximos, em que cada região representa um *cluster*.

Algumas das recentes propostas de algoritmos de agrupamento buscam solucionar desafios específicos da análise de agrupamento. Um desses desafios é o grande volume de dados que tem sido produzido com as tecnologias atuais, tanto em termos de número de objetos, quanto em relação à dimensionalidade. Para resolvê-los, técnicas de *biclustering* e abordagem de agrupamento distribuído também têm sido propostas. Por exemplo, o algoritmo *BeeCluster*, é um algoritmo inspirado em técnicas de inteligência de enxames, cuja proposta é resolver o problema de agrupamento para dados distribuídos, sem a necessidade de inicialização de parâmetros (SANTOS; BAZZAN, 2009). O algoritmo *BeeCluster* utiliza um modelo matemático para formar grupos distribuídos de agentes com características similares. Os atributos dos dados constituem no conjunto de características dos agentes. O algoritmo inspira-se no comportamento das abelhas para formar grupos de agentes que representam dados a serem agrupados, de maneira que agentes com características similares pertençam ao mesmo grupo.

Outros problemas que podemos destacar na análise de agrupamentos, é a influência negativa que atributos irrelevantes dos objetos podem causar no resultado final do agrupamento, e também o fato de que algoritmos tradicionais não conseguem revelar estruturas de tipos diferentes presentes nos dados. Diante destes desafios, o algoritmo MVGEN (do inglês, *Multi-View Generative Model*) (GüNNEMANN; FäRBER; SEIDL, 2012) permite realizar o agrupamento selecionando um conjunto de atributos dos objetos. Com este algoritmo é possível identificar os atributos mais relevantes, e com isso, encontrar os melhores pontos de vista que definem as partições encontradas nestes subconjuntos dos dados.

Mesmo diante da diversidade de técnicas, não existe uma técnica de agrupamento que possa resolver todos os problemas de agrupamento (JAIN; MURTY; FLYNN, 1999), ou seja, não existe uma técnica específica que possa identificar todas as estruturas presentes ou relevantes, em qualquer conjunto de dados. Alguns algoritmos necessitam que parâmetros de inicialização sejam informados, como o número de grupos (*clusters*), ou a densidade máxima ou mínima de cada grupo, e com isso, limita-se a descoberta de estruturas com número de *clusters* previamente estabelecidos. Além disso, cada algoritmo pode apresentar habilidade para encontrar *clusters* de tipos específicos, de acordo com seu critério de agrupamento, e com isso, pode não apresentar bom desempenho para determinados conjuntos de dados, que contenham *clusters* de outros tipos. Por exemplo, alguns algoritmos podem apresentar habilidade em encontrar *clusters* esféricos, porém, nem sempre os dados apresentam *clusters* com essa conformação.

Diante das limitações que os algoritmos de agrupamento apresentam e por não existir um conhecimento prévio dos dados que favoreça a escolha de um algoritmo e seus parâmetros ideais, várias abordagens de combinação de agrupamentos (*ensembles*), agrupamento multi-objetivo e de seleção de partições têm sido propostas e têm se mostrado robustas para a descoberta das estruturas presentes nos dados.

Nas próximas seções, serão abordadas as técnicas que permitem identificar conjuntos de soluções, resultados mais robustos ou soluções novas que não seriam identificadas por métodos tradicionais, sendo elas: combinação de agrupamento (*ensemble*), agrupamento multi-objeto e de seleção de partições.

1.1 Técnicas de interesse

Nesta seção são apresentadas as técnicas de interesse deste trabalho, cuja característica principal é tentar diminuir uma das limitações das técnicas de agrupamento tradicionais. Na revisão da literatura, nota-se que as técnicas mais recentes fazem uso das características positivas de outras abordagens, combinando-as com o objetivo de superar os problemas de cada uma e obter soluções de melhor qualidade.

1.1.1 Combinação de Agrupamento

As técnicas mais tradicionais de combinação de agrupamento, *Cluster Ensembles*, consistem em técnicas que permitem obter uma solução que seja o consenso entre um conjunto de partições base, de modo a fornecer soluções robustas e de melhor qualidade do que os algoritmos tradicionais de agrupamento (FACELI et al., 2011).

De acordo com Topchy, Jain e Punch (2003), o processo de combinação de agrupamentos baseia-se em aplicar uma função consenso em um conjunto de n partições chamadas de partições base $\Pi_I = (\pi_1, \pi_2, ..., \pi_n)$ resultantes de várias aplicações de um ou mais algoritmos de agrupamento em um determinado conjunto de dados X, de maneira a obter uma partição consenso π_S que apresenta melhor qualidade que as partições base.

A melhor qualidade da partição consenso está relacionada ao objetivo que se deseja atingir. Alguns dos principais objetivos estão relacionados à:

- Robustez, ou seja, obtenção de uma partição consenso mais robusta que as partições base;
- Novidade, ao se obter uma partição inédita não presente no conjunto de partições base;
- Consistência ao obter uma partição consenso que esteja em concordância com as partições base;
- Estabilidade ao se obter soluções de agrupamento com menor sensibilidade a ruídos, *outliers*, variações de amostragem ou à variabilidade dos algoritmos (FACELI et al., 2011).

Existe uma grande variedade de funções consenso definidas na literatura, cada uma apresentando uma metodologia específica para realizar a combinação. Dentre elas podemos citar:

- Funções baseadas em co-associação, que calculam a similaridade entre pares de objetos, pelo número de grupos compartilhados entre eles em todas as partições base. A similaridade representa a força de co-associação entre os pares de objetos e é representada por uma matriz de co-associação (FRED; JAIN, 2005);
- Funções baseadas em grafo ou hipergrafo, ou seja, funções que extraem o consenso das partições base, utilizando métodos de manipulação de grafos ou hipergrafos (STREHL; GHOSH, 2002);
- Funções baseadas em probabilidade, ou seja, funções consenso baseadas em uma solução probabilística, onde se calcula a probabilidade de um objeto estar presente

em um partição, baseado na informação presente no conjunto de partições base (TOPCHY; JAIN; PUNCH, 2003).

Os métodos de combinação de agrupamentos selecionados para os experimentos deste trabalho foram: CSPA – (do inglês, *Clusterbased Similarity Partitioning Algorithm*), HGPA – (do inglês, *HiperGraph-Partitioning Algorithm*) e MCLA – (do inglês, *metaCLustering Algorithm*), algoritmos proposto por Strehl e Ghosh (2002) baseados em grafos e hipergrafos, o algoritmo BCE – (do inglês, *Bayesian Cluster Ensembles*) proposto por Wang, Shan e Banerjee (2011) que é um método baseado em probabilidade e o método CTS – (do inglês, *Connected-triple-based similarity*) parte do *framework LinkClue* de Iamon e Garrett (2010), baseado em co-associação.

A seguir, são brevemente descritas as características desses métodos, assumindo que todos podem ser aplicados a um mesmo conjunto de partições base.

Dentre as características dos métodos, o algoritmo CSPA gera uma matriz de similaridade utilizando as partições base e, em uma próxima etapa, essa matriz será usada para construção de um grafo, onde os pesos das arestas são dados pelos valores da matriz. Após este processo, o grafo de similiaridade é particionado pelo algoritmo METIS – (do inglês, *Multilevel Graph Partitioning Algorithm*) (KARYPIS et al., 1999), em um número de grupos balanceados cujo número é definido pelo usuário, onde o objetivo é dividir o grafo eliminando as arestas com o menor peso.

No MCLA cada *cluster* das partições base é considerado um vértice de um meta-grafo. Após a construção do meta-grafo, é realizada a combinação dos grupos, particionando o meta-grafo também usando o algoritmo METIS. O objetivo deste processo é encontrar os grupos das partições base que são correspondentes, gerando meta-grupos. Ao unir os *clusters*de cada meta-grupo, combina as hiperarestas em uma única metahiperaresta para cada meta-grupo. Cada objeto pode pertencer a mais de uma meta-grupo. Para cada meta-hiperaresta é realizado o cálculo do vetor de associação descrevendo o nível de associação de cada objeto com o meta-grupo. Esse nível de associação é obtido a partir da média dos vetores que representam as hiperarestas de um meta-grupo. Cada objetivo é associado ao *meta-cluster* para qual ele possui o maior valor de associação. A partição dos objetos indicada pelos *meta-clusters* é a partição π_S resultante do *ensemble*.

Já a função consenso do algoritmo HGPA trata-se de um problema de particionamento de um hipergrafo, onde cada hiperaresta representam *clusters*. Se existir um caminho entre dois vértices no hipergrafo, os objetos representados por esses vértices estarão no mesmo grupo na partição consenso. O particionamento do hipergrafo é realizado pelo HMETIS que é uma extensão do METIS para hipergrafos, que realiza corte nas hiperarestas, até que um número de grupos pré-definidos seja obtido na partição final consenso.

Diferente dos métodos anteriores, o *ensemble* BCE apresenta uma abordagem *bayesiana* baseada em probabilidade em sua função consenso (WANG; SHAN; BANERJEE, 2011). No processo de combinação do BCE, um objeto pode ser relacionado a várias possibilidades de combinação em seu processo generativo. Ao final, a função consenso irá associar o objeto ao

2 Estudo da influência das partições base nos ensembles

Com o intuito de avaliar a influência dos conjuntos de partições bases nos resultados dos métodos de *ensemble* e *ensemble multiobjetivo*, foram realizados experimentos com esses *ensembles*, aplicando-os em diferentes conjuntos de partições base, com o objetivo de avaliá-los com respeito a sua capacidade de identificar partições de qualidade a partir de diferentes cenários iniciais. Para isso, foi utilizado um conjunto de dados artificial, chamado de *2sp2glob*, e a partir desse conjunto de dados foram gerados artificialmente diversos conjuntos de partições base contemplando as propriedades de interesse para a análise. Conforme ilustra a Figura 1, o conjunto de dados artificial *2sp2glob* apresenta 2000 objetos, contém 4 *clusters* com 500 objetos cada, sendo 2 *clusters* em formato de espiral e 2 *clusters* em formato globular.

Os resultados desse estudo foram apresentados na 22nd International Conference on Neural Information Processing (ICONIP2015) (FACELI et al., 2015). Os conjuntos de dados e os conjuntos de partições base (BP) utilizados neste trabalho estão disponíveis em http://lasid.sor.ufscar.br/2sp2globBPCollection/

Nas próximas seções, serão descritas as características dos conjuntos de partições base empregados nos experimentos, os protocolo dos experimentos e os resultados obtidos.

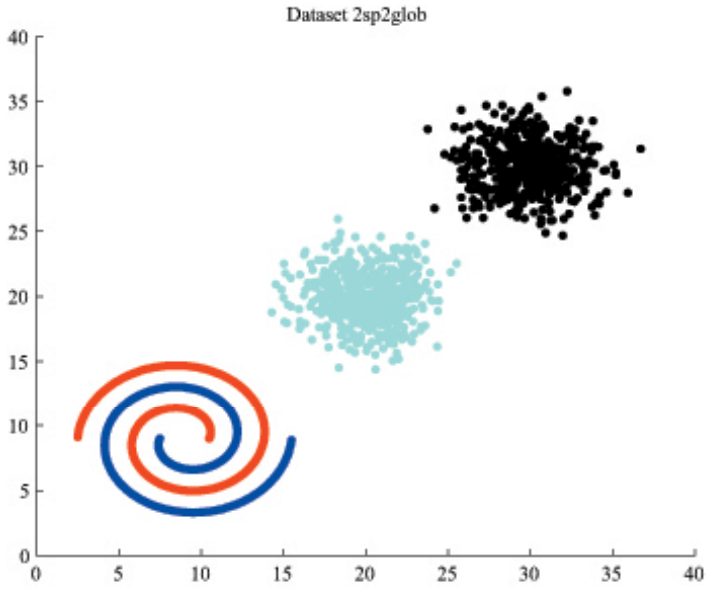

Figura 1 – Conjunto de Dados 2sp2glob

2.1 Conjuntos de partições base

Para a criação das partições base, foram gerados artificialmente três conjuntos de partições a partir do conjunto de dados *2sp2glob*, cada um contendo algumas propriedades de interesse para esta análise, tais como grande número de subdivisões dos *clusters* reais, número de *clusters* próximo ao número da partição real, partições geradas de maneira aleatória. A Tabela 1 contém um resumo das características dos conjuntos de partições, que serão detalhados a seguir.

Tabela 1 – Conjuntos Artificiais de Partições Base

Conjunto	**Descrição**	**Nº Cluster (k)**
Conjunto 1	12 partições (muitas subdivisões)	15 a 24
Conjunto 2	12 partições (contendo 1 *cluster* completo)	10 a 22
Conjunto 3	12 partições (poucas subdivisões)	5 a 8

O conjunto 1 é composto por 12 partições geradas artificialmente contendo entre 15 e 24 *clusters*. Todas as partições contêm diversas subdivisões de todos os *clusters* reais. Dentre as várias partições, há subdivisões que se complementam e se sobrepõem parcialmente.

O conjunto 2 também apresenta 12 partições contendo entre 10 e 22 *clusters*. Porém, diferente do conjunto 1, cada uma das partições apresenta 1 *cluster* completo e correto como na partição real, e os demais *clusters* representam subdivisões dos *clusters* reais. Na Figura 2, a partição possui um *cluster* completo com formato globular. Os *clusters* de todas essas partições foram criados de maneira que também exista sobreposição entre pedaços dos *clusters*, como no conjunto 1.

O conjunto 3 contém 12 partições onde o número de *clusters* varia entre 5 a 8. Cada uma dessas partições contém um ou mais *cluster* completo e correto, e também subdivisões dos *clusters* originais, como pode ser observado na Figura 3, porém em menor número do que os conjuntos 1 e 2. O número de *clusters* de cada partição é próximo ao número de *clusters* da partição real, k = 4.

Com base nos conjuntos 1, 2 e 3, foram criados 9 conjuntos de partições base (BP), $BP1$ a $BP9$. A seguir serão descritos cada um desses conjuntos, juntamente com o objetivo de utilizá-los nas análises.

- **BP 1:** Em BP1 foram utilizadas apenas as partições do conjunto 1, em que nenhum dos *clusters* estava completo (todos apresentavam subdivisões). O intuito neste caso foi analisar a capacidade dos métodos em encontrar uma partição com os *cluster* reais completos como partição consenso, a partir de informações parciais fornecidas nas partições base.

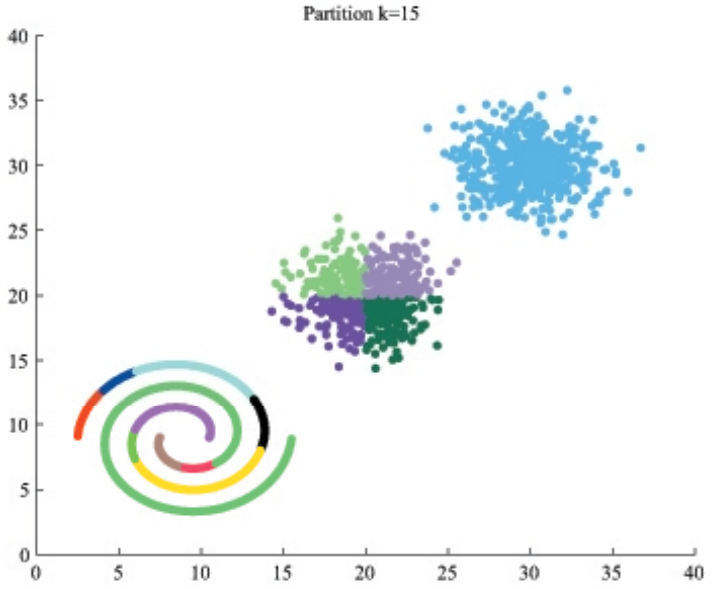

Figura 2 – Partição k=15

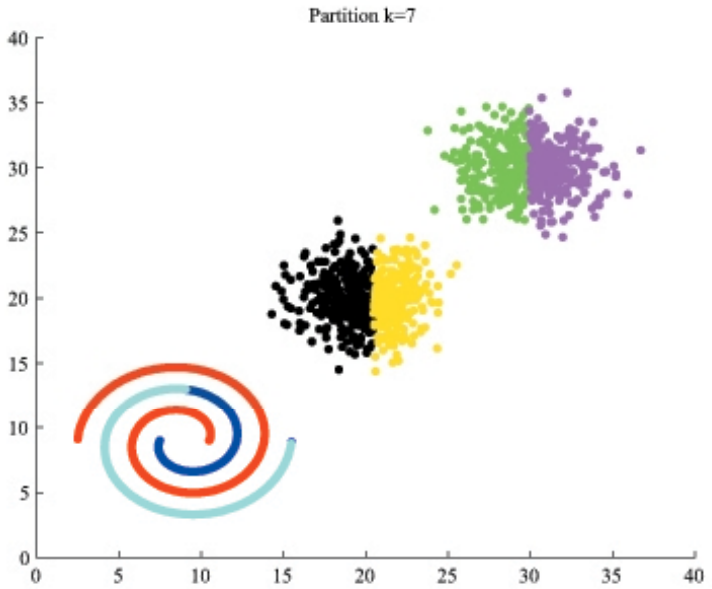

Figura 3 – Partição k=7

- **BP 2:** Em BP2 foram utilizadas apenas as partições do conjunto 2, que apresentavam um dos *clusters* correto e completo. Neste caso, o objetivo foi analisar se a presença do *cluster* completo nas partições poderia influenciar o resultado da partição consenso.

- **BP 3:** Em BP 3 foram utilizadas as partições do conjunto 3, com *clusters* com poucas subdivisões e *clusters* completos. Neste caso, o intuito foi avaliar se os *ensembles* apresentariam melhor desempenho tendo a presença de partições mais parecidas com a partição real, do que o obtido com partições apresentando muitas subdivisões.

- **BP 4, 5 e 6:** esses conjuntos de partições correspondem aos conjuntos BP 1, 2 e 3, respectivamente, com o acréscimo da partição real. Com isso, buscou-se analisar o quanto as técnicas conseguem usar essa informação para melhorar a qualidade do consenso, ou a perde no processamento que faz.

- **BP 7, 8 e 9:** esses conjuntos de partições também correspondem aos conjuntos BP 1, 2 e 3, respectivamente, com o acréscimo de duas partições aleatórias em cada caso. Essas partições foram geradas com a distribuição aleatória dos objetos nos *clusters*, gerando partições de baixíssima qualidade quando comparadas com a partição real. Em cada conjunto, foram adicionadas 2 partições com números de *clusters* iguais ao menor e ao maior k do conjunto inicial, respectivamente. Em BP 7 foram acrescentadas partições 15 e 24, em BP 8 foram acrescentadas partições com 10 e 22 *clusters* e em BP 9 foram acrescentadas partições com 5 e 8 *clusters*. Com isso, buscou-se avaliar o quanto cada técnica é influenciada pela existência de partições de baixa qualidade entre as partições base.

2.2 Protocolo dos experimentos

Nos experimentos realizados com os métodos de combinação de agrupamento, foram realizadas execuções dos métodos BCE - (do inglês, *Bayesian Cluster Ensembles*)(WANG; SHAN; BANERJEE, 2011), CSPA - (do inglês, *Clusterbased Similarity Partitioning Algorithm*), HGPA - (do inglês, *HiperGraph-Partitioning Algorithm*) e MCLA - (do inglês, *metaCLustering Algorithm*)(STREHL; GHOSH, 2002), com diferentes valores do parâmetros k, ou seja, número de *clusters* a ser encontrado. Em virtude da necessidade de comparar o desempenho dos métodos, foi informado o número de partições, iniciando com k = 4, que é o número de *clusters* da partição real, até k = 8. Foi calculado o índice *ARI* entre cada partição obtida e a partição real. Para se avaliar os resultados, foi selecionada a partição com o maior *ARI* dentre as partições obtidas com cada técnica, uma vez que o objetivo era avaliar se as técnicas eram capazes de identificar soluções novas e de qualidade e não explorar suas variações de parâmetros ou outras características.

Nos experimentos realizados com o MOCLE - (do inglês, *Multi-Objective Clustering Ensemble*) (FACELI et al., 2009), o operador de recombinação selecionado foi o MCLA, e foram utilizados os parâmetros default do algoritmo L = 5 e G = 100 relacionados à conectividade e ao número de gerações. Estes valores foram definidos após testes preliminares com diversas combinações de valores desses parâmetros e a realização do teste estatístico Friedman (DEMSAR, 2006) o qual demonstrou não haver diferença significativa entre os resultados do método obtidos através de diferentes valores de parâmetros ($p > 0{,}05$). Para realizar os experimentos, foram realizadas 30 execuções do MOCLE, e foram calculados o índice *ARI* das partições obtidas. O próximo passo foi selecionar a melhor

partição de cada execução, e realizar o cálculo da média e do desvio padrão do conjunto formado pelas das 30 partições selecionadas. Cada uma das técnicas foi aplicada da maneira descrita a cada um dos conjuntos de partições base.

2.3 Análise dos Resultados

Para efeito de comparação entre os resultados obtidos pelas técnicas de agrupamento, utilizou-se o índice Rand Ajustado - ARI (HUBERT; ARABIE, 1985) para avaliação das partições em relação a partição real. E para avaliar a diferença entre os métodos e os conjuntos de partições base, foi aplicado o teste de *Friedman* (DEMSAR, 2006).

A Tabela 2 apresenta o resultado das melhores partições obtidas por cada algoritmo de acordo com o índice ARI calculado. Na coluna $BPID$, estão identificados os conjuntos de partições base utilizados nos experimentos, a coluna BP apresenta o valor de ARI da melhor partição de cada conjunto de partições base.

Pôde-se observar que nas BP 4, 5 e 6, onde foi incluída a partição real, a maioria dos algoritmos de *ensemble* apresentou resultado do ARI superior, se comparado aos resultados dos experimentos iniciais, BP 1, 2 e 3. Demonstrando que partições de melhor qualidade influenciam positivamente o resultado da função consenso. Já os experimentos com o algoritmo multi-objetivo, demonstrou que o MOCLE não foi influenciado, mantendo o resultado médio do ARI muito próximo aos resultados dos experimentos anteriores.

Nos experimentos realizados com as BP 7, 8 e 9, pode-se observar que ao inserir partições aleatórias no conjunto de partições base, o resultado dos *ensembles* foi influenciado de maneira negativa, gerando partições que se comparadas aos experimentos anteriores, apresentam baixa qualidade, conforme ARI calculado.

Quando observa-se os resultados da Tabela 2, pode-se constatar que nos diferentes experimentos com as BP 1, 2, 7 e 8, o MOCLE apresentou valores superiores a melhor partição base utilizada no experimento. Isto significa que ele foi capaz de descobrir partições de alta qualidade que não existia no conjunto de partições base (novidade).

Tabela 2 – Resultado das melhores partições - ARI

BP ID	BP	BCE	CSPA	MCLA	HGPA	LinkClue	MOCLE
BP1	0.5204	0.5134	0.5078	0.5165	0.5285	0.4985	0.9736 (0.0465)
BP2	0.7226	0.5678	0.5743	0.6021	0.5832	0.5602	0.9678 (0.0532)
BP3	0.9354	0.6084	0.6112	0.5991	0.6002	0.5873	0.9232 (0.1156)
BP4	1	0.5208	0.5122	0.5321	0.5397	0.5011	0.9743 (0.0343)
BP5	1	0.6032	0.5832	0.5973	0.5899	0.5532	0.9433 (0.0723)
BP6	1	0.6008	0.6218	0.6199	0.6078	0.5993	0.9671 (0.0854)
BP7	0.5204	0.3832	0.3944	0.4231	0.3755	0.3933	0.9429 (0.0935)
BP8	0.7226	0.3784	0.3983	0.3678	0.3813	0.3631	0.9534 (0.0843)
BP9	0.9354	0.3988	0.3786	0.4009	0.4023	0.3931	0.9734 (0.0721)

Com o intuito de verificar se os dados amostrados fornecem evidência suficiente de que as diferenças observadas nos dados não são meramente casuais. Foram realizadas análises estatísticas considerando os algoritmos como os tratamentos e as BPs como blocos. A hipótese testada foi, se não existe diferença entre os algoritmos $H0$, e se existe diferença estatisticamente significativa entre pelo menos dois algoritmos $H1$.

Ao aplicar o teste de *Friedman*, obteve-se p-valor = $0.003 < 0.05$. Sendo rejeitada $H0$ e podendo-se concluir que existe diferente entre pelos menos 2 algoritmos ou entre um algoritmo e o conjunto de BPs.

A fim de verificar quais algoritmos diferem entre si, foi aplicado o pós-teste de *Nemenyi*. Os resultados confirmaram que o MOCLE apresenta performance superior em relação aos algoritmos de *ensemble* tradicionais. Também foi constatado que os *ensembles* tradicionais (BCE, CSPA, MCLA, HGPA e LinkClue) não apresentaram diferenças entre si.

Em outra análise, foi comparada a influência das diferentes condições iniciais (BPs) na performance dos algoritmos. Foram consideradas as BPs como tratamentos e os algoritmos de *ensemble* como blocos. Sendo testada a hipotese nula $H0$, onde as condições iniciais não influenciaram o resultado dos algoritmos, e $H1$ onde pelo menos duas condições inicias influenciam a performance dos algoritmos.

Aplicando o teste de *Friedman*, obteve-se p-valor de $0.004 < 0{,}05$. Desta forma, rejeitou-se a $H0$, concluindo que, pelo menos, duas BPs apresentaram um desempenho diferente nos *ensembles* tradicionais.

Desta forma, foi aplicado o pós-teste de *Nemenyi*, através dele constatou-se que a $BP6$ teve um desempenho diferente das BPs 7, 8 e 9; e que a $BP3$ apresenta diferença em relação as $BPs8e9$. Ou seja, as BPs que continham partições aleatórias influenciaram o desempenho dos *ensembles* tradicionais, especificamente quando comparado as BPs que continham menor número de subdivisões ($BP3$ e $BP6$).

2.4 Conclusões

Os resultados observados em relação aos métodos tradicionais de *ensemble* demonstram baixa performance em todos os experimentos realizados com os conjuntos de partições base (BPs).

Nos experimentos realizados com BPs que continham partições aleatórias, foi constatada a influência negativa dessas partições no resultado final dos *ensembles*. Mesmo os experimentos em que a BP continha a partição real, não foi suficiente para que o *ensemble* pudesse recuperar ou fornecer um resultado próximo a estrutura real.

Nos experimentos realizados com o método de *ensemble* multi-objetivo, foram constatadas soluções de alta qualidade, mesmo em situações em que a BP não continha a partição real, ou seja, continha apenas a informação parcial dos dados, os experimentos demonstraram que a condição inicial não influenciou negativamente o resultado final.

Desta forma os resultados dos experimentos demonstraram que quando comparado aos métodos de *ensemble* tradicionais, o método de *ensemble* multi-objetivo demonstrou ser mais adequado para dados que contêm estruturas heterogéneas (*clusters* de formas e tamanhos diferentes).

3 Métodos e Experimentos

Neste trabalho, além da análise da influência das partições base nos *ensembles*, também foram feitos experimentos com diversos métodos avançados de agrupamento. O objetivo foi avaliá-los quanto a sua capacidade de produzir diferentes estruturas a partir dos conjuntos de dados. Assim, foram avaliados métodos de combinação de agrupamento tradicionais e multi-objetivo (*ensembles*), uma técnica de agrupamento multi-objetivo e uma estratégia de seleção de partições. Para isso, este métodos foram aplicados em diferentes conjuntos de dados de dados e os conjuntos de soluções resultantes foram avaliadas de diversas maneiras.

O objetivo desses experimentos foi avaliar a capacidade das técnicas em encontrar partições inéditas e mais robustas que o conjunto de partições base utilizados e também verificar a capacidade das técnicas em identificar a maioria dos *clusters* relevantes.

Nas próximas seções serão abordados os conjuntos de dados utilizados neste trabalho, bem como o processo de geração das partições base, execução dos métodos de agrupamento e os índices de validação empregados.

3.1 Conjuntos de dados

Nas análises realizadas, foram utilizados 37 conjuntos de dados com diferentes número de objetos, dimensionalidade, formatos e possibilidades de agrupamento (partições conhecidas).

Dentre os conjuntos de dados utilizados, 15 deles são conjuntos de dados produzidos artificialmente com diferentes propriedades de interesse para avaliação dos métodos de agrupamento e 22 correspondem a dados reais de áreas como a Medicina e Bioinformática.

Na Tabela 3 estão relacionadas as características dos conjuntos de dados, sendo n o número de objetos, d a dimensionalidade (número de atributos), np^{TP} número de partições reais e $K^{\pi^j \in \Pi_{TP}}$ número de *clusters* de cada $\pi^j \in \Pi_{TP}$ e nc^{TP} número de *clusters* distintos em C_{TP}.

Conjuntos de Dados Artificiais:

`atom`, `engyTime`, `hepta`, `lsun`, `target`, `tetra`, `twoDiamonds` e `wingNut` fazem parte da *Suite Fundamental Clustering Problems* , disponível em <http://www.uni-marburg.de/fb12/datenbionik/data?language_sync=1>. Esses conjuntos possuem uma diversidade de tipos de *clusters*, utilizados como *benchmark* de algoritmos de agrupamento (ULTSCH, 2005);

Tabela 3 – Conjuntos de Dados Reais e Artificiais

Tipo	**Conjunto de Dados**	n	d	np^{TP}	$K^{\pi^j \in \Pi_{TP}}$	nc^{TP}
Artificial	`atom`	800	3	1	2	2
	`ds2c2sc13`	588	2	3	2, 5, 13	19
	`ds3c3sc6`	905	2	2	3, 6	8
	`ds4c2sc8`	485	2	2	2, 8	10
	`engyTime`	4096	2	1	2	2
	`gaussian`	60	600	1	3	3
	`hepta`	212	3	1	7	7
	`lsun`	400	2	1	3	3
	`monkey`	4000	2	4	8,5,3,2	14
	`simulated6`	60	600	1	6	6
	`spiralsquare`	2000	2	2	2, 6	8
	`target`	770	2	1	6	6
	`tetra`	400	3	1	4	4
	`twoDiamonds`	800	2	1	2	2
	`wingNut`	1016	2	1	2	2
Real	`armstrong`	72	1081	2	2,3	4
	`chowdary`	104	182	1	2	2
	`contractions`	98	27	1	2	2
	`dyrskjot`	40	1203	1	3	3
	`eTongueSugar`	375	6	2	2,3	5
	`glass`	214	9	3	2, 5, 6	9
	`golub`	72	3571	4	2, 3, 2, 4	10
	`gordon`	181	1626	1	2	2
	`iris`	150	4	1	3	3
	`laryngeal1`	213	16	1	2	2
	`laryngeal2`	692	16	1	2	2
	`laryngeal3`	353	16	2	2,3	4
	`libras`	360	90	2	8,15	21
	`lung`	197	1000	1	4	4
	`miRNAcancer`	218	217	6	3, 20, 4, 9, 2, 2	40
	`respiratory`	85	17	1	2	2
	`segmentation`	2310	19	1	7	7
	`su`	174	1571	1	10	10
	`voice3`	238	10	2	2,3	4
	`voice9`	428	10	2	2,9	10
	`weaning`	302	17	1	2	2
	`yeoh`	248	2526	2	2, 6	7

`gaussian` e `simulated6` são conjuntos artificiais que simulam dados de expressão gênica de alta dimensionalidade. Disponível em <http://www.broadinstitute.org/cgi-bin/cancer/publications/view/87> (MONTI et al., 2003);

`ds2c2sc13`, `ds3c3sc6`, `ds4c2sc8`, `spiralsquare` e `monkey` foram criados para a finalidade de explorar a diversidade de tipos de *clusters* em estruturas heterogêneas e a existência de múltiplas partições reais. Cada um dos conjuntos contem pelo menos duas estruturas, representando diferentes níveis de refinamento da mesma informação. `ds2c2sc13`, `ds3c3sc6`, `ds4c2sc8` e `spiralsquare` foram previamente descritos em Faceli et al. (2010); `spiralsquare` foi construído a partir de dois conjuntos de dados descritos em Handl e Knowles (2007) e `monkey`.

Conjuntos de Dados Reais:

`contractions`, `laryngeal1`, `laryngeal2`, `laryngeal3`, `respiratory`, `voice3`, `voice9` e `weaning` apresentam dados reais do domínio da Medicina e foram disponibilizados pelo *Pattern Recognition Group of School of Computer Science* da Universidade de Bangor <http://pages.bangor.ac.uk/~mas00a/activities/real_data.htm>.

`glass`, `iris`, `libras` e `segmentation` fazem parte do repositório *UCI Machine Learning* <http://archive.ics.uci.edu/ml/> (NEWMAN et al., 1998).

`armstrong`, `chowdary`, `dyrskjot`, `golub`, `gordon`, `lung`, `miRNAcancer`, `su` e `yeoh` fazem parte do domínio da Bio-informática e foram descritos originalmente em Armstrong et al. (2002) , Chowdary et al. (2006), Dyrskjøt et al. (2003), Golub et al. (1999), Gordon et al. (2002), Bhattacharjee et al. (2001), Lu et al. (2005), Su et al. (2001), Yeoh et al. (2002) respectivamente. Neste trabalho foram usadas as versões dos conjuntos de dados descritos por Faceli et al. (2009).

`eTongueSugar` foi desenvolvido a partir de uma combinação do conjuntos de dados *E-Tongue Sugar Collections* v.1 <http://www.dcomp.sor.ufscar.br/talmeida/etonguesugar/index.html>, descrito em Sakata et al. (2012). Refere-se a avaliação da qualidade do açúcar.

Os conjuntos de dados citados neste trabalho estão disponíveis, como parte do *benchmark* para avaliação de *clusters* do LASID - *Clusters Evaluation Benchmark* <http://lasid.sor.ufscar.br/clustersEvaluationBenchmark/>

3.2 Experimentos

Nesta seção será abordado o protocolo dos experimentos realizados neste trabalho. Inicialmente, será detalhado o processo de obtenção das partições base, execução das técnicas de interesse e objetivo dos experimentos, obtenção de partições finais e a extração dos *clusters* individuais. Também, será abordado o procedimento escolhido para análise e avaliação dos resultados obtidos (partições e *clusters*).

3.2.1 Partições Base - Iniciais

Os conjuntos de partições base Π_I utilizados neste trabalho foram gerados previamente em outros trabalhos desenvolvidos no LASID a partir de algoritmos tradicionais de agrupamento (*Average-link, Centroid-link, Complete-link, k-means, Single-link e Shared Nearest Neighbors*), com configurações e parâmetros variados, com o intuito de produzir conjuntos diversificados de partições e *clusters*. Como resultado, para cada conjunto de dados, obteve-se um conjunto de partições com diferentes números de *clusters*.

Os conjuntos de partições utilizados neste trabalho estão disponíveis para acesso no *website* do LASID - *Clusters Evaluation Benchmark* <http://lasid.sor.ufscar.br/clustersEvaluationBenchmark/>

Estas partições foram utilizadas no processo de agrupamento dos métodos de *ensemble*, *ensemble* multi-objetivo e seleção de partições.

3.2.2 Aplicação dos Métodos de Interesse

Nos experimentos realizados foram empregados os *ensembles* BCE (WANG; SHAN; BANERJEE, 2011), LinkClue (IAMON; GARRETT, 2010) com a matriz de similaridade CTS e função consenso *Single Link*, CSPA, HGPA e MCLA (STREHL; GHOSH, 2002). Todos os *ensembles* foram executados diversas vezes, cada vez solicitando um número de *clusters* diferentes, dentro de um intervalo de interesse. Mais especificamente foi usado um intervalo, iniciando com k = ao menor k das partições reais (estruturas conhecidas), até 2 vezes o k da partição real com maior número de *clusters*.

Foi calculado o índice *ARI* entre cada partição obtida e a partição real. Para se avaliar os resultados, foi selecionada a partição com o maior *ARI* dentre as partições obtidas com cada técnica, uma vez que o objetivo era avaliar se as técnicas eram capazes de identificar soluções novas e de qualidade e não explorar suas variações de parâmetros ou outras características.

Para os experimentos realizados com o método de combinação multi-objetivo MOCLE (FACELI et al., 2009), o operador de recombinação selecionado foi o método MCLA, e utilizados os valores default dos parâmetros L = 5 e G = 100, relacionados respectivamente à conectividade e ao número de gerações. Para o intervalo de k, foi empregado o mesmo descrito para os *ensembles*.

Na execução do método de agrupamento multi-objetivo MOCK, define-se o parâmetro L, este parâmetro é utilizado no processo de inicialização, mutação e nas definições de conectividade. Nos experimentos realizados foi adotado o valor *default* L = 10, que permite a detecção de *clusters* mesmo em conjuntos de dados pequenos (HANDL; KNOWLES, 2007). Também foram configurados os parâmetros relacionados ao número de gerações (*number of generations*) = 250 e ao número de soluções iniciais (*initial solutions* = 50).

Por não serem determinísticos e por gerarem um conjunto de partições, a seleção das melhores partições dos métodos multi-objetivo foi realizada da seguinte maneira:

- Foram realizadas 30 execuções dos métodos multi-objetivo;
- Foi calculado o índice *ARI* de cada partição obtida em relação a cada partição real dos conjuntos de dados;

- Foram selecionadas as partições em cada uma das execuções que apresentaram o melhor valor de *ARI* (maiores valores);
- Foi calculada a média das melhores partições selecionadas;

O método de seleção de partições ASA (SAKATA et al., 2010), possui um parâmetro referente ao número de partiçoes idênticas a considerar no processo de seleção, neste trabalho foram utilizados os resultados previamente obtidos pelo ASA em execuções disponíveis no *Cluster Evaluation Benchmark*, não tendo sido realizada a execução do método para este trabalho, já que as configurações do *benchmark* já eram adequadas para as análises a serem realizadas aqui.

4 Resultados

Neste capítulo, serão apresentados os resultados dos experimentos realizados. Será feita uma análise comparativa do desempenho dos métodos em relação aos conjuntos de dados e partições base descritos no Capítulo 4.

As análises realizadas não se limitam a avaliar somente as partições, mas consideram os *clusters* como soluções. O intuito é investigar a capacidade das técnicas em recuperar estruturas (partições e *clusters*) de alta qualidade presentes nos conjuntos de soluções.

Para efeito das análises, serão utilizados os resultados dos índices de validação ARI e InD, e também será empregado o teste estatístico de *Friedman* para verificar se os dados amostrados fornecem evidência suficiente de que as diferenças observadas nos dados não são meramente casuais.

4.1 Análise no Contexto de Partições

Nesta seção, serão avaliadas as partições obtidas através dos métodos de combinação de agrupamento (*ensemble*), agrupamento multi-objetivo, *ensemble* multi-objetivo e seleção de partições.

O resultado do índice ARI será utilizado para análises de desempenho dos métodos em cada um dos conjuntos de dados. Para realizar tais análises foram observadas a qualidade das partições obtidas e a porcentagem de partições reais recuperadas.

4.1.1 Capacidade de Recuperação das Partições Reais

As Tabelas 4 e 5 apresentam os resultados dos experimentos dos métodos de combinação de agrupamento, multi-objetivo e de seleção de partições, em relação ao valor do índice ARI das melhores partições, conforme descrito no Capítulo 4. Nas tabelas estão relacionados os conjuntos de dados utilizados e o valor de ARI da melhor partição dentre as iniciais. Para cada partição real π_{TP} conhecida de cada conjunto de dados, foram destacados os resultados dos métodos que apresentaram o melhor valor (maior ARI).

Nesta análise, pôde-se observar que o método de *ensemble* multi-objetivo conseguiu gerar soluções de alta qualidade, mesmo em situações em que o conjunto de partições iniciais continha partições de baixa qualidade. Observa-se também que, na maior parte dos experimentos com *ensembles*, os melhores resultados do índice ARI obtido pelos métodos foram inferiores a melhor partição inicial (π_I).

Um exemplo é o conjunto de dados `ds4c2sc8`, onde o ARI da melhor partição

inicial $\pi_1 = 0{,}346$, e a melhor partição gerada pelo MOCLE apresentou ARI $\pi_f = 0{,}821$. Já os métodos de *ensemble* apresentaram resultadores inferiores a $\pi_1 = 0{,}346$, sendo o resultado do método MCLA $\pi_f = 0{,}203$ e LinkClue $\pi_f = 0{,}132$.

O MOCK utiliza somente o conjunto de dados em seu processo de agrupamento, e também conseguiu gerar partições de melhor qualidade do que o conjunto de partições iniciais gerados pelos algoritmos tradicionais de agrupamento. O método de seleção ASA, conseguiu selecionar partições de qualidade a partir do conjunto de partições iniciais.

Nota-se que nenhum dos métodos de *ensemble* foi capaz de recuperar alguma das partições reais dos conjuntos de dados. Já os métodos multi-objetivo e de seleção de partições foram capazes de encontrar partições reais em sua execução.

Observa-se que, os valores médio, máximo e mínimo das técnicas de *ensemble* apresentam valores parecidos, e os resultados dos métodos multi-objetivo e de seleção de partições, apresentam valores superiores aos *ensembles*.

Tabela 4 – ARI Métodos - Dados Artificiais

Dataset	TP	π_I	CSPA	HGPA	MCLA	BCE	LinkClue	MOCK	MOCLE	ASA
atom	π_1	1	0,495	0,375	0,413	0,386	0,364	1	1	1
ds2c2sc13	π_1	1	0,479	0,403	0,432	0,375	0,326	1	1	1
	π_2	0,997	0,554	0,465	0,503	0,387	0,339	0,802	0,982	0,872
	π_3	1	0,356	0,334	0,496	0,404	0,410	0,893	0,931	1
ds3c3sc6	π_1	0,590	0,357	0,385	0,404	0,557	0,406	0,990	0,853	0,590
	π_2	0,899	0,556	0,452	0,413	0,395	0,384	0,807	0,816	0,899
ds4c2sc8	π_1	0,346	0,304	0,264	0,203	0,297	0,132	0,867	0,821	0,346
	π_2	0,869	0,376	0,354	0,410	0,325	0,376	0,798	0,800	0,815
engyTime	π_1	0,815	0,545	0,436	0,718	0,376	0,311	1	1	0,815
gaussian	π_1	1	0,398	0,384	0,512	0,265	0,401	0,932	1	1
hepta	π_1	0,971	0,421	0,686	0,352	0,402	0,282	0,897	0,965	0,971
lsun	π_1	1	0,412	0,392	0,284	0,374	0,365	1	1	0,997
monkey	π_1	0,870	0,554	0,486	0,643	0,386	0,397	0,790	0,991	0,808
	π_2	0,855	0,476	0,443	0,613	0,385	0,387	0,823	0,803	0,777
	π_3	0,521	0,434	0,503	0,439	0,384	0,587	0,887	0,783	0,521
	π_4	0,834	0,502	0,634	0,523	0,400	0,394	0,702	0,799	0,746
simulated6	π_1	0,985	0,413	0,385	0,375	0,297	0,354	0,876	0,921	0,985
spiralsquare	π_1	1	0,479	0,395	0,405	0,385	0,375	0,843	1	1
	π_2	0,666	0,407	0,416	0,523	0,453	0,396	0,721	0,920	0,666
target	π_1	1	0,374	0,453	0,427	0,485	0,453	1	1	1
tetra	π_1	1	0,451	0,442	0,413	0,424	0,445	1	1	1
twoDiamonds	π_1	1	0,523	0,414	0,610	0,398	0,392	0,834	1	1
wingNut	π_1	1	0,436	0,452	0,387	0,394	0,695	1	1	1
Média		0,879	0,448	0,433	0,456	0,388	0,390	0,889	0,930	0,883
Máximo		1,000	0,556	0,686	0,718	0,557	0,695	1,000	1,000	1,000
Mínimo		0,346	0,304	0,264	0,203	0,265	0,132	0,702	0,783	0,346

Tabela 5 – ARI Métodos - Dados Reais

Dataset	TP	π_I	CSPA	HGPA	MCLA	BCE	LinkClue	MOCK	MOCLE	ASA
armstrong	π_1	0,694	0,403	0,507	0,521	0,422	0,390	0,576	0,723	0,602
	π_2	0,494	0,450	0,584	0,419	0,389	0,338	0,634	0,694	0,407
chowdary	π_1	0,065	0,030	0,045	0,043	0,042	0,043	0,096	0,125	0,062
contractions	π_1	0,397	0,331	0,296	0,301	0,209	0,285	0,464	0,547	0,397
dyrskjot	π_1	0,628	0,506	0,514	0,538	0,520	0,528	0,588	0,755	0,628
eTongueSugar	π_1	0,054	0,114	0,121	0,183	0,190	0,166	0,058	0,062	0,054
	π_2	0,704	0,432	0,398	0,302	0,283	0,359	0,497	0,685	0,639
glass	π_1	0,671	0,433	0,578	0,395	0,430	0,345	0,682	0,754	0,666
	π_2	0,559	0,498	0,375	0,334	0,320	0,299	0,697	0,785	0,547
	π_3	0,263	0,420	0,392	0,512	0,443	0,310	0,512	0,498	0,256
golub	π_1	0,943	0,432	0,654	0,410	0,294	0,312	0,785	0,876	0,876
	π_2	0,727	0,454	0,392	0,493	0,402	0,367	0,689	0,776	0,666
	π_3	0,314	0,420	0,390	0,372	0,586	0,410	0,601	0,589	0,314
	π_4	0,865	0,411	0,420	0,398	0,384	0,432	0,564	0,675	0,798
gordon	π_1	0,174	0,123	0,137	0,143	0,102	0,140	0,295	0,334	0,174
iris	π_1	0,821	0,400	0,428	0,642	0,469	0,366	0,896	0,975	0,759
laryngeal1	π_1	0,100	0,162	0,159	0,178	0,190	0,110	0,193	0,234	0,085
laryngeal2	π_1	0,099	0,137	0,185	0,150	0,195	0,101	0,201	0,283	0,099
laryngeal3	π_1	0,303	0,296	0,213	0,245	0,202	0,197	0,284	0,397	0,303
	π_2	0,136	0,105	0,123	0,130	0,116	0,090	0,301	0,375	0,136
libras	π_1	0,343	0,413	0,387	0,294	0,354	0,382	0,721	0,771	0,325
	π_2	0,218	0,394	0,401	0,407	0,323	0,255	0,576	0,533	0,217
lung	π_1	0,642	0,432	0,463	0,634	0,382	0,333	0,545	0,683	0,642
miRNAcancer	π_1	0,415	0,334	0,306	0,314	0,390	0,323	0,497	0,597	0,409
	π_2	0,661	0,446	0,433	0,385	0,342	0,284	0,476	0,643	0,550
	π_3	0,594	0,375	0,334	0,473	0,294	0,304	0,398	0,407	0,471
	π_4	0,301	0,243	0,210	0,296	0,238	0,235	0,387	0,421	0,235
	π_5	0,300	0,206	0,202	0,223	0,134	0,190	0,401	0,398	0,300
	π_6	0,221	0,202	0,195	0,204	0,130	0,191	0,196	0,323	0,221
respiratory	π_1	0,123	0,113	0,113	0,110	0,114	0,113	0,194	0,284	0,123
segmentation	π_1	0,403	0,344	0,434	0,392	0,397	0,345	0,498	0,563	0,403
su	π_1	0,578	0,430	0,404	0,385	0,343	0,320	0,576	0,602	0,547
voice3	π_1	0,125	0,100	0,109	0,103	0,110	0,097	0,136	0,225	0,083
	π_2	0,114	0,099	0,094	0,100	0,076	0,096	0,113	0,178	0,083
voice9	π_1	0,092	0,502	0,595	0,492	0,452	0,397	0,221	0,213	0,092
	π_2	0,102	0,522	0,453	0,410	0,393	0,340	0,129	0,145	0,102
weaning	π_1	0,082	0,096	0,073	0,095	0,043	0,024	0,108	0,241	0,082
yeoh	π_1	0,941	0,632	0,523	0,493	0,434	0,402	0,897	0,967	0,907
	π_1	0,215	0,221	0,244	0,199	0,235	0,190	0,495	0,586	0,215
Média		0,389	0,322	0,328	0,327	0,292	0,264	0,439	0,521	0,364
Máximo		0,943	0,632	0,654	0,642	0,586	0,528	0,897	0,975	0,907
Mínimo		0,054	0,030	0,045	0,043	0,042	0,024	0,058	0,062	0,054

4.1.2 Análise Estatística

Com o intuito de verificar se houve diferença significativa entre os métodos, foi aplicado o teste Friedman (DEMSAR, 2006).

O teste foi aplicado nos resultados de *ARI* obtidos nos experimentos com os conjuntos de dados reais e artificiais, testando as hipóteses H0: não há diferença entre os métodos vs. H1: pelo menos dois métodos diferem entre si. O nível de significância adotado foi de 5%.

A estatística observada resultou no p-valor de 0,0008. Como o p-valor é menor que 0,05, rejeitou-se H0 ao nível de 5% de significância e concluiu-se que há diferença entre pelo menos dois métodos.

Como os resultados de *ARI* entre as técnicas de um mesmo tipo foram bastante semelhantes para cada conjunto de dados, foi aplicado o teste de *Friedman* somente entre os resultados dos *ensembles*, e depois somente entre os resultados dos métodos multi-objetivo e de seleção de partições, a fim de verificar quais métodos apresentam diferenças entre si.

No teste aplicado entre os *ensembles*, a estatística observada resultou no p-valor de 0,703. Como o p-valor é maior que 0,05, não se rejeita H0, isto é, não há evidências de que exista diferença, ao nível de 5% de significância, entre os métodos comparados.

Já o teste aplicado entre os métodos multi-objetivo e de seleção de partições, a estatística observada resultou no p-valor de 0,007. Como o p-valor é menor que 0,05 rejeitou-se H0 ao nível de 5% de significância e concluiu-se que há diferença entre pelo menos dois métodos. Desta forma foi aplicado o pós teste de *Nemenyi*. Os resultados do teste (*p-valor*) da comparação dois a dois, conclui que o método *MOCLE* difere dos demais métodos ao nível de 5% de significância, e que os demais não diferem entre si.

4.1.3 Tamanho dos conjuntos de soluções

Diferente dos métodos de combinação de agrupamento, cuja execução resulta em uma única partição consenso, os métodos multi-objetivo e de seleção de partições apresentam como resultado final um conjunto de partições ao final da execução.

A Tabela 6 apresenta os números de partições dos conjuntos soluções obtidos com o MOCLE, MOCK e ASA. Em cada execução desses métodos, foram obtidos conjuntos de partições (Π_S) contendo partições com diferentes números de *clusters*. A coluna Π_{TP} apresenta o número de partições reais de cada conjuntos de dados, a coluna Π_I apresenta o conjunto de partições iniciais (π_I) geradas através dos algoritmos tradicionais de agrupamento.

De acordo com a Tabela 6, o método MOCLE resultou o maior número de partições na soma dos conjuntos de soluções (Π_S), um total de 632 partições. Porém, o método ASA apresentou valor muito próximo ao MOCLE, um total de 604 partições. O método MOCK apresentou um número menor de partições na soma dos conjuntos de soluções, 35% menor que o MOCLE, em virtude dos parâmetros utilizados em sua execução.

Tabela 6 – Conjuntos de soluções - Partições

Tipo	Dataset	Π_{TP}	Π_I	MOCK	MOCLE	ASA
Artificial	atom	1	21	7	13	11
	ds2c2sc13	3	147	13	41	21
	ds3c3sc6	2	65	19	30	25
	ds4c2sc8	2	89	20	29	29
	engyTime	1	18	24	32	16
	gaussian	1	45	8	5	18
	hepta	1	77	6	7	25
	lsun	1	29	9	12	10
	monkey	4	75	15	27	22
	simulated6	1	58	8	5	15
	spiralsquare	2	85	9	25	18
	target	1	55	13	18	24
	tetra	1	53	7	17	13
	twoDiamonds	1	24	7	7	12
	wingNut	1	17	10	15	5
Real	armstrong	2	27	6	5	8
	chowdary	1	15	4	6	4
	contractions	1	17	8	9	16
	dyrskjot	1	25	10	7	16
	eTongueSugar	2	25	6	8	11
	glass	3	55	13	21	15
	golub	4	49	9	11	25
	gordon	1	15	6	9	9
	iris	1	37	11	12	8
	laryngeal1	1	18	8	7	10
	laryngeal2	1	15	13	15	10
	laryngeal3	2	25	9	11	15
	libras	2	122	15	39	16
	lung	1	40	12	12	17
	miRNAcancer	6	218	22	46	36
	respiratory	1	22	5	4	9
	segmentation	1	40	11	34	12
	su	1	95	19	35	23
	voice3	2	32	7	7	11
	voice9	2	96	17	32	35
	weaning	1	24	7	8	12
	yeoh	2	55	17	11	22
	Total	62	1925	410	632	604

Para efeito de comparação do desempenho dos métodos, foram resumidos o número das melhores partições recuperadas por cada método em relação aos conjuntos de dados. Nesta análise, os *ensembles* não foram incluídos, pois nenhum dos métodos conseguiu recuperar a partição real, e na maior parte dos casos, os valores de *ARI* calculados nos experimentos, apresentaram valores inferiores a 0,70.

De acordo como a Tabela 7, no contexto das análises realizadas, o método MOCLE foi o que apresentou o maior número de partições reais recuperadas, ao menos parcialmente (*ARI* superior a 0,7), atingindo um total de 51,61% das partições recuperadas, porém, MOCK e ASA apresentaram valores muito próximos.

Tabela 7 – Partições Recuperadas pelos métodos

Tipo	Dataset	np^{TP}	ARI = 1			ARI > 0,7		
			MOCK	MOCLE	ASA	MOCK	MOCLE	ASA
Artificial	atom	1	1	1	1	1	1	1
	ds2c2sc13	3	1	1	2	3	3	3
	ds3c3sc6	2	0	0	0	2	2	2
	ds4c2sc8	2	0	0	0	2	2	2
	engyTime	1	1	1	0	1	1	1
	gaussian	1	0	1	1	1	1	1
	hepta	1	0	0	0	1	1	1
	lsun	1	1	1	0	1	1	1
	monkey	4	0	0	0	4	4	3
	simulated6	1	0	0	0	1	1	1
	spiralsquare	2	0	1	1	2	2	1
	target	1	1	1	1	1	1	1
	tetra	1	1	1	1	1	1	1
	twoDiamonds	1	0	1	1	1	1	1
	wingNut	1	1	1	1	1	1	1
Real	armstrong	2	0	0	0	0	1	0
	chowdary	1	0	0	0	0	0	0
	contractions	1	0	0	0	0	0	0
	dyrskjot	1	0	0	0	0	1	0
	eTongueSugar	2	0	0	0	0	0	0
	glass	3	0	0	0	0	2	0
	golub	4	0	0	0	1	2	2
	gordon	1	0	0	0	0	0	0
	iris	1	0	0	0	1	1	1
	laryngeal1	1	0	0	0	0	0	0
	laryngeal2	1	0	0	0	0	0	0
	laryngeal3	2	0	0	0	0	0	0
	libras	2	0	0	0	1	1	0
	lung	1	0	0	0	0	0	0
	miRNAcancer	6	0	0	0	0	0	0
	respiratory	1	0	0	0	0	0	0
	segmentation	1	0	0	0	0	0	0
	su	1	0	0	0	0	0	0
	voice3	2	0	0	0	0	0	0
	voice9	2	0	0	0	0	0	0
	weaning	1	0	0	0	0	0	0
	yeoh	2	0	0	0	1	1	1
	Total	62	7	10	9	27	32	25

A Figura 4 apresenta o resumo do desempenho dos métodos MOCK, MOCLE e ASA, em relação ao percentual de partições recuperadas, cujo o valor de *ARI* calculado seja maior que 0,7 e igual a 1. O MOCLE obteve a maior porcentagem de partições recuperadas integralmente e parcialmente.

Os experimentos demonstraram que poucas partições reais foram descobertas pelos métodos, ou seja, através da avaliação de partições, pôde-se constatar que os métodos não conseguiram recuperar informação de qualidade, principalmente quando trata-se dos *ensembles*, que apresentaram partições de baixa qualidade. Este fato, motiva as análises realizadas na próxima seção, onde busca-se analisar os *clusters* individualmente, considerando-os como conjuntos de soluções relevantes presentes nas partições.

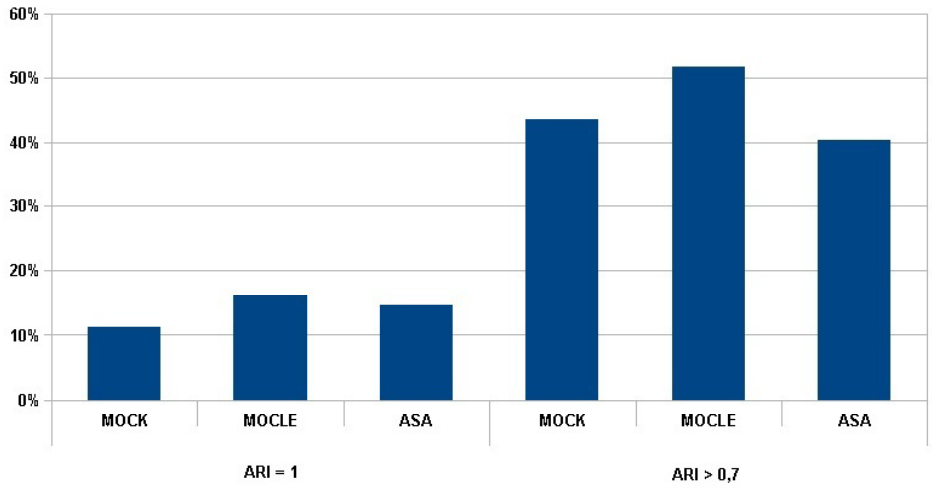

Figura 4 – Partições Recuperadas pelos métodos

4.2 Análise no Contexto de Clusters

Nesta seção, serão realizadas análises voltadas aos *clusters*, considerando os *clusters* extraídos das partições como conjunto de soluções. O intuito é avaliar a capacidade dos métodos em encontrar os *clusters* relevantes nos dados e verificar se nesse contexto há um melhor desempenho das técnicas.

4.2.1 Capacidade de Recuperação dos Clusters Reais

A avaliação do desempenho foi realizada utilizando o índice InD, que permite a análise da similaridade entre os *clusters* das partições reais e os *cluster* presentes nas partições geradas pelos métodos. Em virtude do volume de *clusters* reais, não serão fornecidos os valores de InD, mas sim, o número e o percentual de *clusters* reais descobertos.

A Tabela 8 apresenta o número total de *cluster* reais (C_{TP}) e os totais de *clusters* recuperados pelos métodos MOCK, MOCLE e ASA, para cada conjunto de dados. A contagem representa o número total de *clusters* descobertos cujo valor do InD seja maior que 0,7 e igual a 1. Nesta análise, não foram incluídos os *ensembles*, pois o número de *clusters* recuperados é muito pequeno quando comparado aos métodos multi-objetivo e de seleção de partições.

Analisando o desempenho dos métodos, nota-se que MOCK conseguiu encontrar integralmente 15% dos *clusters* presentes nas partições reais e o MOCLE 22,30%. Já o método ASA, nesta análise, apresentou o melhor desempenho com 28% dos *clusters* reais recuperados.

Tabela 8 – Clusters Recuperados MO e Seleção de Partições

Tipo	Dataset	C_{TP}	InD = 1			InD > 0,7		
			MOCK	MOCLE	ASA	MOCK	MOCLE	ASA
Artificial	atom	2	1	2	2	2	2	2
	ds2c2sc13	20	14	15	19	14	16	19
	ds3c3sc6	9	0	0	0	5	5	5
	ds4c2sc8	10	2	1	1	3	8	8
	engyTime	2	0	0	0	1	2	2
	gaussian3	3	1	2	3	2	3	3
	hepta	7	1	4	4	2	5	7
	lsun	3	3	3	3	3	3	3
	monkey	18	3	2	7	5	8	11
	simulated6	6	2	5	5	4	6	6
	spiralsquare	8	1	2	4	4	6	8
	target	6	1	5	6	5	6	6
	tetra	4	2	4	4	2	4	4
	twoDiamonds	2	0	2	2	2	2	2
	wingNut	2	1	2	2	2	2	2
Real	armstrong	5	0	0	0	3	2	3
	chowdary	2	0	0	0	1	0	0
	contractions	2	0	0	0	1	1	1
	dyrskjot	3	0	0	0	1	2	2
	eTongueSugar	5	0	0	0	2	2	2
	glass	13	0	0	0	3	3	3
	golub	11	2	2	1	3	7	7
	gordon	2	1	0	0	1	1	1
	iris	3	1	1	1	3	3	3
	laryngeal1	2	0	0	0	1	0	0
	laryngeal2	2	0	0	0	1	1	1
	laryngeal3	5	0	0	0	0	1	1
	libras	23	0	0	0	1	0	1
	lung	4	1	1	1	3	4	4
	miRNAcancer	40	4	6	10	8	14	22
	respiratory	2	0	0	0	0	0	0
	segmentation	7	0	1	1	2	1	1
	su	10	1	0	0	2	4	4
	voice9	11	1	0	0	1	1	1
	voice3	5	0	0	0	2	0	2
	weaning	2	0	0	0	0	0	0
	yeoh	8	0	0	0	1	2	2
	Total	269	43	60	76	96	127	149

A Figura 5 resume o desempenho somente dos métodos de *ensemble* em relação ao percentual de *clusters* reais recuperados ($InD = 1$) ou recuperados parcialmente ($InD > 0{,}7$), os valores são muito inferiores aos valores obtidos através dos métodos MOCK, MOCLE e ASA.

Os resultados da análise de *clusters* confirmam o baixo desempenho dos métodos de *ensemble*, assim como já foi constatado na análise das partições através do índice ARI. Ou seja, mesmo observando-se detalhadamente o conteúdo das partições, observa-se que os *ensembles* perdem muito das informações presentes nas partições base.

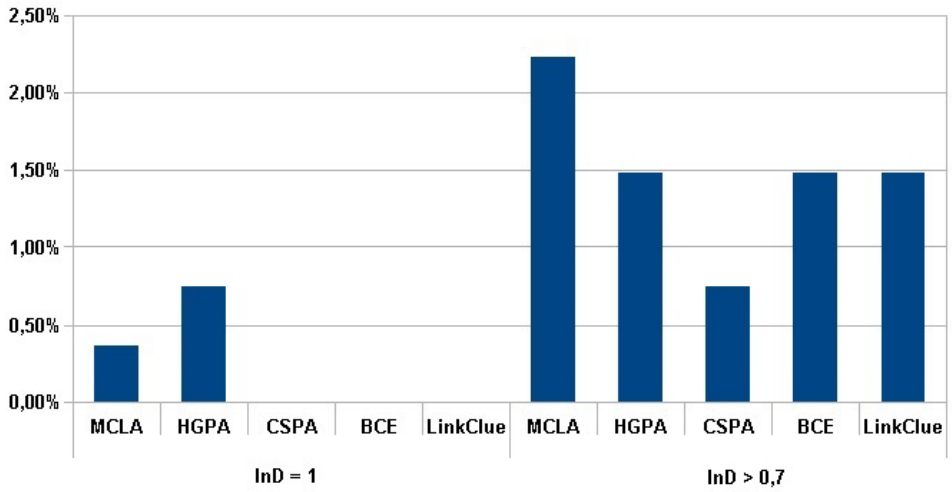

Figura 5 – Percentual de Clusters Recuperados - Ensemble

4.2.2 Tamanho dos conjuntos de soluções

Na Tabela 9, pode-se observar o volume de *clusters* extraídos das partições obtidas através dos métodos que geram conjuntos de soluções. A tabela apresenta os conjuntos de dados (*Dataset*), o número de *clusters* reais C_{TP}, o número de *clusters* extraídos das partições iniciais C_I, e o número de *clusters* obtidos através dos métodos MOCK, MOCLE e ASA.

Ao analisar os *clusters* individualmente, observa-se a existência de um grande volume de partições que apresentam informação redundante. Um exemplo é o conjunto de dados `miRNAcancer`, que possui 40 *clusters* reais, ao extrair os *clusters* c_S obtidos pelo método ASA, nota-se que foram gerados 18 vezes mais *clusters* que o número de *clusters* reais C_{TP} em seu conjunto de solução C_S, ou seja, o total de 744 *clusters*. Deste total, pôde-se constatar que 329 *clusters* são redudantes, ou seja, se repentem em uma ou mais partições. Porém, mesmo identificando os *clusters* distintos, ainda assim é um número muito grande de informação que não apresenta relevância. Este fato, dificultaria este tipo de análise para um pesquisador do domínio dos dados.

Tabela 9 – Clusters extraídos do Conjunto de Soluções

Tipo	Dataset	C_{TP}	C_I	MOCK	MOCLE	ASA
Artificial	atom	2	65	14	38	34
	ds2c2sc13	20	1923	214	663	267
	ds3c3sc6	9	457	135	208	160
	ds4c2sc8	10	800	165	286	216
	engyTime	2	57	48	4554	49
	gaussian	3	189	24	18	78
	hepta	7	575	42	113	136
	lsun	3	115	27	37	34
	monkey	18	675	227	267	156
	simulated6	6	416	48	63	90
	spiralsquare	8	471	71	148	131
	target	6	385	78	148	153
	tetra	4	266	28	108	64
	twoDiamonds	2	70	14	28	37
	wingNut	2	50	20	42	15
Real	armstrong	5	111	24	22	38
	chowdary	2	45	8	11	13
	contractions	2	51	16	19	49
	dyrskjot	3	100	30	30	73
	eTongueSugar	5	100	27	19	46
	glass	13	385	148	166	89
	golub	11	228	73	80	121
	gordon	2	45	12	34	25
	iris	3	132	33	49	30
	laryngeal1	2	57	16	26	32
	laryngeal2	2	45	26	45	33
	laryngeal3	5	100	43	61	62
	libras	23	2308	341	647	303
	lung	4	197	48	82	80
	miRNAcancer	40	4394	702	1056	744
	respiratory	2	60	10	9	27
	segmentation	7	420	77	442	131
	su	10	1045	190	419	249
	voice3	5	125	34	33	41
	voice9	11	943	177	333	319
	weaning	2	66	14	36	37
	yeoh	8	385	124	110	164
	Total	269	17856	3328	10450	4326

5 Conclusão

Nesta dissertação foi realizada uma análise exploratória das características das novas abordagens de agrupamento, foram realizados experimentos com métodos tradicionais de *ensemble*, agrupamento multi-objetivo, *ensemble* multi-objetivo e método de seleção de partições.

No Capítulo 3, foi realizado o estudo da influência das partições base nos *ensembles* tradicionais e *ensemble* multi-objetivo (FACELI et al., 2015). Esta análise foi motivada por trabalhos como (HANDL; KNOWLES, 2007) e (FACELI; CARVALHO; SOUTO, 2007) que já indicavam que os *ensembles* tradicionais levam a perda de informação. Nos estudos realizados, foram realizadas análises com mais profundidade, demonstrando que de fato os métodos tradicionais de *ensemble* demonstraram baixa performance em todos os experimentos realizados com os diversos conjuntos de partições base. Por outro lado, nos experimentos realizados com o método de *ensemble* multi-objetivo, foram constatadas soluções de alta qualidade, mesmo em situações em que continha apenas a informação parcial dos dados, os experimentos demonstraram que a condição inicial não influenciou negativamente o resultado final.

O outro estudo realizado, teve como objetivo avaliar a capacidade das técnicas em relação a recuperar as informações existentes nos dados. E para isto, foram realizadas investigações nos dois contextos: partições, que é a forma tradicional de análise e *clusters* para verificar internamente se as partições recuperadas contém mais informações relevantes do que a análise de partições demonstra. E para realizar tais análises foram observadas a qualidade das partições e dos *clusters* (ARI e InD), a porcentagem de informações reais (partições e *clusters*) realmente recuperadas, nos dois contextos, e o volume de informações irrelevantes que cada técnica produz.

Nos experimentos realizados, pôde-se concluir que os métodos de *ensemble* demonstraram resultados inferiores aos resultados obtidos pelos métodos de agrupamento multi-objetivo, *ensemble* multi-objetivo e método de seleção de partições.

Em relação a qualidade e a quantidade das partições recuperadas, nenhum dos métodos de *ensemble* conseguiu recuperar a partição real integralmente. Já os métodos MOCK, MOCLE e ASA, que geram um conjunto de soluções, conseguiram recuperar a partição real em alguns experimentos, sendo que o MOCLE apresentou melhor desempenho recuperando 16,13% das partições reais. Mesmo assim, a qualidade da informação extraída é bastante subestimada quando é avaliada pela análise partições.

Em relação a recuperação de *clusters*, de acordo com os resultados, o MOCK conseguiu encontrar integralmente 15% dos *clusters* presentes nas partições reais e o

MOCLE 22,30%. Já o método ASA, nesta análise, apresentou o melhor desempenho com 28% dos *clusters* reais recuperados. Os ensembles tiveram um resultado muito inferior aos métodos MOCK, MOCLE e ASA, em torno de 1,5% dos clusters recuperados parcialmente ($InD > 0{,}7$).

Observa-se que apesar dos resultados apresentados pelos métodos que geram conjuntos de soluções terem sido superiores aos *ensembles*, tais métodos geram um grande volume de informação redundante, ou seja, *clusters* que se repetem em várias partições, dificultando a análise do especialista do domínio dos dados.

As limitações deste trabalho estão relacionadas a investigação de outros algoritmos além dos que foram utilizados nas análises. Neste trabalho foram realizados experimentos somente com algoritmos de autores que disponibilizaram seus códigos fontes.

Possíveis trabalhos futuros, podem estar relacionados ao desenvolvimento de um *framework* que facilite o processo de quantificar e avaliar os *clusters* de uma forma mais simplificada, possibilitando que especialistas do domínio dos dados possam realizar as análises neste nível de refinamento.

Referências

ARMSTRONG, S. A. et al. MLL Translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, v. 30, n. 1, p. 41–47, 2002. Citado na página 41.

BERKHIN, P. A Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data*, Springer Berlin Heidelberg, p. 25–71, 2006. Citado na página 25.

BHATTACHARJEE, A. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma sub-classes. *Proc. Natl. Acad. Sci. USA*, v. 98, n. 24, p. 13790–13795, 2001. Citado na página 41.

CHOWDARY, D. et al. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *Journal of Molecular Diagnostics*, v. 8, n. 1, p. 31–39, 2006. Citado na página 41.

DEMSAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, 2006. Citado 3 vezes nas páginas 34, 35 e 48.

D'HAESELEER, P. How does gene expression clustering word. *Nature Biotechnology 23*, p. 1499 – 1501, 2005. Citado na página 21.

DUDA, R. O.; HART, P. E. Pattern Classification and Scene Analysis. Wiley, 1973. Citado na página 21.

DYRSKJØT, L. et al. Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics*, v. 33, n. 1, p. 90–96, 2003. Citado na página 41.

EVERITT, B. S. et al. Cluster analysis. Wiley, 2009. Citado na página 21.

FACELI, K.; CARVALHO, A. C. P. L. F. de; SOUTO, M. C. P. de. Multi-objective Clustering Ensemble. *Int. J. Hybrid Intell. Syst.*, v. 4, p. 145–156, 2007. Citado 2 vezes nas páginas 22 e 57.

FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. LTC, 2011. Citado 2 vezes nas páginas 21 e 27.

FACELI, K.; SAKATA, T. C. Multiple solutions in cluster analysis: partitions x clusters. *Machine Learning*, 2015. Citado na página 22.

FACELI, K. et al. Impact of Base Partitions on Multi-objective and Traditional Ensemble Clustering Algorithms. *International Conference on Neural Information Processing*, 2015. Citado 3 vezes nas páginas 22, 31 e 57.

FACELI, K. et al. Partitions selection strategy for set of clustering solutions. *Neurocomputing*, v. 73, n. 1618, p. 2809 – 2819, 2010. Citado na página 40.

FACELI, K. et al. Multi-objective clustering ensemble for gene expression data analysis. *Neurocomputing*, Elsevier Science Publishers B. V., v. 72, n. 13-15, p. 2763–2774, 2009. Citado 3 vezes nas páginas 34, 41 e 42.

FRED, A.; JAIN, A. Combining multiple clusterings using evidence accumulation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 27, n. 6, p. 835–850, 2005. Citado na página 27.

GOLUB, T. R. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, v. 286, n. 5439, p. 531–537, 1999. Citado na página 41.

GORDON, G. J. et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, v. 62, n. 17, p. 4963–4967, 2002. Citado na página 41.

GüNNEMANN, S.; FäRBER, I.; SEIDL, T. Multi-view clustering using mixture models in subspace projections. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 132–140, 2012. Citado na página 26.

HANDL, J.; KNOWLES, J. An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, v. 11, n. 1, p. 56–76, 2007. Citado 5 vezes nas páginas 21, 22, 40, 42 e 57.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, n. 1, p. 193–218, 1985. Citado na página 35.

IAMON, N.; GARRETT, S. Linkclue: A MATLAB package for link-based cluster ensembles. *Journal of Statistical Software*, v. 36, n. 1, 2010. Citado 2 vezes nas páginas 28 e 42.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. Citado na página 21.

JAIN, A. K.; FLYNN, P. J. Image segmentation using clustering. *In Advances in Image Understanding: A Festschrift for Azriel Rosenfeld*, IEEE Press, 1996. Citado na página 21.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. *ACM Comput. Surv.*, ACM, v. 31, p. 264–323, 1999. Citado 2 vezes nas páginas 25 e 26.

KARYPIS, G. et al. Multilevel hypergraph partitioning: applications in VLSI domain. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, v. 7, n. 1, p. 69–79, 1999. Citado na página 28.

KOHONEN, T. Self-organizing maps. *Springer Series in Information Sciences*, v. 30, 1995. Citado na página 25.

LU, J. et al. MicroRNA expression profiles classify human cancers. *Nature*, v. 435, p. 834–838, 2005. Citado na página 41.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *University of California Press*, University of California Press, 1966. Citado na página 25.

MONTI, S. et al. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, v. 52, n. 1-2, p. 91–118, 2003. Citado na página 40.

NEWMAN, D. et al. *UCI Repository of machine learning databases*. 1998. <http://www.ics.uci.edu/~mlearn/MLRepository.html> [Acessado em 08/2015]. University of California, Irvine, Dept. of Information and Computer Sciences. Citado na página 41.

SAKATA, T. et al. The Assessment of the Quality of Sugar Using Electronic Tongue and Machine Learning Algorithms. *11th International Conference on Machine Learning and Applications (ICMLA)*, v. 1, p. 538–541, 2012. Citado na página 41.

SAKATA, T. et al. Improvements in the partitions selection strategy for set of clustering solutions. *Proceedings of the 11th Brazilian Symposium on Neural Networks, (SBRN'2010)*, p. 49–54, 2010. Citado 2 vezes nas páginas 22 e 43.

SANTOS, D.; BAZZAN, A. A biologically-inspired distributed clustering algorithm. *Swarm Intelligence Symposium, 2009. SIS '09. IEEE*, p. 160–167, 2009. Citado na página 25.

SRIVASTAVA, J. et al. Web usage mining: Discovery and application of usage patterns from web data. *ACM SIGKDD Explorations*, p. 12–23, 2000. Citado na página 21.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. A comparison of document clustering techniques. *In KDD Workshop on Text Mining*, 2000. Citado na página 21.

STREHL, A.; GHOSH, J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, v. 3, p. 583–617, 2002. Citado 4 vezes nas páginas 27, 28, 34 e 42.

SU, A. I. et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research*, v. 61, n. 20, p. 7388–7393, 2001. Citado na página 41.

TOPCHY, A.; JAIN, A. K.; PUNCH, W. Combining multiple weak clusterings. *Proceedings of the Third IEEE International Conference on Data Mining*, IEEE Computer Society, p. 331–, 2003. Citado 2 vezes nas páginas 27 e 28.

ULTSCH, A. Clustering with som: U*c. *Workshop on Self-Organizing Maps*, Paris, France, p. 75–82, 2005. Citado na página 39.

WANG, H.; SHAN, H.; BANERJEE, A. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, v. 4, n. 1, p. 54–70, 2011. Citado 4 vezes nas páginas 28, 29, 34 e 42.

XU, R.; WUNSCH D., I. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, v. 16, n. 3, p. 645–678, 2005. Citado na página 21.

YEOH, E.-J. et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, v. 1, n. 2, p. 133–143, 2002. Citado na página 41.