
Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

# Rotulação Automática de *Clusters* Baseados em Análise de Filogenias

**Francisco Neto Carvalho de Araújo**

**Número de Ordem PPGCC: M001**

**Teresina-PI, Março de 2018**

Francisco Neto Carvalho de Araújo

# Rotulação Automática de *Clusters* Baseados em Análise de Filogenias

**Dissertação de Mestrado** apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Vinicius Ponte Machado

Coorientador: Antonio Helson Mineiro Soares

Teresina-PI

Março de 2018

Francisco Neto Carvalho de Araújo
Rotulação Automática de *Clusters* Baseados em Análise de Filogenias/ Francisco Neto Carvalho de Araújo. – Teresina-PI, Março de 2018-
48 p. : il. (algumas color.) ; 30 cm.

Orientador: Vinicius Ponte Machado

Dissertação (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, Março de 2018.

1. Aprendizagem de Máquina. 2.Agrupamento. I. Dr. Vinicius Ponte Machado. II. Universidade Federal do Piauí. III. Departamento de Computação. IV. Rotulação Automática de Grupos Baseados em Análise de Filogenias

CDU 02:141:005.7

Francisco Neto Carvalho de Araújo

# Rotulação Automática de *Clusters* Baseados em Análise de Filogenias

**Dissertação de Mestrado** apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 19 de Março de 2018:

**Vinicius Ponte Machado**
Orientador

**Antonio Helson Mineiro Soares**
Co-Orientador

**Alexandre Cláudio Botazzo Delbem**

**Laurindo de Sousa Britto Neto**

**Ricardo de Andrade Lira Rabêlo**

Teresina-PI
19 de Março de 2018

*Aos meus pais Marcos e Edineide, e à minha esposa Jullianna,*
*por me apoiarem em todos os momentos.*

# Agradecimentos

Agradeço aos meus pais, Marcos e Edineide, por todo o amor, carinho, confiança, apoio, atenção e compreensão ao longo de toda a minha vida, em especial durante a minha longa jornada acadêmica.

À minha esposa Jullianna, por todo o companheirismo, carinho, amor, por sempre estar ao meu lado apoiando em todas as decisões e sendo minha principal incentivadora e exemplo de dedicação.

Agradeço ao meu orientador, Vinicius Ponte Machado, por todos os conselhos, pela paciência, pelo incentivo, pela aceitação como orientando e ajuda nesse período pelo qual passamos por três projetos de iniciação científica até o momento atual do programa de mestrado.

Agradeço ao meu co-orientador, Antonio Helson Mineiro Soares, pelo incentivo, apoio, confiança e por todo o suporte fundamental para tornar o desenvolvimento deste trabalho possível.

Aos meus amigos Francisco Júnior, Hugo Santos, Luís Guilherme, Wender, além dos demais colegas de mestrado que compartilharam e compartilham das vivências ocorridas até aqui. Agradeço também a todos os companheiros do LINA pela parceria e conhecimento compartilhado.

Agradeço a todos os professores do PPgCC, em especial aos professores Rodrigo de Melo Sousa Veras e Ricardo de Andrade Lira Rabêlo, cujos papéis foram fundamentais no transcorrer do desenvolvimento da pesquisa, pelos conselhos, pela colaboração e pelo incentivo para sempre seguir adiante.

Aos professores Antonio Helson Mineiro Soares, Laurindo de Sousa Britto Neto e Ricardo de Andrade Lira Rabêlo, Rodrigo de Melo Sousa Veras e Alexandre C. B. Delbem por aceitarem fazer parte da banca de avaliação desta qualificação.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro para realização deste trabalho de pesquisa.

*"O maior inimigo do sucesso*
*é o medo do fracasso."*
*(Autor Desconhecido)*

# Resumo

O agrupamento (clusterização) é uma das principais técnicas de reconhecimento de padrões. Essa técnica consiste em identificar grupos (*clusters*) de elementos em um determinado conjunto de dados, levando em consideração métricas que permitam determinar a semelhança entre eles. Os elementos presentes nesses conjuntos de dados (*data sets*) frequentemente são descritos por meio de atributos, os quais podem assumir valores de diversos tipos, exigindo métodos eficientes na tarefa de detectar correlações entre dados de tipos complexos (ou mistos). No entanto, o processo de clusterização não fornece informações claras que permitam inferir as características de cada *cluster* formado, ou seja, o resultado do processo de clusterização não permite que os *clusters* tenham seu significado facilmente compreendido. A rotulação de dados visa identificar essas características e permitir então que se tenha a plena compreensão dos *clusters* resultantes. Neste trabalho propõe-se a utilização em conjunto de métodos de Aprendizagem de Máquina não supervisionada e supervisionada para as tarefas de agrupamento e rotulação de dados, respectivamente. Os algoritmos DAMICORE e sua nova versão, o DAMICORE-2 (ambos reconhecidamente eficientes) foram utilizados para detectar *clusters* que posteriormente foram submetidos ao Método de Rotulação Automática de *clusters* (MRA), obtendo taxas de acerto média, entre todos os conjuntos de dados, de 86,75%.

**Palavras-chaves**: Aprendizagem de Máquina. Agrupamento. Rotulação Automática.

# Abstract

Clustering is one of the main techniques of pattern recognition. This technique consists of organizing the elements of a given set into groups (clusters) taking into account some metric that allows to determine the similarity in them. These datasets often describe the elements that compose them by means of attributes that can take values of several types, requiring efficient methods in the task of detectiong correlations between complex (or mixed) type data. However, the clustering process does not provide clear information to infer the characteristics of each clusters formed, ie, the result of the clustering process does not allow clusters to have their meaning easily understood. Data labeling aims at identifying these characteristics and then allowing full understanding of the resulting clusters. In this work we propose the joint use of unsupervised and supervised Machine Learning methods for data grouping and labeling tasks, respectively. For that, we used well-known algorithms. The DAMICORE algorithms and their new version, DAMICORE-2, were used to form clusters that were later submitted to the Automatic Labeling Method (ALM), obtaining average hit rates that reached 86.75%.

**Keywords**: Machine Learning. Clustering. Pattern Recognition.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| AM | Aprendizagem de Máquina |
| DAMICORE | *DAta MIning of COde REpository* |
| EFD | *Equal Frequency Discretization* |
| EWD | *Equal Width Discretization* |
| FA | *Fast Newman Algorithm* |
| IA | Inteligência Artificial |
| MRA | Método de Rotulação Automática |
| NCD | *Normalized Compression Distance* |
| NID | *Normalized Information Distance* |
| NJ | *Neighbor Joining* |
| RNA | Rede Neural Artificial |

# Sumário

# Introdução

A rápida popularização do uso de computadores para informatizar diversos setores da sociedade resultou no expressivo crescimento das bases de dados. Pesquisadores passaram então a utilizar técnicas de reconhecimento de padrões, por meio da detecção de correlações entre os dados, que pudessem trazer à tona conhecimentos relevantes e úteis, potencialmente contidos nessas bases (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Uma das principais técnicas de reconhecimento de padrões é o agrupamento (clusterização) o qual tem sido considerado como uma das mais relevantes dentre as existentes na área de aprendizagem de máquina não supervisionada (um paradigma do Aprendizado de Máquina). O processo de clusterização visa identificar (detectar) grupos (*clusters*) de elementos em um conjunto de objetos, levando em consideração métricas que permitam medir a correlação existente entre os atributos que os descrevem.

Han, Kamber e Pei (2011) afirmam que o processo de clusterização tem sido largamente utilizado em áreas como inteligência empresarial, podendo ser utilizado para organizar uma grande quantidade de clientes em grupos, facilitando o desenvolvimento de estratégias de negócio que melhorem a gestão de relacionamento com os clientes; reconhecimento de padrões em imagens, para, por exemplo, aumentar a precisão de sistemas de reconhecimento de caracteres manuscritos; biologia; segurança; entre outras.

Embora tenha sido um dos focos principais dos pesquisadores, o processo de clusterização não fornece informações que permitam inferir as características de cada *cluster* formado. Isso se deve a limitações das métricas utilizadas (ANAYA-SÁNCHES; PONS-PORRATA; BERLANGA-LLAVORI, 2008), uma vez que, sozinhas, não deixam claras as características predominantes em cada *cluster*. A rotulação de dados visa identificar essas características e permitir então que se tenha a compreensão dos *clusters* resultantes.

A rotulação de um *cluster* busca resumir sua definição, ou seja, descrevê-lo em função de seus atributos mais relevantes – ou seja, aqueles que são determinantes para o agrupamento – e suas respectivas faixas de valores, a fim de melhor compreendê-lo. Assim, esse conjunto de valores representa uma definição para um *cluster* qualquer – isto é, um rótulo – capaz de fornecer ao especialista um melhor entendimento sobre os dados.

É comum a presença de uma diversidade de tipos de dados em uma mesma base, o que torna a inferência de uma correlação entre eles um processo geralmente não trivial e relativamente complexo. O método DAMICORE (do inglês, *DAta MIning of COde REpository*) (SANCHES; CARDOSO; DELBEM, 2011) mostrou ser capaz de encontrar correlações em bases de dados de tipos mistos (registros com diferentes tipos de dados), com base na análise de filogenias (estruturas que representam o relacionamento evolucionário

entre espécies).

Neste trabalho, utilizou-se também uma nova versão do DAMICORE, chamada DAMICORE-2 (descrito na Seção 1.3). Ambos os métodos têm como base a reconstrução de árvores filogenéticas, que são estruturas hierárquicas que representam os relacionamentos entre os elementos. O agrupamento resultante é, portanto, hierárquico. Isso permite que o Método de Rotulação Automática (MRA) (LOPES; MACHADO; RABÊLO, 2016) seja utilizado para rotular não somente os *clusters* obtidos, mas também os nós internos, os quais representam um conhecimento mais generalizado, ou seja, representam uma quantidade maior de elementos correlacionados em estruturas denominadas *super-clusters*, definidas na Seção 2.2.

A existência de um rótulo permite a identificação de quais características definem um *cluster*. Lopes, Machado e Rabêlo (2016) apontam que um rótulo pode ser útil para a identificação das características em problemas que necessitam de ações corretivas, podendo até mesmo apontar a intensidade dessas ações, baseada nos valores encontrados. Assim, a compreensão de *clusters* por meio de rótulos pode contribuir de várias maneiras com a elaboração da solução ou otimização de um problema.

Em Lopes, Machado e Rabêlo (2016) é proposta a utilização de Redes Neurais Artificiais (RNA) para identificar quais os atributos relevantes, e suas respectivas faixas de valores que, juntos, formam o rótulo de um determinado *cluster*. Ou seja, as RNAs determinam as características predominantes pelas quais os elementos foram alocados em um mesmo *cluster*. A abordagem proposta obteve taxa de acerto média de 74,13%, ao ser aplicada em *clusters* formados pelo algoritmo *K-means* (MACQUEEN, 1967).

O método de clusterização utilizado é um dos fatores de maior influência sobre a acurácia da rotulação. Assim, quanto maior for a semelhança intra-*cluster* (quanto mais semelhantes forem os elementos de um *cluster*, entre si), maior será a capacidade dos rótulos encontrados de definirem os *clusters*. Neste trabalho, apresenta-se a utilização do MRA para rotular os *clusters* formados pelo DAMICORE e pelo DAMICORE-2. Com isso, aferiu-se a eficiência do MRA em rotular *clusters* baseados em reconstrução de filogenias.

## Objetivo Geral

Apresentar uma metodologia para rotulação de *clusters* baseados em análise de filogenias, utilizando o MRA, permitindo identifica as principais características de cada um deles.

## Objetivos Específicos

- Apresentar um método para codificação de valores discretos;

- Rotular os nós folhas da filogenia resultante;
- Rotular os nós internos da filogenia resultante.

## Contribuições Científicas

No decorrer do desenvolvimento deste trabalho foram publicados e/ou aceitos artigos relacionados à abordagem proposta com os resultados preliminares obtidos. As publicações foram:

- Araújo, F.; Soares, A.; Machado, V.; Rabêlo, R. Rotulação automática de *clusters* baseados em análise de filogenias. In: *X Encontro Unificado de Computação - ENUCOMP*, 2017. p. 489-496. ISBN: 978-85-8320-201-1 331.
- Araújo, F.; Machado, V.; Soares, A.; Veras, R. Automatic Cluster Labeling Based on Phylogram Analysis. In: *International Joint Conference on Neural Networks - IJCNN*, 2018. (Aceito)

## Estrutura Organizacional

O restante deste trabalho está organizado da seguinte maneira: no Capítulo 1 tem-se o referencial teórico com descrição dos métodos utilizados para as tarefas de clusterização e de rotulação, e a apresentação de alguns trabalhos relacionados; no Capítulo 2 é descrita a metodologia de realização dos testes e as métricas de avaliação utilizadas; no Capítulo 3 são apresentados os resultados e é feita uma breve comparação com os resultados do trabalho original; por fim, no Capítulo 4 são apresentadas as conclusões obtidas, além de listar possíveis trabalhos futuros.

# 1 Referencial Teórico

A seguir são apresentados os conceitos de Aprendizado de Máquina, bem como de seus dois principais paradigmas: Aprendizado supervisionado e Aprendizado não supervisionado. Posteriormente, apresenta-se o funcionamento básico dos métodos DAMICORE, DAMICORE-2 e MRA, em que os dois primeiros são utilizados para agrupamento e o terceiro para rotulação automática de *clusters*.

## 1.1 Aprendizagem de Máquina

A linguagem e o aprendizado, juntamente com a criação artística, a tomada de decisão ética e a responsabilidade social são apontadas como habilidades essencialmente humanas e, ao longo dos anos, têm sido um dos maiores desafios para o progresso da Inteligência Artificial (IA). Uma das razões que tornam essas áreas de pesquisa tão complexas, embora de suma importância, é que englobam muitas outras habilidades inteligentes humanas. Assim, a reivindicação da criação de uma Inteligência Artificial deve abordar questões sobre Linguagem Natural, Raciocínio Automático e Aprendizagem de Máquina (AM).

A capacidade de aprender deve fazer parte de qualquer sistema que reivindique possuir inteligência em um sentido geral. Segundo Luger (2013) os agentes inteligentes artificiais devem ser capazes de se modificarem ao longo do curso de suas interações com o mundo, bem como pela experiência de seus próprios estados e processos internos. O aprendizado é um dos pontos de maior importância para aplicações de IA.

Feigenbaum e MacCorduck (1983) identificaram as limitações presentes na engenharia do conhecimento como o maior obstáculo para o uso em larga escala de sistemas inteligentes. Uma solução para esse problema seria os programas começarem com uma quantidade mínima de conhecimento e aprenderem a partir de exemplos ou de suas próprias explorações do domínio.

Simon (1981) define o aprendizado como qualquer mudança em um sistema que melhore o seu desempenho na segunda vez em que ele repetir a mesma tarefa ou outra tarefa sobre a mesma população. Na mesma linha, Mitchell (1997) define Aprendizagem de Máquina como o ramo da IA que estuda o desenvolvimento de sistemas capazes de aprender com a experiência. Isto é, um sistema que realiza determinada tarefa e obtém informações que o permitam realizá-la com melhor desempenho em execuções futuras. Assim, a área de AM lida com o estudo de métodos computacionais de forma a permitir que programas de computadores obtenham melhoria na realização de uma tarefa por meio

de suas experiências anteriores.

Luger (2013) aponta a existência de quatro paradigmas de AM: simbólico, conexionista, genético e estocástico. No aprendizado simbólico um conjunto de símbolos representa as entidades e relações de um domínio do problema. Os algoritmos de aprendizado simbólico tentam inferir generalizações novas, válidas e úteis que podem ser expressas usando esses símbolos.

A abordagem conexionista representa o conhecimento como padrões de atividade em redes de pequenas unidades de processamento individuais, inspiradas na arquitetura do cérebro de animais. Os algoritmos genéticos e evolucionários fazem uso de uma população de soluções candidatas para um problema, avaliando-as segundo sua habilidade de solução e as combinando para produzir soluções cada vez melhores. Por fim, a abordagem estocástica propõe que a experiência de situações em um domínio condiciona as expectativas do conhecedor em interpretar novos dados nesse domínio.

Quanto ao modo, a AM é tipicamente dividida em três categorias: supervisionado, não supervisionado e por reforço. Os dois primeiros são usados neste trabalho e serão detalhados a seguir. No aprendizado por reforço, um agente (inteligente) é localizado em um ambiente e recebe realimentação desse contexto (LUGER, 2013). Esse aprendizado requer que o agente crie uma política que o permita interpretar o resultado de suas ações sobre o ambiente.

### 1.1.1 Aprendizagem Supervisionada

A Aprendizagem Supervisionada possui a tarefa de inferir uma função a partir de dados previamente rotulados em um conjunto de treinamento. O conjunto de treinamento consiste em exemplos. Cada exemplo é formado por um par entrada-saída, em que a entrada é tipicamente um vetor, contendo os valores das variáveis que descrevem os exemplos e a saída, o rótulo previamente estabelecido. Aprendizagem Supervisionada implica necessariamente a existência de entrada e a indicação de uma saída que possa ser aprendida para ocorrer o processo de aprendizagem (BRAGA; CARVALHO; LUDEMIR, 2007). Isto é, a classe à qual cada amostra no conjunto de treinamento pertence é previamente conhecida.

A indução, que é o aprendizado de uma generalização a partir de um conjunto de exemplos, é uma das tarefas de aprendizado mais fundamentais. O aprendizado de conceito é um problema típico de aprendizado indutivo em que, fornecidos exemplos de um determinado conceito, tenta-se inferir uma definição que permitirá à máquina reconhecer corretamente futuras ocorrências daquele conceito. Um algoritmo de Aprendizagem Supervisionada analisa um conjunto de treinamento e infere (induz) uma função matemática, a qual pode ser utilizada para mapear novos exemplos. A aprendizagem supervisionada, portanto, usa padrões para identificar os valores do rótulo em dados

adicionais não rotulados.

Os métodos utilizados nas etapas de treinamento e teste podem variar e devem ser ajustados conforme a técnica aplicada. Nesse contexto, a etapa de treinamento consiste, basicamente, em utilizar parte dos dados para a elaboração de um modelo. Essa parte dos dados é conhecida como conjunto de treinamento. Os valores de entrada são atribuídos ao modelo e seu resultado é então comparado à resposta esperada. O modelo produzido é reajustado sempre que houver erro ou até que se atinja uma margem de erro aceitável. Quando essa condição for finalmente satisfeita – ou que se atinja uma quantidade máxima de iterações – segue-se para a etapa de testes na qual a parte restante dos dados (denominada conjunto de teste) será utilizada para medir a acurácia do modelo produzido. As amostras presentes no conjunto de treinamento e conjunto de teste são mutuamente exclusivas.

### 1.1.2 Aprendizagem não Supervisionada

Contrário ao que acontece na Aprendizagem Supervisionada, no AM não Supervisionado inexiste a ideia de classe desejada, requerendo que o próprio algoritmo avalie os conceitos. A própria ciência pode ser um bom exemplo para ilustrar esse tipo de aprendizado. Um cientista, sem intermédio de um supervisor, propõe hipóteses a partir de observações sobre um experimento realizado, avaliando-as posteriormente de acordo com critérios pré-estabelecidos. Conforme Russell e Norvig (2003), o Aprendizado não supervisionado envolve a aprendizagem de padrões na entrada, quando não são fornecidos valores de saídas específicas.

Uma das principais tarefas de AM a fazer uso da aprendizagem não supervisionada é o agrupamento, também denominado de formação de categorias, ou seja, de discernir múltiplas categorias em uma coleção de objetos. O problema é não supervisionado pois, os rótulos das categorias não são conhecidos. O problema de agrupamento se inicia com uma coleção de objetos (conjunto de dados) não classificados e um meio de medir a sua similaridade. O objetivo é organizar os objetos de maneira a satisfazer algum padrão de qualidade, com a maximização da similaridade de objetos de um mesmo grupo e da dissimilaridade entre objetos de grupos distintos.

## 1.2 DAMICORE

Uma filogenia é uma representação, em forma de árvore, do relacionamento de espécies com a mesma origem. O termo Árvore Filogenética tem sido usado tanto para filogenias obtidas de dados morfológicos quanto para as obtidas de sequências genéticas. Neste trabalho, filogenias são reconstruídas com o objetivo de determinar *clusters* de objetos (filos) correlacionando esses dados.

Usualmente, Árvores Filogenéticas (um grafo acíclico conectado) são árvores binárias onde suas folhas representam espécies (SOARES; RABÊLO; DELBEM, 2017). Assim, folhas são identificadas com o nome da espécie correspondente. A Figura 1 mostra uma mesma filogenia, que destaca os relacionamentos evolucionários entre um conjunto de plantas, com possíveis clados (grupos de espécies evolucionariamente relacionadas) circundados por linhas tracejadas. As folhas representam espécies existentes, enquanto os nós internos indicam ancestrais hipotéticos ou espécies extintas.

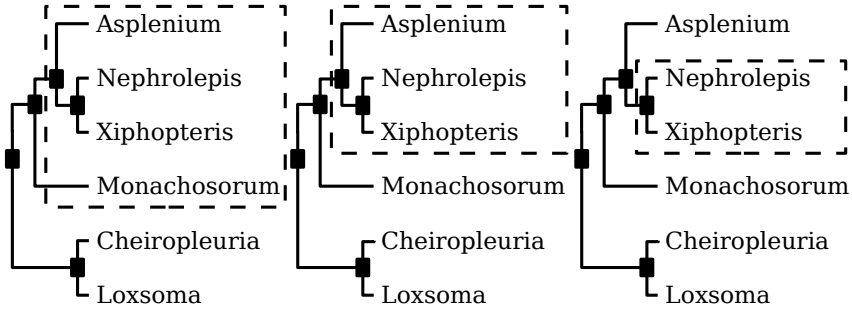



Figura 1 – Possíveis clados (linhas tracejadas) que podem ser obtidos de uma mesma filogenia.

Fazendo uso de filogenias, o DAMICORE é um método de detecção de correlação de dados que une algoritmos largamente utilizados, produzindo resultados eficientes, como demonstrado em Sanches, Cardoso e Delbem (2011). O método utiliza um conjunto de técnicas de várias áreas do conhecimento (Teoria da Computação, Bioinformática e Física) de forma a extrair informações por meio de uma métrica universal e robusta. O DAMICORE surge como um método de identificação de correlação entre dados de tipos diversos, procedimento relativamente complexo para a maioria dos algoritmos de clusterização. Além disso, uma de suas principais características é a inexistência da necessidade de informar ao algoritmo a quantidade de *clusters* na qual os elementos devem ser alocados.

Um diagrama resumindo todas as etapas do DAMICORE pode ser visto na Figura 2. O DAMICORE recebe como parâmetros o arquivo contendo os elementos a serem agrupados, o tamanho do problema (número de elementos) e o número de atributos que descrevem os elementos. A execução do DAMICORE é iniciada pelo cálculo da Matriz de Distância usando como métrica a NCD (do inglês, *Normalized Compression Distance*) (CILIBRASI; VITÁNYI, 2005), a qual calcula uma razão de distância entre os dados determinando a semelhança entre os valores das variáveis (atributos) com base nos tamanhos de seus dados compactados. A NCD tem sido aplicada com sucesso em áreas como a genética, literatura, música e astronomia. Além disso, essa abordagem não requer nenhum conhecimento específico do domínio da aplicação.

A NCD é baseada em outra métrica chamada NID (do inglês, *Normalized Information Distance*) (LILLO-CASTELLANO et al., 2013), que considera a semelhança entre as

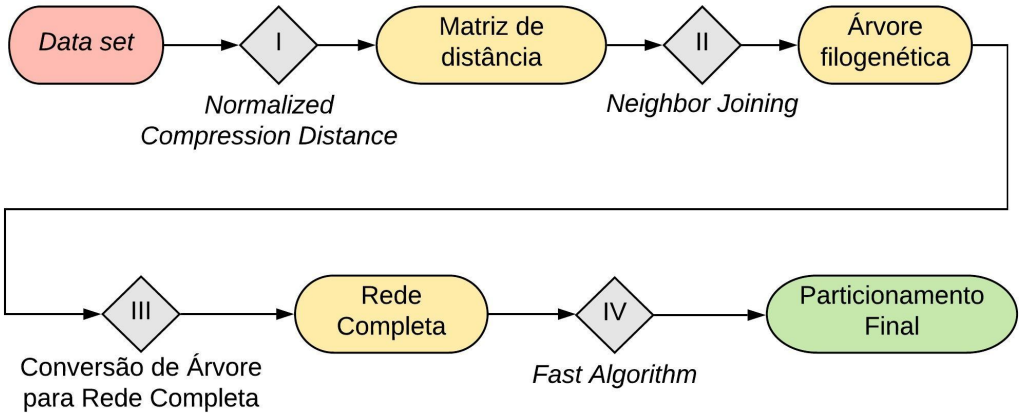



Figura 2 – Diagrama resumindo o funcionamento do DAMICORE.

variáveis de acordo com a característica dominante que elas compartilham. No entanto, a NID utiliza diretamente o conceito de complexidade de Kolmogorov (LI; VITÁNYI, 1997) no cálculo da distância, que é computacionalmente inviável para amostras grandes. A NCD substitui o cálculo da complexidade de Kolmogorov por uma aproximação obtida a partir de um algoritmo de compressão. Na prática, a distância entre dois dados $X$ e $Y$ em NCD é um número positivo variando entre $[0; 1 + \varepsilon]$, que representa o quão diferente $X$ e $Y$ são, e o parâmetro $\varepsilon$ é um limitante superior para o erro do compressor usado. O valor da distância entre $X$ e $Y$ é dado pela Equação 1.1:

$$D_{NCD}(X,Y) = \frac{C(XY) - min\{C(X), C(Y)\}}{max\{C(X), C(Y)\}}, \tag{1.1}$$

em que $C(XY)$ é o tamanho obtido após concatenação de $X$ e $Y$ seguida de sua compressão, $C(X)$ e $C(Y)$ são os tamanhos de $X$ e $Y$ comprimidas, respectivamente.

Considerando $C(Y) \geq C(X)$, a Equação 1.1 pode ser reescrita na forma da Equação 1.2:

$$D_{NCD}(X,Y) = \frac{C(XY) - C(X)}{C(Y)}, \tag{1.2}$$

isto é, a distância $D_{NCD}(X,Y)$ entre $X$ e $Y$ pode ser interpretada como o incremento resultante da compressão de $Y$ usando informações prévias sobre a compressão de $X$, expressando a diferença de tamanho entre as duas versões comprimidas. A Figura 3 exemplifica o cálculo da Matriz de Distância para um conjunto de amostras correspondente a parte de registros de ocorrências de acidentes de trânsito na rodovia BR-116 (município de Cajati) Km 509 ao Km 519, fornecidos pela concessionária AutoPista Régis Bittencourt.

A partir da Matriz de Distância, calculada pela NCD, é reconstruída uma Árvore Filogenética (ou Filogenia) (CANCINO; DELBEM, 2007) (que representa relações hie-

Base de Dados

| Número da Ocorrência | Objeto1- Tipo de Acidente | Objeto2- Causa Provável | Objeto3- Km |
|---|---|---|---|
| 0 | Saida de Pista | Mal subito do motorista | 511 |
| 1 | Saida de Pista | Derrapagem | 515 |
| 2 | Saida de Pista | Imprudência Condutor | 513 |
| 3 | Tombamento | Levou Fechada | 517 |
| 4 | Tombamento | Levou Fechada | 516 |
| 5 | Tombamento | Levou Fechada | 515 |
| 6 | Capotamento | Imprudência Condutor | 515 |
| 7 | Capotamento | Imprudência Condutor | 515 |
| 8 | Colisão Lateral | Imprudência Condutor | 514 |

(a)

NCD

Matriz Distância

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,000 | 1,754 | 1,111 | 1,899 | 1,907 | 1,914 | 1,829 | 1,872 | 1,889 |
| 1 | 1,754 | 0,000 | 1,683 | 1,966 | 1,967 | 1,984 | 1,622 | 1,933 | 1,948 |
| 2 | 1,111 | 1,683 | 0,000 | 1,888 | 1,957 | 1,943 | 1,750 | 1,864 | 1,923 |
| 3 | 1,899 | 1,966 | 1,888 | 0,000 | 1,789 | 1,895 | 1,988 | 1,803 | 1,949 |
| 4 | 1,907 | 1,967 | 1,957 | 1,789 | 0,000 | 1,984 | 1,918 | 1,865 | 1,939 |
| 5 | 1,914 | 1,984 | 1,943 | 1,895 | 1,984 | 0,000 | 1,914 | 1,930 | 1,955 |
| 6 | 1,829 | 1,622 | 1,750 | 1,988 | 1,918 | 1,914 | 0,000 | 1,670 | 1,672 |
| 7 | 1,872 | 1,933 | 1,864 | 1,803 | 1,865 | 1,930 | 1,670 | 0,000 | 1,857 |
| 8 | 1,889 | 1,948 | 1,923 | 1,949 | 1,939 | 1,955 | 1,672 | 1,857 | 0,000 |

(b)

Figura 3 – Matriz Distância calculada por meio da NCD.

rárquicas entre os indivíduos, uma vez que a estrutura de uma árvore é intrinsecamente hierárquica) usando o algoritmo NJ (do inglês, *Neighbor Joining*) (SAITOU; NEI, 1987). A Figura 4 mostra uma árvore reconstruída a partir da Matriz de Distância da Figura 3. A saída do NJ é, por sua vez, convertida do formato *Newick*[1] para o formato de Matriz de Adjacências. A Figura 5 ilustra essa conversão. O uso do formato *Newick* e sua conversão para Matriz de Adjacências possibilita também que diversos algoritmos de reconstrução de árvores possam ser considerados nessa etapa do DAMICORE.

Matriz Distância

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,000 | 1,754 | 1,111 | 1,899 | 1,907 | 1,914 | 1,829 | 1,872 | 1,889 |
| 1 | 1,754 | 0,000 | 1,683 | 1,966 | 1,967 | 1,984 | 1,622 | 1,933 | 1,948 |
| 2 | 1,111 | 1,683 | 0,000 | 1,888 | 1,957 | 1,943 | 1,750 | 1,864 | 1,923 |
| 3 | 1,899 | 1,966 | 1,888 | 0,000 | 1,789 | 1,895 | 1,988 | 1,803 | 1,949 |
| 4 | 1,907 | 1,967 | 1,957 | 1,789 | 0,000 | 1,984 | 1,918 | 1,865 | 1,939 |
| 5 | 1,914 | 1,984 | 1,943 | 1,895 | 1,984 | 0,000 | 1,914 | 1,930 | 1,955 |
| 6 | 1,829 | 1,622 | 1,750 | 1,988 | 1,918 | 1,914 | 0,000 | 1,670 | 1,672 |
| 7 | 1,872 | 1,933 | 1,864 | 1,803 | 1,865 | 1,930 | 1,670 | 0,000 | 1,857 |
| 8 | 1,889 | 1,948 | 1,923 | 1,949 | 1,939 | 1,955 | 1,672 | 1,857 | 0,000 |

NJ

Árvore

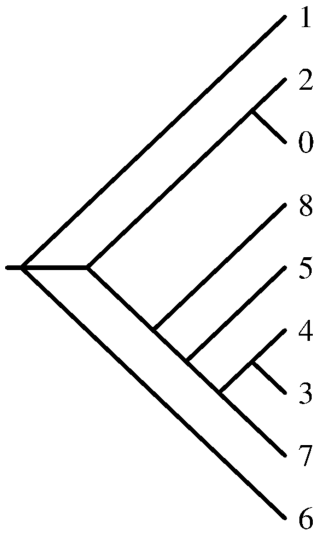



Figura 4 – Árvore Filogenética reconstruída pelo NJ.

Sobre a Matriz de Adjacências obtida é aplicado o FA (do inglês, *Fast Newman Algorithm*) (NEWMAN; GIRVAN, 2004), um algoritmo de detecção de estruturas de comunidades da área de Redes Complexas (DUCH; ARENAS, 2005). O FA realiza o Particionamento Final das variáveis do problema, contemplando todos os nós da Árvore Filogenética. Por fim, são removidos os nós internos, restando apenas os nós folhas,

[1] usualmente empregado por ferramentas de bioinformática.

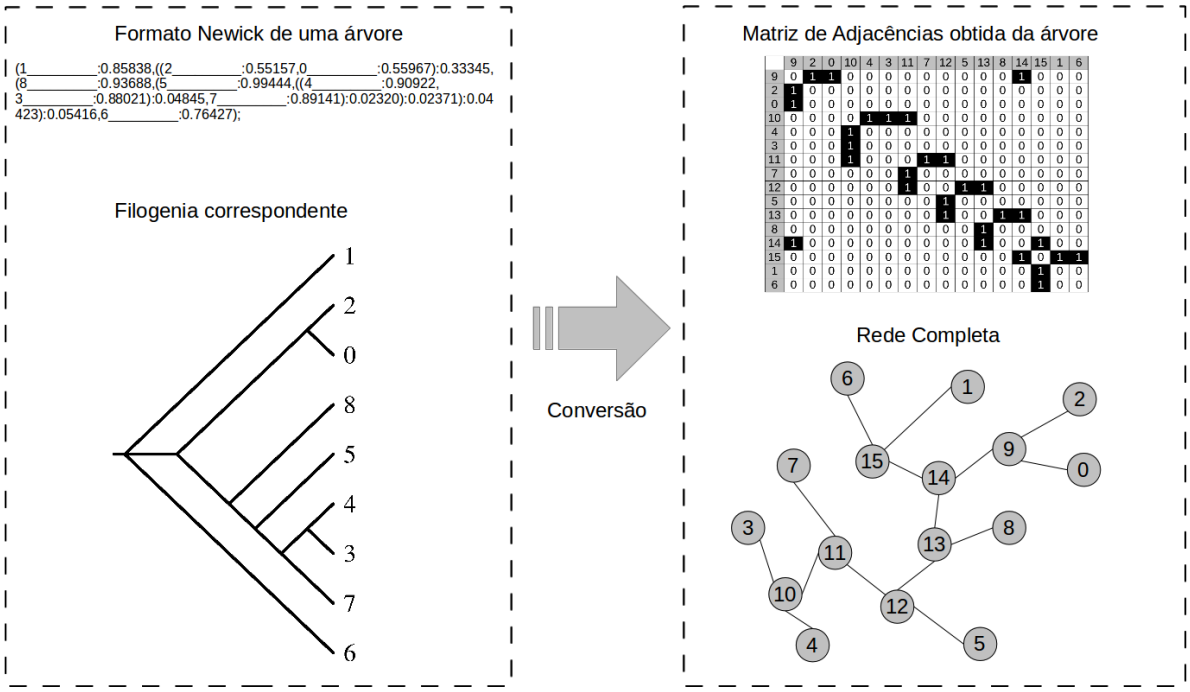



Figura 5 – Conversão de filogenia no formato *Newick* para Matriz de Adjacências. Neste exemplo, nós com índices maiores que 8 são nós internos da filogenia.

que representam as variáveis do problema, como ilustra a Figura 6. Como resultado do Particionamento são obtidas as comunidades detectadas, as quais representam os *clusters* formados.

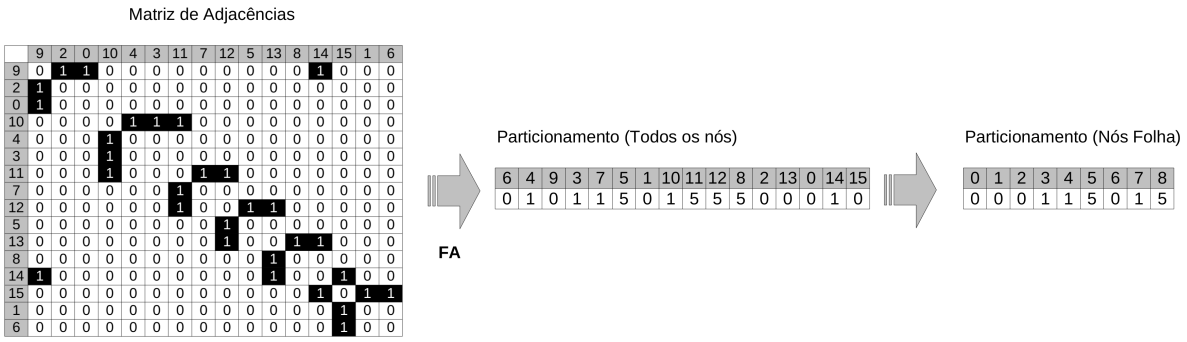



Figura 6 – Particionamento Final gerado a partir da Matriz de Adjacências (Figura 5) usando o FA.

## 1.3 DAMICORE-2

O DAMICORE-2 é uma nova versão do DAMICORE. A Figura 7 apresenta o diagrama geral do DAMICORE-2 contendo cada uma de suas etapas, descritas a seguir. Além dos mesmos parâmetros requeridos pelo DAMICORE, o DAMICORE-2 requer,

também, o número de Árvores Filogenéticas que deverão ser reconstruídas e o número de Redes ao qual cada uma dará origem.

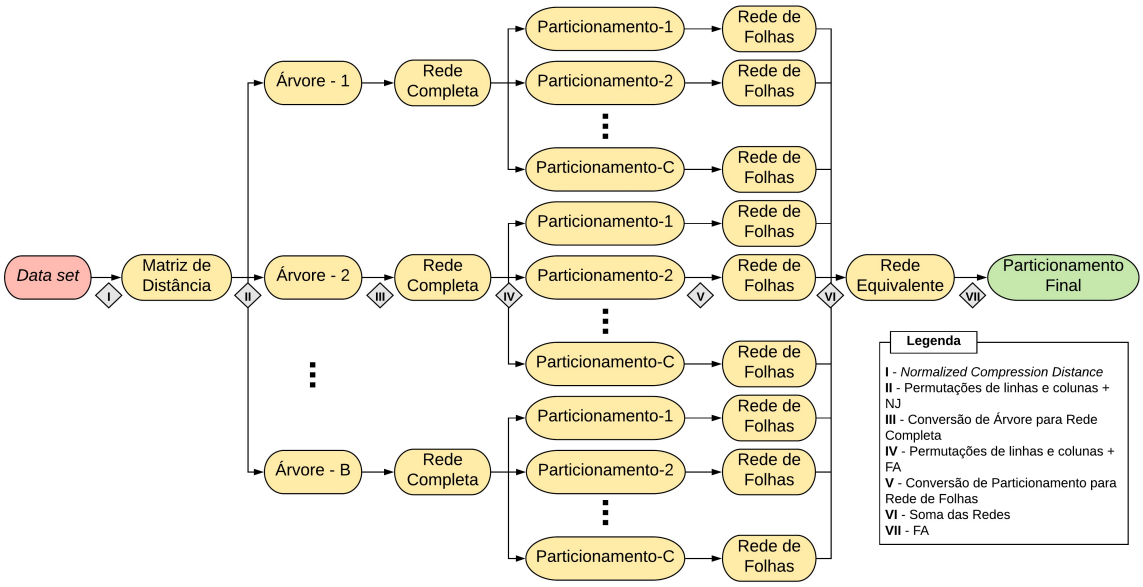



Figura 7 – Diagrama resumindo o DAMICORE-2.

Assim como sua versão anterior, o DAMICORE-2 inicia a sua execução calculando a Matriz de Distâncias por meio da NCD. A partir desse ponto se iniciam as diferenças entre as duas versões. Como o NJ é um algoritmo do tipo guloso, a filogenia reconstruída pode não ser a ótima. No entanto, geralmente é uma aproximação adequada da filogenia ótima. Assim, no DAMICORE-2, as colunas e linhas da Matriz de Distâncias são permutadas fazendo com que o NJ reconstrua diferentes árvores sub-ótimas para cada permutação. Um aspecto importante dessas diferentes árvores é a presença de subestruturas comuns. A Figura 8 mostra três árvores diferentes originadas da mesma Matriz de Distâncias por meio de permutações de colunas e linhas, destacando as subestruturas que se preservaram nas três.

Cada uma das árvores reconstruídas é convertida do formato *Newick* para o formato de Matriz de Adjacências, exatamente como ocorre no DAMICORE. Um conjunto de possíveis particionamentos das variáveis do problema é, então, determinado para cada Matriz, por meio do algoritmo FA. O FA, assim como o NJ, é um algoritmo guloso, produzindo particionamentos diferentes a cada execução. Para lidar com esse aspecto do FA, várias permutações são geradas produzindo um conjunto de particionamentos, a partir dos quais se pode eliminar as partições menos frequentes. Ao fim desse processo os nós internos são removidos do particionamento, mantendo-se apenas os nós folhas. A Figura 9 mostra particionamentos derivados de três permutações da Matriz de Adjacências.

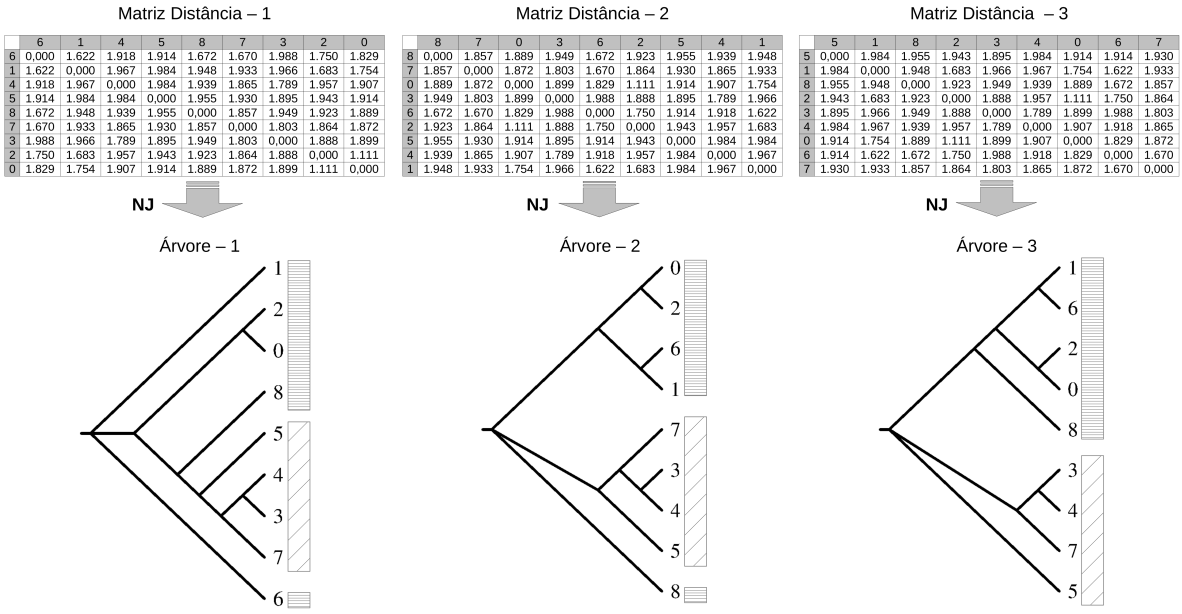



Figura 8 – Diferentes árvores geradas pelo NJ permutando aleatoriamente linhas e colunas de uma mesma Matriz Distância. As faixas iguais de uma filogenia para a outra destacam subestruturas comuns.

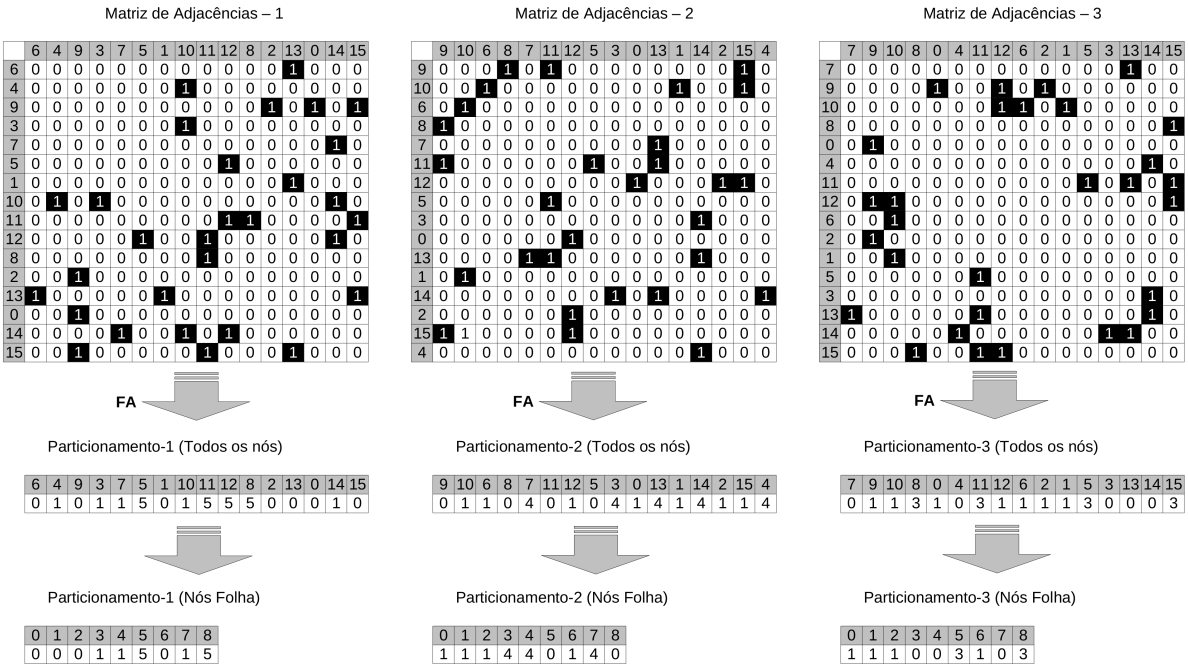



Figura 9 – Diferentes particionamentos gerados a partir das três permutações (Figura 8) sobre uma mesma Matriz de Adjacências (Figura 3) usando o FA.

Em seguida, os particionamentos contendo apenas os nós folhas, são convertidos para o formato de Matriz de Adjacências, agora representando um grafo chamado de Rede de Folhas. Nesse ponto, considera-se que todos os nós que constituem uma mesma partição formam um clique de grafo (DIESTEL, 2006), como exemplifica a Figura 10. Essa conversão se faz necessária para o passo seguinte, em que todas as Matrizes de Adjacências

são somadas (por meio de soma lógica ou operador booleano OU (LIPSCHUTZ; LIPSON, 2004)). O resultado é também uma Matriz de Adjacências denominada Rede Equivalente. A Figura 11 mostra a Rede Equivalente obtida a partir da soma das Redes de Folhas.

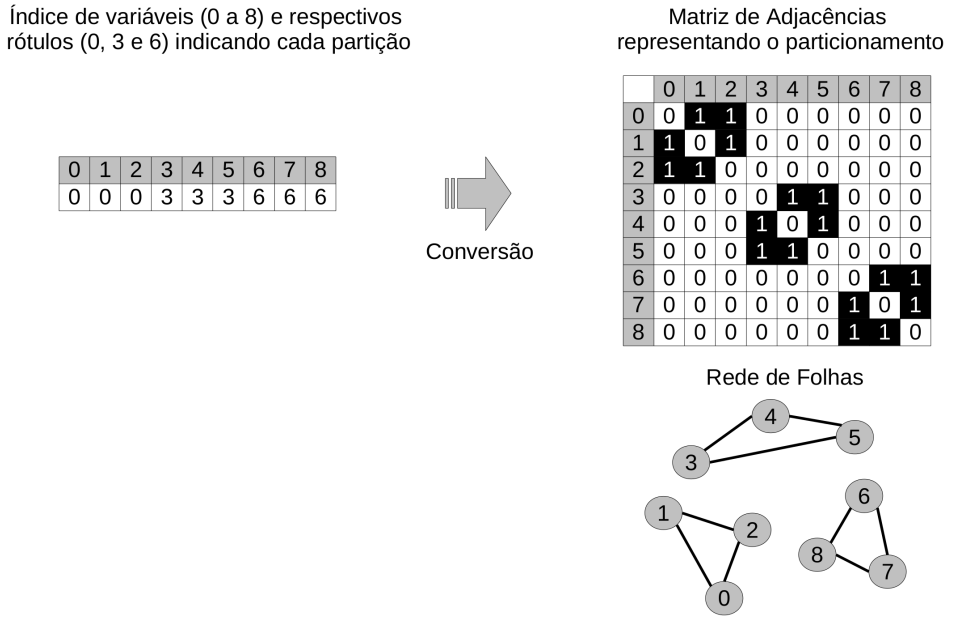



Figura 10 – Exemplo de conversão entre partições (vetor de rótulos), representação em Matriz de Adjacências e a Rede de Folhas correspondente.

Por fim, o FA é aplicado sobre a Rede Equivalente obtendo o Particionamento Final das variáveis do problema, resultando na identificação das comunidades existentes, como exemplifica a Figura 12. A metodologia empregada pelo DAMICORE-2 permite detectar correlações mesmo com a presença de Redes de baixa qualidade (por exemplo, com a ligação errônea de duas partições). A soma dessas Matrizes possibilita entender uma das propriedades do DAMICORE-2, a de que partições menores podem se tornar maiores (ou formar clados) ao se verificar consensos entre essas partições; ou simplesmente, podem se tornar mais representativas ou confiáveis por meio da inclusão de arestas entre seus nós, ao mesmo tempo em que reduz a interferência de ligações incorretas.

A inovação do DAMICORE-2 em utilizar um conjunto de filogenias, para compor um particionamento dos grupos das variáveis, permite que o método tenda a fugir da aleatoriedade de resultados subótimos do NJ e FA, buscando um consenso desses resultados, por meio dos quais se pode abstrair um resultado mais consistente e menos flutuante a cada execução e com isso, descobrir novidades nas bases de dados.

## 1.4 Rotulação Automática de *Clusters*

Existem diversas pesquisas acerca do problema de clusterização, entretanto, poucas são as que focam em rotular os *clusters* resultantes. Tzerpos (2001) aponta que, em

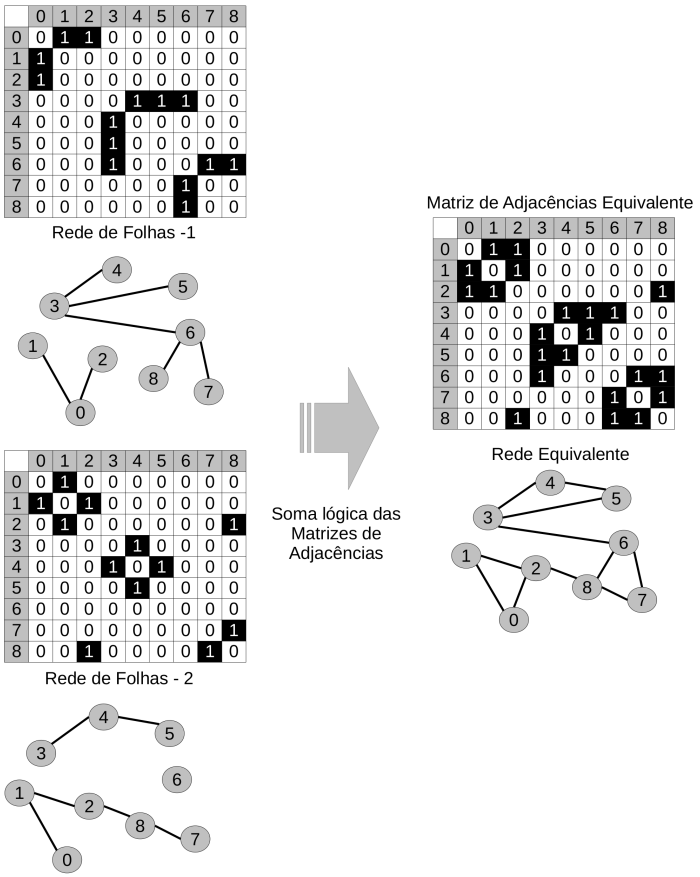



Figura 11 – Redes de Folhas e Rede Equivalente obtidas a População Inicial.

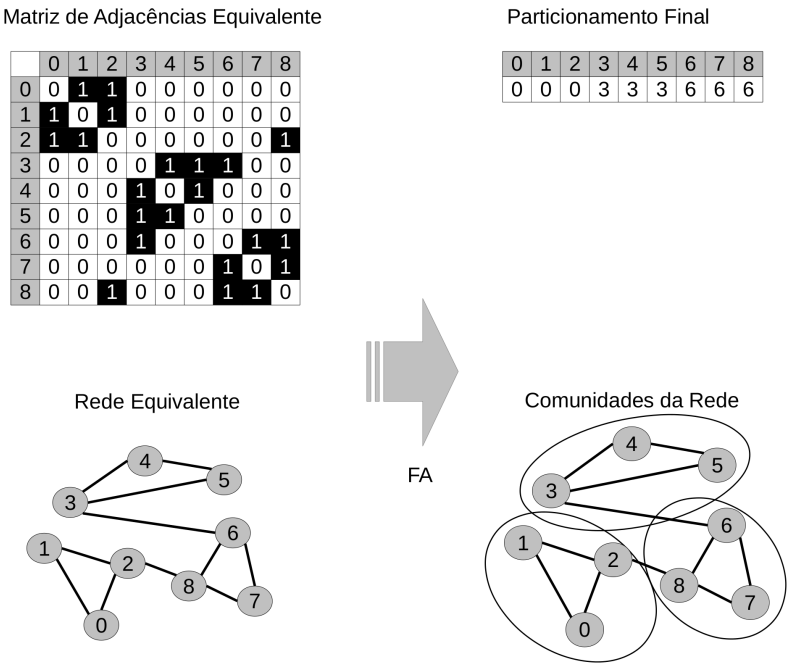



Figura 12 – Identificação das comunidades através da Rede Equivalente.

um esforço para maximizar o desempenho e precisão de algoritmos que lidam com o agrupamento, muitos pesquisadores negligenciaram o fato de que o objetivo principal era, a princípio, a compreensão dos *clusters* formados. Os seus esforços foram voltados para a satisfação de critérios como, por exemplo, a maximização dos graus de similaridade e dissimilaridade entre elementos intra e extra-*clusters*, respectivamente.

Ainda conforme Tzerpos (2001), a compreensão dos *clusters* se deve, principalmente, aos valores assumidos pelos atributos mais relevantes de seus elementos. Assim, os atributos relevantes, acompanhados de seus respectivos valores, representam uma definição para um *cluster*, ou seja, um rótulo, facilitando o trabalho de especialistas ao estudar e interpretar os dados. Chen, Chuang e Chen (2008) e Eltoft e deFigueiredo (1998) se referem à rotulação como o problema de atribuir um rótulo – isto é, um *cluster* – a um elemento desconhecido. Ou seja, se referem ao já conhecido problema de classificação. Lopes, Machado e Rabêlo (2016) define formalmente o problema de rotulação como segue.

**Problema da rotulação.** *Sendo $C$ um conjunto de clusters definido como $C = \{c_1, \ldots, c_K \mid K \geqslant 1\}$, de tal modo que cada cluster contenha um conjunto de elementos $c_i = \{\vec{e}_1, \ldots, \vec{e}_{n^{(c_i)}} \mid n^{(c_i)} \geqslant 1\}$ que podem ser representados por meio de um vetor de atributos definidos em $\mathbb{R}^m$ e expresso por $\vec{e}_j^{(c_i)} = (a_1, \ldots, a_m)$ e ainda que $c_i \cap c_{i'} = \emptyset$ com $1 \leqslant i, i' \leqslant K$ e $i \neq i'$; o objetivo é apresentar um conjunto de rótulos $R = \{r_{c_i}, \ldots, r_{c_k}\}$ no qual cada rótulo específico é dado por um conjunto de pares de valores, atributo e seu respectivo intervalo, $r_{c_i} = \{(a_1, [p_1, q_1]), \ldots, (a_{m^{(ci)}}, ]p_{m^{(ci)}}, q_{m^{(ci)}}[)\}$ capaz de expressar o cluster $c_i$ associado.*

Da definição acima temos que: $K$ é o número de *clusters*; $c_i$ é um *cluster* qualquer; $n^{(c_i)}$ é o número de elementos do *cluster* $c_i$; $\vec{e}_j^{(c_i)}$ se refere ao $j$-ésimo elemento pertencente ao *cluster* $c_i$; $m$ é a dimensão do problema, ou seja, a quantidade de atributos; $r_{c_i}$ é o rótulo referente ao *cluster* $c_i$; $]p_{m^{(ci)}}, q_{m^{(ci)}}]$ representa o intervalo de valores do atributo $a_{m^{(ci)}}$ em que $p_{m^{(ci)}}$ é o limite inferior e $q_{m^{(ci)}}$ é o limite superior; e, por fim, $m^{(ci)}$ é a quantidade de atributos presente em um rótulo referente ao *cluster* $c_i$.

### 1.4.1 Trabalhos Relacionados

Propostas de rotulação de dados baseadas em árvores de decisão como, por exemplo, $C4.5$ e $ID3$ (QUINLAN, 1986), apresentam regras que podem tornar a extração de informações bastante complexa, ou até mesmo inviável. Essas regras dizem respeito ao problema como um todo, e não de forma individual para cada *cluster*. Além disso, essas regras se dispõem misturadas em várias condições, envolvendo os diversos valores de seus atributos.

Embora classificar um elemento desconhecido fazendo uso de árvores de decisão seja uma tarefa simples, bastando verificar as regras de forma hierárquica, até encontrar

o *cluster* ao qual será associado, dificilmente as regras apresentadas serão capazes de representar um *cluster* específico.

A derivação de uma organização hierárquica de conceitos foi proposta por Sanderson e Croft (1999) com o intuito de rotular um conjunto de documentos sem o uso de dados de treinamento ou técnicas de agrupamento padrão. Aqui, um *cluster* é definido por um conjunto de palavras e frases. Os autores utilizam a frequência dos termos entre os documentos para criar um conceito hierárquico definindo rótulos monotéticos (apenas um termo).

Alguns trabalhos como: Glover et al. (2002); Chuang e Chien (2004); Maqbool e Babri (2005), se caracterizam por tratar exclusivamente com informações textuais. O trabalho de Treeratpituk e Callan (2006) se soma aos que objetivam rotular documentos com base no seu conteúdo textual e lida com *clusters* hierárquicos, de modo que os *clusters* podem ser divididos em *sub-clusters* recursivamente. Treeratpituk e Callan (2006) apontam que, embora existam diversos trabalhos relacionados a agrupamento hierárquico, poucos tem o objetivo de defini-los. Além disso, os descritores de *clusters* geralmente falham em fornecer uma descrição compreensiva, que muitas vezes ainda necessitam ser avaliados por um especialista.

O trabalho de Anaya-Sánches, Pons-Porrata e Berlanga-Llavori (2008) é mais um exemplo de rotulação baseada em dados textuais, ao propor uma abordagem para agrupar e rotular documentos, baseada na frequência das palavras. Divergindo da maioria dos trabalhos voltados para a rotulação de dados, o método proposto por Solana-Cipres et al. (2009) usa os conceitos da lógica *fuzzy* para rotular, em tempo real, objetos em imagens de câmeras de segurança em ambientes monitorados.

Yeganova, Comeau e Wilbur (2010) também tem como objetivo a rotulação de textos, ao aplicar técnicas de AM para rotular siglas presentes no contexto da literatura biomédica. Outro exemplo é a abordagem proposta por Cuayáhuitl, Dethlefs e Hastie (2014), que utiliza AM para rotulação em aplicações de processamento de linguagem natural.

Como visto acima, os trabalhos mencionados existentes focam, principalmente, a rotulação de informações textuais, não tendo sido encontrado nenhum trabalho, além do proposto por Lopes, Machado e Rabêlo (2016), envolvendo a rotulação de *clusters* no que diz respeito a apresentar uma definição para se obter conhecimento em relação a atributos numéricos relevantes.

### 1.4.2 Rotulação Automática de *Clusters* - MRA

Lopes, Machado e Rabêlo (2016) propõe a utilização de métodos de aprendizado supervisionado para gerar rótulos automaticamente para *clusters* obtidos a partir da

execução de algoritmos de agrupamento, baseados em aprendizado não supervisionado. De acordo com o método proposto, qualquer algoritmo que faça uso de aprendizado supervisionado pode ser utilizado para rotular *clusters* resultantes do agrupamento realizado por qualquer dos algoritmos de aprendizado não-supervisionado. A Figura 13 mostra todas as etapas do método proposto.

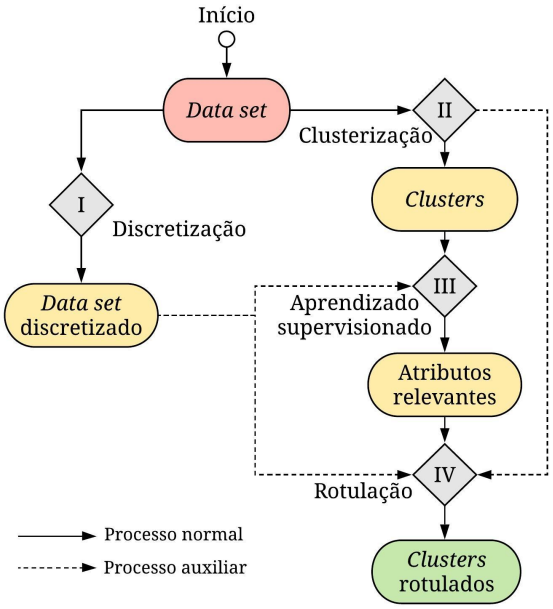



Figura 13 – Modelo proposto por Lopes, Machado e Rabêlo (2016).

O MRA recebe como parâmetro de entrada um conjunto de dados. O método inicia a partir do processo de discretização, representado pela Etapa I da Figura 13. A discretização consiste em atribuir valores discretos para os atributos que podem assumir uma grande variedade de valores em um determinado domínio. Essa Etapa só é aplicada em casos nos quais os atributos possuem, originalmente, valores contínuos. Com isso se espera que o algoritmo de aprendizagem supervisionada utilizado na Etapa III esteja apto a identificar os possíveis relacionamentos existentes entre os atributos com menor complexidade e, consequentemente, podendo obter um aumento significativo na acurácia do rótulo.

A Etapa II corresponde ao processo de geração de *clusters*. Ou seja, o agrupamento, que consiste na associação de elementos em *clusters*, a partir de um conjunto fornecido como entrada. Essa etapa foi introduzida na metodologia para contextualizar um problema real, no qual se tem apenas o conjunto de dados como entrada, fazendo-se necessário formar grupos a partir das amostras inicialmente fornecidas. Nessa Etapa é utilizado o conjunto de dados original e não o conjunto com valores discretizados. Esse é utilizado apenas nas Etapas III e IV, em que ocorre a fase de rotulação. Cabe ressaltar que se parte do princípio de que qualquer algoritmo utilizando o paradigma de aprendizado não

supervisionado seja capaz de lidar com a tarefa de agrupamento.

De posse dos *clusters* obtidos na Etapa anterior se dá início à Etapa III, o processo de rotulação propriamente dito. Redes Neurais Artificiais (RNA) do tipo *Perceptron* Multi-Camadas são utilizadas para obter os atributos relevantes para os *clusters*. Para cada atributo dos elementos de um dado *cluster* é criada uma RNA. Essas RNAs apresentam como saída o valor estimado do atributo avaliado (atributo classe) e como entrada os valores dos demais atributos. As RNAs de um mesmo grupo trabalham com os mesmos elementos variando somente a maneira como esses elementos são utilizados, entrada ou saída, como ilustrado pela Figura 14.

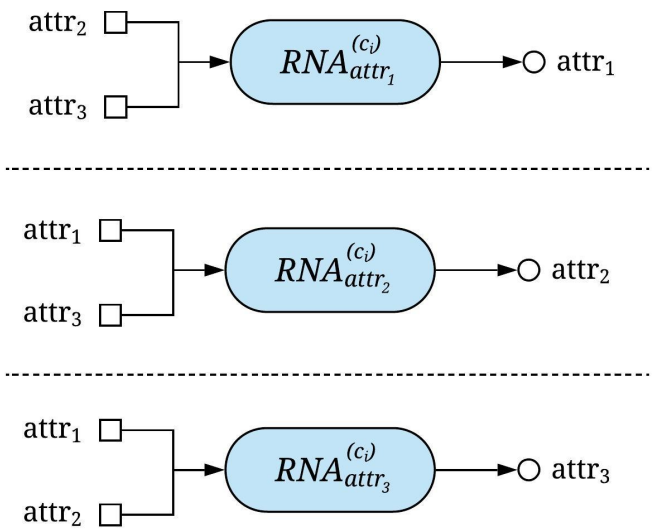



Figura 14 – RNAs para seleção de atributos de um *cluster*.

Cada RNA é criada de forma a representar e avaliar a importância de um atributo em relação aos demais, para cada grupo. Por exemplo, a $RNA^{(c_i)}_{attr_1}$ avalia a relevância do atributo $attr_1$ em relação aos atributos $attr_2$ e $attr_3$ para o *cluster* $c_i$, em que $i$ representa o índice do *cluster*. A porcentagem de acerto de uma RNA ao avaliar um atributo em relação a um determinado *cluster* indica o quão relevante o atributo é. Assume-se, portanto, que a quantidade de acerto (em %) de uma rede indica se existe relação entre os valores de entrada e o de saída. Assim, um atributo é relevante se puder ter seu valor determinado como uma combinação dos valores dos demais atributos. Os atributos de saída das RNAs com as maiores taxas de acerto em cada grupo, junto com suas respetivas faixas de valores, constituirão o rótulo.

Um parâmetro de variação $v$ é utilizado para eliminar ambiguidades entre rótulos de diferentes grupos, assim, todos os atributos – bem como suas faixas de valores – que obtiveram uma taxa de acerto até uma diferença de $v$ da taxa de acerto máxima são incluídos no rótulo, e os demais são descartados. Como exemplo, caso $v$ tenha o valor 5 e

a maior taxa de acerto para um determinado grupo tenha sido de 95%, todas as RNAs com taxa de acerto a partir de 90% serão selecionadas para compor o rótulo.

A Etapa IV consiste em calcular os intervalos de valores para os atributos selecionados na etapa anterior. Nos casos em que não há ocorrência de valores contínuos, os valores escolhidos são os de maior frequência no *cluster*, assim se espera representar a maioria dos elementos. Já para os casos em que houve o processo de discretização é calculado um intervalo de valores que corresponde aos limites da faixa correspondente aos valores de maior frequência.

# 2 Materiais e Métodos

Para a realização dos experimentos foram utilizados cinco *data sets* obtidos do *UCI Machine Learning*[1]. Esse repositório *online* mantém e disponibiliza uma coleção de dados composta por 381 *data sets*, de vários domínios, para que sejam utilizados pela comunidade científica na análise empírica de algoritmos de Aprendizado de Máquina. Todos os conjuntos de dados selecionados possuem o número de classes conhecido previamente. Mais detalhes acerca de cada um deles são dados a seguir, assim como serão apresentados o método proposto e a metodologia de avaliação.

## 2.1 *Data sets*

**Iris**[2] **- Identificação de plantas.** Este *data set* contém 3 classes com 50 elementos cada e se refere à identificação de plantas. Cada classe corresponde a um tipo específico da planta Iris (FISHER, 1936). Os 150 elementos do *data set* são descritos por 4 características cujos valores são contínuos: comprimento da pétala (CP), largura da pétala (LP), comprimento da sépala (CS) e largura da sépala (LS).

O *data set* Iris possui três *clusters* bem distintos, isto é, os valores dos atributos em *clusters* diferentes são visivelmente distintos, com poucas ocorrências do mesmo valor para elementos pertencentes a diferentes *clusters*. Além disso, o *data set* é balanceado, com cada *cluster* possuindo a mesma quantidade de elementos.

**Seeds**[3] **- Identificação de sementes.** O segundo *data set* usado no experimento diz respeito à identificação de sementes (KULCZYCKI; CHARYTANOWICZ, 2011). Esse conjunto de dados é composto por 210 amostras de 3 tipos de sementes de trigo, sendo 70 amostras de cada tipo, portanto é um conjunto com classes balanceadas. Entretanto, frequentemente são encontrados elementos que possuem o mesmo valor para um determinado atributo, pertencendo a *clusters* diferentes, não havendo o mesmo tipo de divisão bem delimitada vista no caso do *data set* Iris.

Os elementos são descritos por 7 atributos os quais representam as características geométricas das sementes: área, perímetro, densidade, comprimento da semente (CS), largura da semente (LS), coeficiente de assimetria (CA) e comprimento do sulco da semente (CSS).

**Wine**[4] **- Identificação de vinhos.** O terceiro *data set*, Wine (AEBERHARD;

[1] http://archive.ics.uci.edu/ml/index.php
[2] http://archive.ics.uci.edu/ml/datasets/Iris
[3] http://archive.ics.uci.edu/ml/datasets/seeds
[4] http://archive.ics.uci.edu/ml/datasets/Wine

COOMANS; VEL, 1992), é o resultado da análise química de vinhos de uma mesma região da Itália, mas provenientes de três diferentes cultivos (três classes), totalizando 178 amostras. Os elementos estão distribuídos entre as classes da seguinte maneira:

- classe 1: 59 amostras;
- classe 2: 71 amostras;
- classe 3: 48 amostras.

A distribuição dos elementos entre as classes revela que o *data set* Wine é levemente desbalanceado. A análise original determinou a presença de treze componentes em cada um dos três vinhos, que por sua vez são os atributos que caracterizam cada uma das instâncias, a saber: álcool, ácido málico (AM), cinzas, alcalinidade das cinzas (AC), magnésio, flavonoides, fenóis não-flavonoides (NF), proantocianidinas, intensidade da cor (IC), matiz, OD280/OD315 de vinhos diluídos (VD) e prolina.

**Vehicle Silhouettes**[5] **- Identificação de Veículos.** O *data set* Vehicles Silhouettes foi criado com o objetivo de reconhecer objetos 3D a partir de uma imagem 2D. Com isso, pretende-se classificar uma determinada silhueta dentre quatro tipos de veículos, utilizando um conjunto de características extraídas das imagens, as quais continham representações dos objetos em vários ângulos distintos (SIEBERT, 1987).

Os 846 elementos que constituem o conjunto de dados podem ser classificados em 4 classes distintas e balanceadas, com a maioria das classes contendo a mesma quantidade de elementos: *bus*, um ônibus de dois andares (240 elementos); *van*, um modelo específico de van fabricado pela montadora Chevrolet (226 elementos); Saab 9000 (240 elementos) e Opel Manta 400 (240 elementos).

Os elementos são descritos por 18 atributos obtidos por meio de um sistema de processamento de imagens, são eles:

- Densidade;
- Circularidade;
- Distância de Circularidade (DC);
- Proporção do Raio (PR);
- Proporção da Tela no Eixo Principal (PTEP);
- Comprimento Máximo da Proporção na Tela (CMPT);

[5] http://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29

- Taxa de Dispersão (TD);
- Alongamento;
- Retangularidade do Eixo Principal (REP);
- Retangularidade do Comprimento Máximo (RCM);
- Variância Escalada ao Longo do Eixo Maior (VEEM);
- Variância Escalada ao Longo do eixo Menor (VEEMe);
- Raio Escalado de Rotação (RER);
- Distorção Sobre o Eixo Maior (DEM);
- Distorção Sobre o Eixo Menor (DEMe);
- Achatamento Sobre o Eixo Maior (AEM);
- Achatamento Sobre o Eixo Menor (AEMe);
- Proporção de Cavidades (PC).

**Glass**[6] **- Identificação de vidros.** O *data set* Glass faz referência à identificação de tipos de vidros e é composto por 214 elementos (amostras de vidros), caracterizados por 9 atributos, definindo seu Índice de Refração (*IR*) e sua composição química em termos das porcentagens dos óxidos (*Na*, *Mg*, *K*, *Al*, *Si*, *Ca*, *Ba* e *Fe*). Os elementos podem ser organizados em 7 *clusters* diferentes quanto à sua destinação de uso e a presença ou não de tratamento térmico (EVETT; SPIEHLER, 1988), assim distribuídos:

- janelas de construção (vidro com tratamento térmico): 70 elementos;
- janelas de construção (vidro sem tratamento térmico): 76 elementos;
- janelas de veículos (vidro com tratamento térmico): 17 elementos;
- janelas de veículos (vidro sem tratamento térmico): 0 elemento;
- recipientes: 13 elementos;
- utensílios de cozinha: 9 elementos;
- faróis: 29 elementos.

Esse é um *data set* desbalanceado, com quantidades bastante diferentes de elementos em cada *cluster*, havendo até mesmo um *cluster* que não possui elementos, no caso, amostras de vidros de veículos do tipo sem tratamento térmico.

[6] http://archive.ics.uci.edu/ml/datasets/Glass+Identification

## 2.2 Método Proposto

A etapa de clusterização é apontada como de fundamental importância para a acurácia dos rótulos encontrados. Assim, espera-se que quanto mais eficiente a técnica de agrupamento utilizada, maior será a acurácia obtida pela rotulação. A abordagem proposta usa os algoritmos DAMICORE e DAMICORE-2 em conjunto com o MRA para rotular *clusters*, o que permite analisar as relações hierárquicas existentes entre esses. A Figura 15 resume as etapas do método proposto.

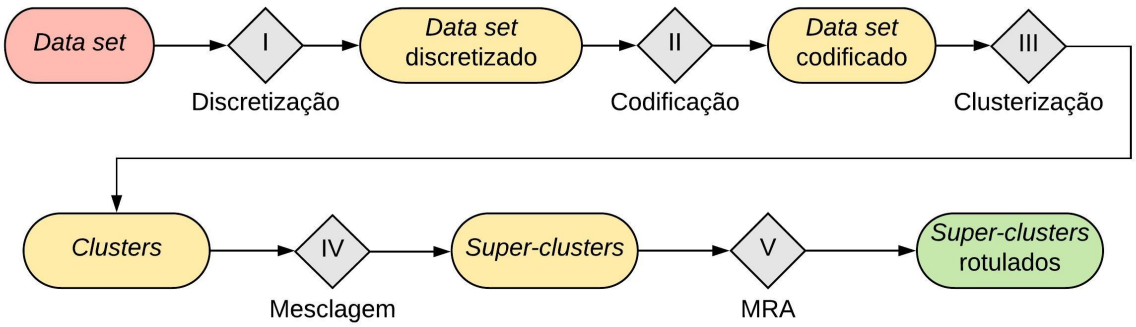
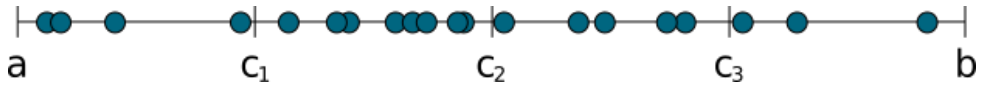



Figura 15 – Fluxograma do método proposto.

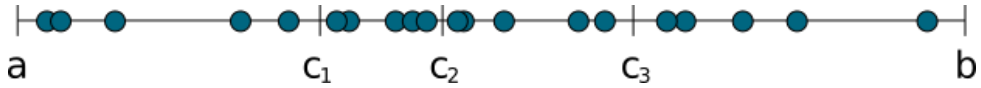
O método se inicia pela etapa de discretização (etapa I). Nesta etapa, todos os atributos que assumem valores contínuos são discretizados utilizando um dos dois métodos não supervisionados mais comuns, de acordo com Kotsiantis e Kanellopoulos (2006), Cerquides e Màntaras (1997) e Dougherty, Kohavi e Sahami (1995): o método EWD (*Equal Width Discretization*), no qual o intervalo de valores assumidos pelo atributo é dividido em faixas de larguras iguais; e o método EFD (*Equal Frequency Discretization*), que divide o intervalo de valores do atributo de forma a alocar a mesma quantidade de valores distintos em cada faixa resultante, ou seja, as faixas podem possuir quantidades de elementos diferentes, porém o número de valores distintos assumidos pelos seus atributos é o mesmo para cada faixa.

As Figuras 16a e 16b exemplificam a aplicação dos métodos EWD e EFD respectivamente, sobre um conjunto com vinte elementos. No exemplo os valores são discretizados em quatro faixas, e os pontos $c_1$, $c_2$ e $c_3$ representam os pontos de corte. A Figura 16a, representa os pontos de corte calculados de forma que todas as faixas de valores tenham a mesma largura. Já a Figura 16b apresenta os pontos de corte definidos de maneira que se mantenha uma quantidade uniforme de valores distintos em cada faixa (cinco elementos por faixa, considerando-se que não há elementos com o mesmo valor para o atributo discretizado, no exemplo em questão).

Para exemplificar o processo foi utilizado um conjunto composto por trinta elementos selecionados aleatoriamente do *data set Iris*, sendo dez de cada uma das três classes




(a) Discretização EWD.




(b) Discretização EFD.

Figura 16 – Exemplificação do uso dos métodos de discretização EWD e EFD sobre um mesmo conjunto de dados.

previamente conhecidas. A Figura 17 apresenta o resultado da etapa de discretização, do tipo EWD, aplicada sobre esse conjunto. *A1*, *A2*, *A3* e *A4* representam os quatro atributos que descrevem os elementos do *data set Iris*, como detalhado na Seção 3.1.

| A1 | A2 | A3 | A4 |
|---|---|---|---|
| 5.0 | 3.6 | 1.4 | 0.2 |
| 5.4 | 3.9 | 1.7 | 0.4 |
| 4.4 | 2.9 | 1.4 | 0.2 |
| 5.4 | 3.9 | 1.3 | 0.4 |
| 4.6 | 3.6 | 1.0 | 0.2 |
| 5.2 | 4.1 | 1.5 | 0.1 |
| 4.9 | 3.1 | 1.5 | 0.1 |
| 5.0 | 3.2 | 1.2 | 0.2 |
| 5.1 | 3.4 | 1.5 | 0.2 |
| 5.0 | 3.5 | 1.6 | 0.6 |
| 4.9 | 2.4 | 3.3 | 1.0 |
| 5.0 | 2.0 | 3.5 | 1.0 |
| 6.1 | 2.9 | 4.7 | 1.4 |
| 5.6 | 2.9 | 3.6 | 1.3 |
| 5.6 | 3.0 | 4.5 | 1.5 |
| 6.2 | 2.2 | 4.5 | 1.5 |
| 6.8 | 2.8 | 4.8 | 1.4 |
| 5.5 | 2.4 | 3.7 | 1.0 |
| 5.6 | 3.0 | 4.1 | 1.3 |
| 5.0 | 2.3 | 3.3 | 1.0 |
| 7.1 | 3.0 | 5.9 | 2.1 |
| 7.6 | 3.0 | 6.6 | 2.1 |
| 6.4 | 2.7 | 5.3 | 1.9 |
| 5.8 | 2.8 | 5.1 | 2.4 |
| 7.2 | 3.2 | 6.0 | 1.8 |
| 7.4 | 2.8 | 6.1 | 1.9 |
| 6.1 | 2.6 | 5.6 | 1.4 |
| 6.7 | 3.3 | 5.7 | 2.5 |
| 6.7 | 3.0 | 5.2 | 2.3 |
| 5.9 | 3.0 | 5.1 | 1.8 |

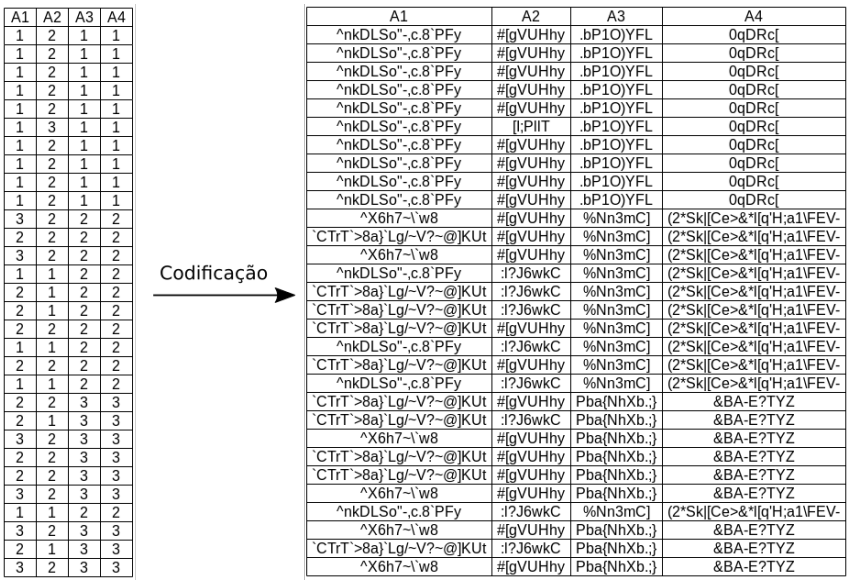
Discretização

| A1 | A2 | A3 | A4 |
|---|---|---|---|
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 3 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 3 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 |
| 3 | 2 | 2 | 2 |
| 1 | 1 | 2 | 2 |
| 2 | 1 | 2 | 2 |
| 2 | 1 | 2 | 2 |
| 2 | 2 | 2 | 2 |
| 1 | 1 | 2 | 2 |
| 2 | 2 | 2 | 2 |
| 1 | 1 | 2 | 2 |
| 2 | 2 | 3 | 3 |
| 2 | 1 | 3 | 3 |
| 3 | 2 | 3 | 3 |
| 2 | 2 | 3 | 3 |
| 2 | 2 | 3 | 3 |
| 3 | 2 | 3 | 3 |
| 1 | 1 | 2 | 2 |
| 3 | 2 | 3 | 3 |
| 2 | 1 | 3 | 3 |
| 3 | 2 | 3 | 3 |

Figura 17 – Etapa de discretização do tipo EWD aplicada a um subconjunto do *data set Iris* durante a fase de pré-processamento.

Na Etapa II, o *data set* discretizado é submetido a um processo de codificação na qual os valores de atributos numéricos são substituídos por códigos alfanuméricos. Essa fase visa reforçar a diferença entre valores como, por exemplo, 1 e 11 – que por vezes são considerados mais próximo que 1 e 2. Isso ocorre devido ao fato de a NCD utilizar algoritmos de compressão, que por sua vez, fazem uso de métodos de ordenação lexicográfica para determinar a similaridade entre os dados a serem comprimidos. Dessa maneira, as etapas I e II permitem ao DAMICORE e ao DAMICORE-2 medir a similaridade dos elementos com maior precisão, contribuindo para a obtenção de agrupamentos mais significantes. A Figura 18 ilustra o resultado dessa Etapa II.

O *data set* codificado é então submetido ao DAMICORE, ou DAMICORE-2, para realização da clusterização, que corresponde à Etapa III da Figura 15. É obtida ao final

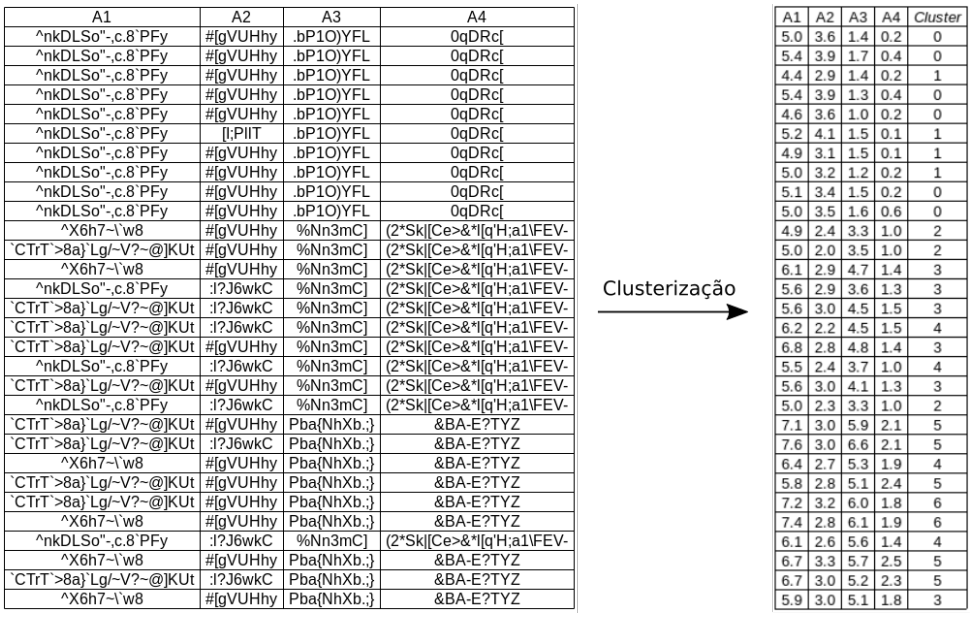



| A1 | A2 | A3 | A4 |
|---|---|---|---|
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 3 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 1 | 2 | 1 | 1 |
| 3 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 |
| 3 | 2 | 2 | 2 |
| 1 | 1 | 2 | 2 |
| 2 | 1 | 2 | 2 |
| 2 | 1 | 2 | 2 |
| 2 | 2 | 2 | 2 |
| 1 | 1 | 2 | 2 |
| 2 | 2 | 2 | 2 |
| 1 | 1 | 2 | 2 |
| 2 | 2 | 3 | 3 |
| 2 | 1 | 3 | 3 |
| 3 | 2 | 3 | 3 |
| 2 | 2 | 3 | 3 |
| 2 | 2 | 3 | 3 |
| 3 | 2 | 3 | 3 |
| 1 | 1 | 2 | 2 |
| 3 | 2 | 3 | 3 |
| 2 | 1 | 3 | 3 |
| 3 | 2 | 3 | 3 |

| A1 | A2 | A3 | A4 |
|---|---|---|---|
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | [l;PlIT | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^X6h7~\`w8 | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^X6h7~\`w8 | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | Pba{NhXb.;} | &BA-E?TYZ |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | Pba{NhXb.;} | &BA-E?TYZ |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |

Figura 18 – Resultado da aplicação da etapa de codificação sobre subconjunto do *data set* *Iris* discretizado.

dessa etapa, uma lista contendo o índice de cada elemento seguido por um número inteiro representando o *cluster* no qual o elemento foi alocado. Esses valores são adicionados ao *data set* original, como exemplificado pela Figura 19.




| A1 | A2 | A3 | A4 |
|---|---|---|---|
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | [l;PlIT | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^nkDLSo"-,c.8`PFy | #[gVUHhy | .bP1O)YFL | 0qDRc[ |
| ^X6h7~\`w8 | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^X6h7~\`w8 | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | Pba{NhXb.;} | &BA-E?TYZ |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| ^nkDLSo"-,c.8`PFy | :l?J6wkC | %Nn3mC] | (2*Sk\|[Ce>&*l[q'H;a1\FEV- |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |
| `CTrT`>8a}`Lg/~V?~@]KUt | :l?J6wkC | Pba{NhXb.;} | &BA-E?TYZ |
| ^X6h7~\`w8 | #[gVUHhy | Pba{NhXb.;} | &BA-E?TYZ |

| A1 | A2 | A3 | A4 | *Cluster* |
|---|---|---|---|---|
| 5.0 | 3.6 | 1.4 | 0.2 | 0 |
| 5.4 | 3.9 | 1.7 | 0.4 | 0 |
| 4.4 | 2.9 | 1.4 | 0.2 | 1 |
| 5.4 | 3.9 | 1.3 | 0.4 | 0 |
| 4.6 | 3.6 | 1.0 | 0.2 | 0 |
| 5.2 | 4.1 | 1.5 | 0.1 | 1 |
| 4.9 | 3.1 | 1.5 | 0.1 | 1 |
| 5.0 | 3.2 | 1.2 | 0.2 | 1 |
| 5.1 | 3.4 | 1.5 | 0.2 | 0 |
| 5.0 | 3.5 | 1.6 | 0.6 | 0 |
| 4.9 | 2.4 | 3.3 | 1.0 | 2 |
| 5.0 | 2.0 | 3.5 | 1.0 | 2 |
| 6.1 | 2.9 | 4.7 | 1.4 | 3 |
| 5.6 | 2.9 | 3.6 | 1.3 | 3 |
| 5.6 | 3.0 | 4.5 | 1.5 | 3 |
| 6.2 | 2.2 | 4.5 | 1.5 | 4 |
| 6.8 | 2.8 | 4.8 | 1.4 | 3 |
| 5.5 | 2.4 | 3.7 | 1.0 | 4 |
| 5.6 | 3.0 | 4.1 | 1.3 | 3 |
| 5.0 | 2.3 | 3.3 | 1.0 | 2 |
| 7.1 | 3.0 | 5.9 | 2.1 | 5 |
| 7.6 | 3.0 | 6.6 | 2.1 | 5 |
| 6.4 | 2.7 | 5.3 | 1.9 | 4 |
| 5.8 | 2.8 | 5.1 | 2.4 | 5 |
| 7.2 | 3.2 | 6.0 | 1.8 | 6 |
| 7.4 | 2.8 | 6.1 | 1.9 | 6 |
| 6.1 | 2.6 | 5.6 | 1.4 | 4 |
| 6.7 | 3.3 | 5.7 | 2.5 | 5 |
| 6.7 | 3.0 | 5.2 | 2.3 | 5 |
| 5.9 | 3.0 | 5.1 | 1.8 | 3 |

Figura 19 – Aplicação da etapa de agrupamento utilizando DAMICORE ou DAMICORE-2 aos dados codificados.

O DAMICORE não requer o conhecimento prévio da quantidade de classes na qual se distribuem os elementos do *data set*. Nesse caso, o número de *clusters* é determinado pelo próprio algoritmo, por meio da quantidade de comunidades detectadas. No entanto, a árvore filogenética reconstruída pelo NJ fornece a relação hierárquica existente entre os elementos de cada *cluster*. Nesse caso, os *clusters* que possuem o mesmo ancestral comum

estão diretamente relacionados entre si. Com isso, o MRA pode ser utilizado para rotular os nós internos significativos (nós internos que são ancestrais comuns a todos os elementos de um *cluster* ou mais). A Figura 20 destaca os nós internos (numerados) para a árvore filogenética reconstruída para o conjunto de dados de exemplo, com os seus respectivos rótulos, determinados pelo MRA.

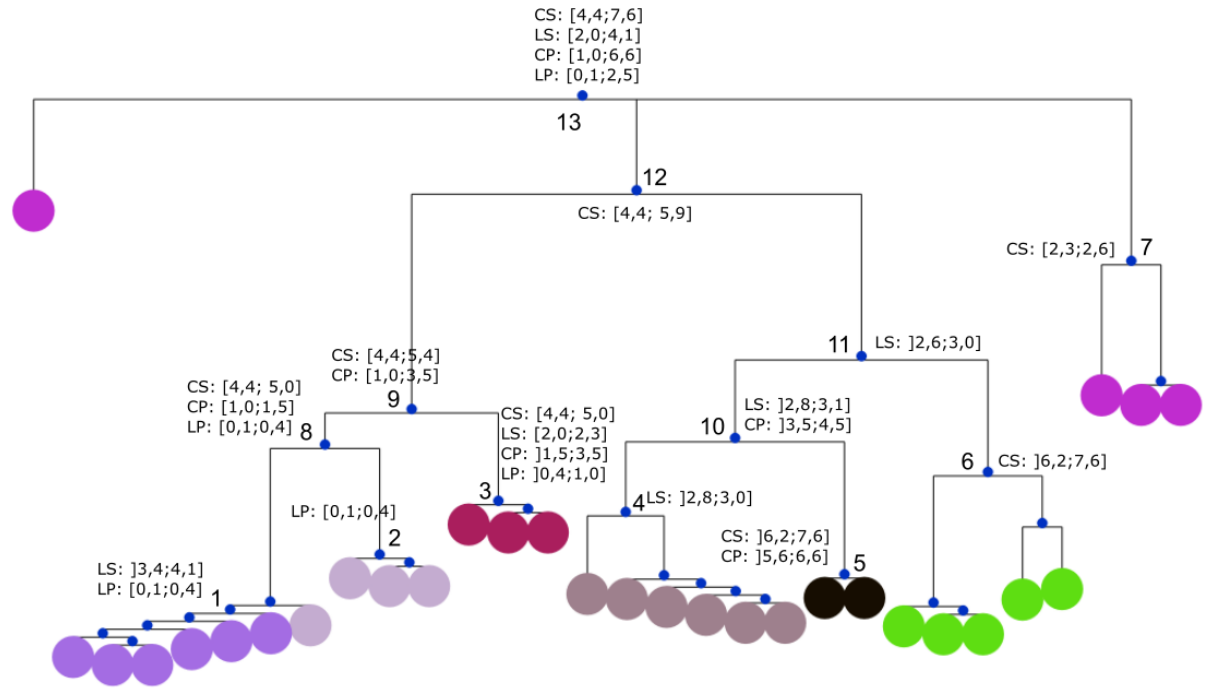



Figura 20 – Nós internos da árvore filogenética com seus respectivos rótulos, determinados pelo MRA.

Resultados empíricos apontam que, frequentemente, o número de *clusters* resultante é superior àquele determinado pela literatura. Em face disso, por meio da árvore filogenética reconstruída, os *clusters* podem ser mesclados até que se obtenha a quantidade desejada, facilitando a compreensão dos rótulos, que diminui quanto maior for o número de nós a serem rotulados. A mesclagem desses *clusters* tem como consequência a formação de *clusters* maiores, os *super-clusters*, os quais representam um conceito mais geral. Com a redução do número de nós os rótulos se tornam mais compreensíveis. Essa tarefa é realizada na Etapa IV da Figura 15. A Figura Figura 21 demonstra os passos dessa Etapa.

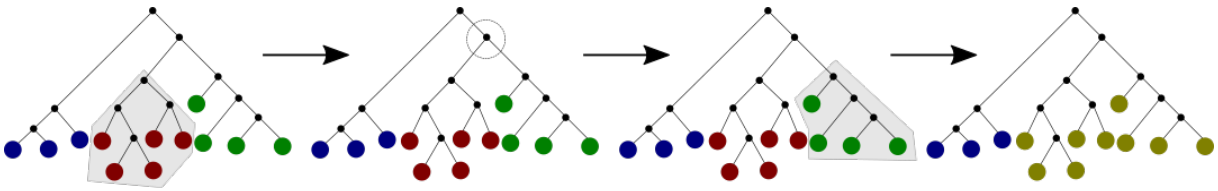


Figura 21 – Exemplificação da etapa de mesclagem.

Finalmente, os *super-clusters* são submetidos ao MRA (Etapa V). Nessa etapa os *super-clusters* são rotulados, permitindo identificar quais os atributos mais relevantes

e quais os seus respectivos intervalos de valores. A análise desses rótulos possibilita compreender a relação existente entre os elementos as razões pelas quais foram alocados no mesmo *cluster*. O resultado dessa etapa pode ser conferido na Figura 22.

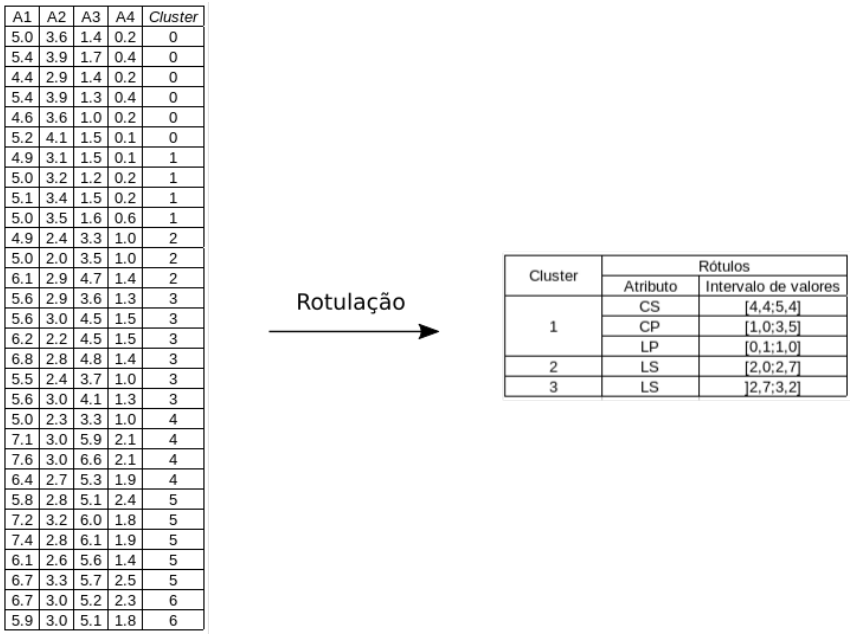



Figura 22 – Resultado final do MRA com os rótulos definidos para os *clusters* resultantes do agrupamento.

## 2.3 Experimentos e Metodologia de Avaliação

Como já afirmado no início do capítulo todos os *data sets* possuem o número de classes conhecido previamente. Esse fato permite que a etapa de mesclagem seja realizada até que se obtenha a quantidade de *super-clusters* igual àquela apontada pela literatura. Além da quantidade de *super-clusters*, a Tabela 1 resume os valores utilizados para o número de faixas da discretização e para o parâmetro de variação em todos os *data sets*. Notadamente o número de faixas de discretização escolhida para cada bateria de testes foi o mesmo número de classes apontado pela literatura.

Tabela 1 – Parâmetros de teste.

| ***Data set*** | ***# super-clusters*** | **Faixas de discretização** | **Variação ($v$)** |
|---|---|---|---|
| Iris | 3 | 3 | 5 |
| Seeds | 3 | 3 | 10 |
| Wine | 3 | 3 | 5 |
| Vehicle Silhouettes | 4 | 4 | 2 |
| Glass | 6 | 6 | 15 |

Os *data sets* foram agrupados primeiramente utilizando o algoritmo K-means e em seguida os *clusters* formados foram rotulados com o MRA, permitindo a comparação

com a abordagem utilizada no trabalho original. Em seguida o processo se repetiu, dessa vez utilizando os métodos DAMICORE e DAMICORE-2 para o agrupamento dos dados, aplicando a abordagem proposta. Foram realizadas dez execuções de cada algoritmo para cada *data set*. A partir dos resultados obtidos foi calculada a Acurácia Média (AM), ou seja, a média aritmética das acurácias de cada execução.

Além disso, usando a Equação 2.1, é calculada a Acurácia Parcial (AP), correspondente à acurácia de cada atributo em cada *cluster*. $AP_{q,r}$ é a acurácia parcial do atributo $q$ em relação ao *cluster* $r$. $T_{q,r}$ é o número de elementos do *cluster* $r$ que se enquadram no intervalo de valores determinado pelo rótulo para o atributo $q$. Finalmente, $N_r$ é o número de elementos no *cluster* $r$.

$$PA_{q,r} = \frac{T_{q,r}}{N_r}, \tag{2.1}$$

A análise dos valores da acurácia parcial de cada atributo diferente presente no rótulo permite o ajuste do parâmetro de variação, adicionando ou removendo atributos. Isso possibilita reduzir a ambiguidade entre os rótulos, mantendo o valor da acurácia dentro de um intervalo aceitável. Além disso, para complementar a avaliação, foi calculado o desvio padrão, como medida de dispersão em torno da média populacional

Por fim, os mesmos procedimentos foram repetidos utilizando o DAMICORE-2 na Etapa de clusterização. Os parâmetros foram ajustados para que o DAMICORE-2 reconstruísse dez Árvores Filogenéticas, em que cada uma dará origem a dez Redes, totalizando, ao final do processo, cem Redes. Esses valores foram utilizados devido à indicação de testes preliminares de que, para a maioria dos *data sets*, com uma configuração de sete árvores e sete redes o algoritmo atinge o que foi nomeado de ponto de saturação de *clusters*, no qual o algoritmo não consegue aumentar sua capacidade de generalização e o número de *clusters* formados se fixa. Com isso, os parâmetros utilizados garantem uma margem capaz de englobar uma maior amplitude de conjuntos de dados.

## 2.4 Considerações Finais

Neste Capítulo foram apresentados a abordagem proposta e a metodologia de avaliação utilizada na realização dos testes. No próximo Capítulo são apresentados os resultados obtidos com relação à rotulação dos *super-clusters*, enfatizando a configuração e parâmetros nas quais foram alcançados.

# 3 Resultados e Discussão

A seguir são apresentados os resultados da aplicação do Método de Rotulação Automática de *clusters* (MRA) proposto por Lopes, Machado e Rabêlo (2016). Foram utilizados cinco conjuntos de dados sobre os quais foram aplicados os algoritmos K-means, DAMICORE e DAMICORE-2, com a finalidade de realizar o agrupamento de dados. Todos os *data sets* são caracterizados por atributos do tipo numérico, os quais podem ser manipulados por RNAs. Além disso, o número de classes existente é previamente estabelecido pela literatura. Os parâmetros foram os mesmos utilizados em Lopes, Machado e Rabêlo (2016), sempre que possível.

As Tabelas 2 a 16 apresentam os resultados obtidos pela aplicação do MRA sobre os *clusters* formados pelo K-means, DAMICORE e DAMICORE-2 para os cinco *data sets* utilizados. Ressalta-se que são apresentados somente os melhores resultados em relação ao método de discretização (EWD ou EFD). Para efeito de ilustração, são incluídas as árvores antes e depois do processo de mesclagem no *data set* Iris para os algoritmos DAMICORE e DAMICORE-2. A árvores são apresentadas em forma circular para melhor visualização.

A leitura das Tabelas citadas acima se dá da seguinte maneira: a primeira coluna contém o índice que identifica os *clusters*, enquanto a segunda contém a quantidade de elementos que o respectivo *cluster* possui. As duas colunas seguintes apresentam os atributos que compõem os rótulos e suas respectivas faixas de valores. A coluna seguinte armazena a relevância de cada atributo para a definição do rótulo do *cluster*. As duas últimas colunas armazenam a quantidade de erros (número de elementos que não satisfazem as condições estabelecidas para cada atributo do rótulo, ou seja, estão fora da faixa de valores estabelecida), e a acurácia parcial (porcentagem de elementos que satisfazem as condições) para cada atributo individualmente. Um elemento é considerado corretamente rotulado se, e somente se, os valores de seus atributos coincidirem com os determinados pelos rótulos.

## 3.1 Iris

Os resultados obtidos são apresentados na Tabela 2. Destaca-se que não houve repetição de rótulos entre os *clusters*. Os atributos CP e LP se fazem presentes tanto no *cluster* 1 quanto no 2 e possuem faixas de valores distintas. Além disso o *cluster* 2 possui o atributo CS constituindo o seu rótulo, sendo que o mesmo não ocorre com o *cluster* 1. O *cluster* 3 se destaca por possuir a menor acurácia parcial, fato devido ao elevado número de erros acarretados pela inclusão do atributo LS. Isso revela que o fato de *clusters* possuírem rótulos distintos não significa que os mesmos foram rotulados corretamente, pois os rótulos

encontrados podem não representar a realidade dos elementos que compõem o *cluster* analisado.

O melhor resultado da aplicação do MRA aos *clusters* formados pelo K-means para os elementos do *data set Iris* foi obtido utilizando discretização por larguras iguais. Na Tabela 2 fica claro que a maior quantidade de elementos desobedecendo os intervalos de valores definidos para os atributos do rótulo ocorre no *cluster* 3, no qual os atributos possuem os menores valores de relevância.

Tabela 2 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo K-means no *data set Iris*.

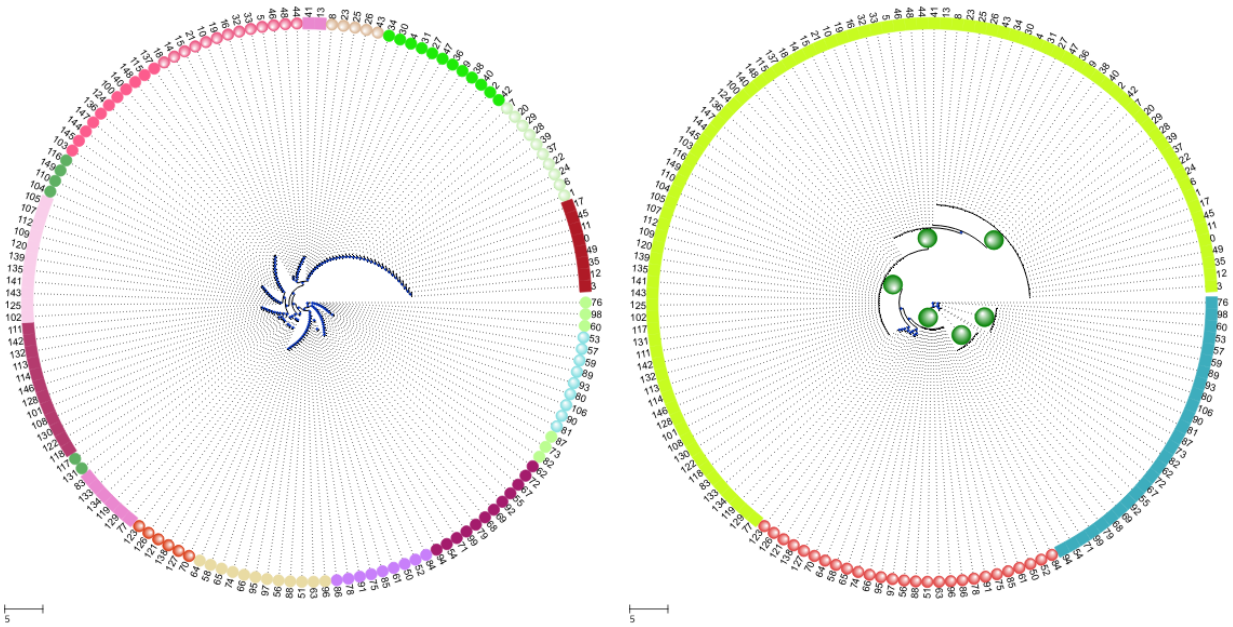
| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 38 | CP | ]4,93;6,9] | 100 | 1 | 97,37 |
| | | LP | ]1,7;2,5] | 93,33 | 3 | 92,11 |
| 2 | 50 | CS | [4,3;5,5] | 100 | 3 | 94 |
| | | CP | [1,0;2.97] | 100 | 0 | 100 |
| | | LP | [0,1;0,9] | 100 | 0 | 100 |
| 3 | 62 | CS | ]5,5;6,7] | 80 | 9 | 85,48 |
| | | LS | [2,0;2,8] | 72 | 24 | 61,29 |

Para este *data set* o DAMICORE definiu uma média de 19 *clusters*. Após a etapa de mesclagem resultaram 3 *clusters*, valor pré-definido pela literatura. A Figura 23a ilustra a distribuição dos elementos nos *clusters* antes da etapa de mesclagem, enquanto a 23b ilustra a distribuição, após o processo de mesclagem, nos *super-clusters*. Os nós preenchidos com gradiente verde representam os pontos em que houve mesclagem de *clusters*.

Assim como no experimento anterior, o método de discretização EWD apresentou os melhores resultados. Mais uma vez, não se observou a presença de *clusters* com rótulos iguais, como pode ser visto na Tabela 3. O *cluster* 3 apresentou a menor acurácia parcial, embora o único atributo que o represente possua relevância de 100%, com isso, o rótulo definido para este *cluster* não é capaz de defini-lo, uma vez que apenas 52,08% dos elementos obedecem ao intervalo definido.

Tabela 3 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE no *data set Iris*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 27 | CS | ]5,2;6,7] | 100 | 3 | 88,89 |
| | | CP | ]2,97;4,93] | 100 | 0 | 100 |
| 2 | 27 | LS | [2,0;2,8] | 100 | 0 | 100 |
| | | CP | ]2,97;4,93] | 100 | 0 | 100 |
| | | LP | ]0,9;1,7] | 100 | 0 | 100 |
| 3 | 96 | CP | [1,0;2,97] | 100 | 46 | 52,08 |



(a) Distribuição dos elementos antes da mesclagem. (b) Distribuição dos elementos após a mesclagem.

Figura 23 – Exemplificação do uso dos métodos de discretização EWD e EFD sobre um mesmo conjunto de dados.

Quanto aos resultados da rotulação do agrupamento formado pelo DAMICORE-2, diferente do ocorrido com os demais, a discretização do tipo EFD alcançou melhores resultados, os quais são exibidos na Tabela 4. O DAMICORE-2 já forneceu de início 3 *clusters*. Portanto a etapa de mesclagem foi omitida e a Figura 24 ilustra a distribuição dos elementos entre os *clusters* fornecida diretamente pelo DAMICORE-2.

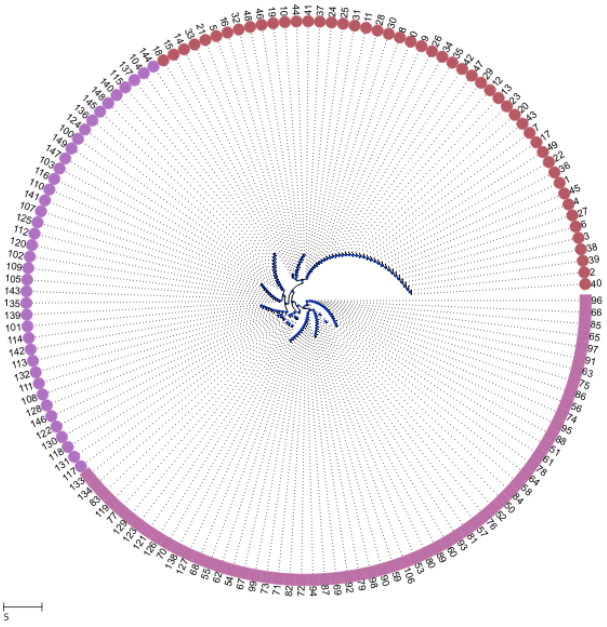


Figura 24 – Distribuição dos elementos nos *clusters* obtidos.

A semelhança da quantidade de elementos por *cluster* com o agrupamento realizado pelo K-means é notável ao se comparar as Tabelas 2 e 4. Entretanto as acurácias parciais obtidas pelos rótulos atribuídos aos *clusters* obtidos pelo DAMICORE-2 são superiores às obtidas pelo K-means, o que resulta em uma acurácia média também superior. Ainda da Tabela 4 destaca-se o fato de os *clusters* 2 e 3 possuírem o mesmo atributo, porém com intervalos de valores distintos, o que elimina a ambiguidade entre eles.

Tabela 4 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE-2 no *data set Iris*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 50 | CS | ]4,3;5,4] | 90 | 5 | 90 |
| | | CP | [1,0;3,8] | 100 | 0 | 100 |
| | | LP | [0,1;1,1] | 100 | 0 | 100 |
| 2 | 60 | LP | ]1,1;1,8] | 95,83 | 8 | 86,67 |
| 3 | 40 | LP | ]1,8;2,5] | 87,50 | 0 | 100 |

## 3.2 Seeds

Os resultados atingiram acurácia superior quando o *data set* Seeds foi discretizado utilizando o método EWD, em todos os experimentos realizados. A Tabela 5 mostra os resultados da rotulação aplicada ao agrupamento realizado pelo K-means, deixando claro o fraco desempenho do método para o agrupamento em questão, obtendo baixos valores de acurácia parcial na grande maioria dos atributos, fazendo com que os rótulos não fossem capazes de definir os *clusters*, resultando na menor acurácia média alcançada pelo K-means entre os cinco experimentos realizados.

Os resultados obtidos para este *data set* utilizando o DAMICORE são exibidos na Tabela 6. O DAMICORE determinou um total de 22 *clusters* para este *data set*. Utilizando como parâmetro o valor apontado pela literatura a etapa de mesclagem reduziu esse número para 3 *clusters*. Os rótulos resultantes foram capazes de definir os *clusters* de maneira única, entretanto a baixa acurácia parcial obtida pelo atributo Área no *cluster* 1, com um total de 42 elementos cujos valores assumidos para o referido atributo não estão de acordo com o determinado pelo rótulo, fizeram com que a acurácia média fosse de apenas 62,81%.

O DAMICORE-2 conseguiu se aproximar mais do número de *clusters* apontados pela literatura, resultando em 6 *clusters*, reduzidos para 3 após a etapa de mesclagem. A análise da Tabela 7 permite visualizar que os *clusters* foram rotulados de maneira distinta. Porém, assim como ocorre no experimento anterior, um dos atributos possui baixa acurácia parcial, neste caso o atributo Perímetro do *cluster* 2, com o valor de apenas 55,45%, levando o MRA a atingir a acurácia média de 72,76%.

Tabela 5 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo K-means no *data set Seeds*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 21 | Área | [10,59;14,12] | 100 | 10 | 52,38 |
| | | Perímetro | [12,41;14,02] | 100 | 10 | 52,38 |
| | | Densidade | [0,81;302,21] | 100 | 1 | 95,24 |
| | | CS | ]4451,69;6675,0] | 100 | 0 | 100 |
| | | CSS | [4,83;2169,22] | 100 | 0 | 100 |
| 2 | 168 | Área | [10,59;14,12] | 97,01 | 93 | 44,64 |
| | | Perímetro | [12,41;14,02] | 98,51 | 101 | 39,88 |
| | | CS | ]4451,69;6675,0] | 100 | 0 | 100 |
| | | CSS | ]4333,61;6498,0] | 100 | 0 | 100 |
| 3 | 21 | Área | [10,59;14,12] | 100 | 7 | 66,67 |
| | | CS | [5,09;2228,39] | 100 | 0 | 100 |
| | | CSS | ]4333,61;6498,0] | 100 | 2 | 90,48 |

Tabela 6 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE no *data set Seeds*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 142 | Área | [10,59;14,12] | 100 | 42 | 70,42 |
| | | Densidade | [0,81;302,21] | 91,23 | 16 | 88,73 |
| 2 | 63 | Perímetro | ]15,64;17,25] | 92 | 8 | 87,3 |
| | | Densidade | [0,81;302,21] | 92 | 8 | 87,3 |
| | | CS | ]4451,6967;6675,0] | 96 | 4 | 93,65 |
| | | LS | ]2689,54;4033,0] | 96 | 5 | 92,06 |
| | | CSS | ]4333,61;6498,0] | 88 | 6 | 90,48 |
| 3 | 5 | Área | ]14,12;17,65] | 100 | 0 | 100 |
| | | Perímetro | ]14,02;15,64] | 100 | 1 | 80 |
| | | Densidade | [0,81;302,21] | 100 | 0 | 100 |
| | | CS | ]4451.7;6675,0] | 100 | 0 | 100 |
| | | LS | ]2689,5;4033,0] | 100 | 0 | 100 |
| | | CA | ]2819,2;5637,6] | 100 | 1 | 80 |
| | | CSS | ]4333,6;6498,0] | 100 | 0 | 100 |

## 3.3 Wine

Diferente do ocorrido com os *data sets* Iris e Seeds, os melhores resultados da rotulação dos *clusters* formados pelo K-means foram obtidos por meio do método de discretização EFD. A Tabela 8 exibe os resultados desse experimento. Os rótulos não se repetiram pois, embora os *clusters* 1 e 2 tenham tido o mesmo atributo selecionado como o mais relevante, os seus intervalos de valores são diferentes. Quanto à acurácia, destaca-se que o rótulo definido para o *cluster* 3 não é atendido por mais da metade dos elementos do *cluster*, alcançando o valor de apenas 37,68% para o atributo OD280/OD315 de vinhos

Tabela 7 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE-2 no *data set Seeds*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 37 | Densidade | [0,81;302.21] | 100 | 1 | 97,3 |
| | | LS | ]2689,5;4033,0] | 93,33 | 2 | 94,59 |
| 2 | 101 | Perímetro | ]15.64;17,25] | 92,50 | 45 | 55,45 |
| | | CS | ]4451,7;6675,0] | 97,50 | 4 | 96,04 |
| | | LS | ]2689,5;4033,0] | 97,50 | 6 | 94,06 |
| | | CSS | ]4333,6;6498,0] | 87,50 | 9 | 91,09 |
| 3 | 72 | Área | [10,59;14,12] | 100 | 0 | 100 |
| | | Perímetro | [12,41;14,02] | 100 | 0 | 100 |
| | | CSS | ]4333,6;6498,0] | 93,10 | 9 | 87,50 |

diluídos, fazendo com que a acurácia média fosse a segunda menor entre os três algoritmos testados.

Tabela 8 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo K-means no *data set Wine*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 47 | Prolina | ]885,0;1680,0] | 100 | 0 | 100 |
| 2 | 62 | Prolina | ]560,0;885,0] | 100 | 4 | 93,55 |
| 3 | 69 | VD | [1,27;2,23] | 85,71 | 43 | 37,68 |
| | | Prolina | [278,0;560,0] | 82,14 | 6 | 88,41 |

Os rótulos definidos para os *clusters* formados pelo DAMICORE (usando discretização do tipo EWD) podem ser vistos na Tabela 9. O algoritmo obteve inicialmente um total de 21 *clusters*, reduzidos a 3 com a etapa de mesclagem. Destacam-se os altos valores de acurácia parcial atingidas pelos rótulos dos *clusters* 2 e 3, em contraste com o *cluster* 1. Esse fato está diretamente relacionado à quantidade de elementos alocados em cada *cluster*. Enquanto o *cluster* 3 possui apenas 9 elementos, o *cluster* 1 teve 131 elementos alocados, fazendo com que a complexidade em se encontrar um rótulo capaz de representar todos os elementos fosse maior, levando ao total de 39 elementos incorretamente rotulados.

A Tabela 10 traz os resultados obtidos quando aplicado o MRA aos *clusters* formados pelo DAMICORE-2, que inicialmente forneceu 6 *clusters*. Todos os *clusters* foram rotulados de maneira única, sendo os *clusters* 1 e 2 distinguidos pelos intervalos de valores do atributo Flavonóides. O maior número de ocorrências de elementos incorretamente rotulados se deu no *cluster* 1, porém apenas 8 em um total de 98 elementos. Com isso, os rótulos definidos para este agrupamento alcançaram as maiores acurácias médias entre os três experimentos realizados com o *data set* Wine, atingindo uma acurácia média de 90,56%.

Tabela 9 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE no *data set Wine*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 131 | IC | [1,28;5,19] | 90,38 | 39 | 70,23 |
| 2 | 38 | Cinzas | ]1,98;2,61] | 100 | 2 | 94,74 |
| | | AC | ]17,07;23,53] | 100 | 0 | 100 |
| | | Flavonóides | [0,34;1,92] | 100 | 0 | 100 |
| | | VD | [1,27;2,18] | 100 | 3 | 92,11 |
| 3 | 9 | Cinzas | ]2.61;3,23] | 100 | 0 | 100 |
| | | AC | ]23,53;30,0] | 100 | 0 | 100 |
| | | Flavonóides | [0,34;1,92] | 100 | 0 | 100 |
| | | Saturação | [0,48;0,89] | 100 | 1 | 88,89 |
| | | VD | [1,27;2,18] | 100 | 0 | 100 |

Tabela 10 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE-2 no *data set Wine*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 98 | Flavonóides | ]1,92;3,5] | 94,87 | 8 | 91,84 |
| 2 | 42 | Flavonóides | [0,34;1,92] | 100 | 1 | 97,62 |
| 3 | 38 | Cinzas | ]1,98;2,61] | 100 | 2 | 94,74 |
| | | AC | ]17.07;23,53] | 100 | 0 | 100 |
| | | Flavonóides | [0,34;1,92] | 100 | 0 | 100 |
| | | VD | [1,27;2,18] | 100 | 3 | 92,11 |

## 3.4 Vehicle Silhouettes

Assim como acontece com o *data set* Wine, o melhor resultado do MRA ao rotular *clusters* fornecidos pelo K-means foi obtidos através da utilização do método EFD para o processo de discretização. Na Tabela 11 são exibidos os resultados da aplicação do MRA na execução deste experimento. Facilmente se percebe que não houve repetição de rótulos entre os *clusters* (os *clusters* 3 e 4 são distinguidos pelos intervalos de valores). Além disso nota-se que a menor acurácia parcial tem o valor de 96,81%, com apenas 10 elementos incorretamente rotulados de um total de 846. Isso demonstra que o rótulos são capazes de definir corretamente e de maneira única cada um dos *clusters* resultantes.

Os resultados da aplicação da rotulação automática sobre os *clusters* formados pelo DAMICORE são exibidos na Tabela 12. O DAMICORE agrupou os 846 elementos do conjunto de dados em 44 *clusters*, utilizando o méto de discretização EWD. A etapa de mesclagem de *clusters* reduziu esse valor para o total de 4 *clusters*. Ao analisar os rótulos não se nota repetições, entretanto o elevado número de elementos incorretamente rotulados no *cluster* 4 fez com que a acurácia média para o experimento em questão atingisse o valor de 83,07%. Além disso, a distribuição dos elementos pelos *clusters* ocorre de maneira

Tabela 11 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo K-means no *data set Vehicle Silhouetes*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 169 | TD | [112,0;146,0] | 100 | 0 | 100 |
| | | REP | [17,0;20,0] | 100 | 0 | 100 |
| | | VEEMe | [184,0;320,0] | 100 | 0 | 100 |
| 2 | 188 | DC | ]92,0;112,0] | 100 | 6 | 96,81 |
| | | Alongamento | [26,0;34,0] | 100 | 0 | 100 |
| 3 | 360 | REP | [17,0;20,0] | 99,31 | 4 | 98,89 |
| 4 | 129 | REP | ]20,0;23,0] | 98,08 | 0 | 100 |

bastante distinta da apontada pela literatura, onde é afirmado que cada classe possui 240 elementos, aproximadamente.

Tabela 12 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE no *data set Vehicle Silhouetes*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 20 | PTEP | [47,0;69,75] | 100 | 0 | 100 |
| | | CMPT | [2,0;1,25] | 100 | 0 | 100 |
| | | REP | [17,0;20,0] | 100 | 0 | 100 |
| | | RCM | ]135,5;153,0] | 100 | 0 | 100 |
| | | DEM | [59,0;78,0] | 100 | 0 | 100 |
| | | AEMe | ]183,5;191,0] | 100 | 2 | 90 |
| 2 | 102 | REP | [17,0;20,0] | 100 | 0 | 100 |
| | | VEEMe | [184,0;392,5] | 100 | 0 | 100 |
| 3 | 475 | CMPT | [2,0;15,25] | 99,47 | 8 | 98,32 |
| 4 | 249 | PTEP | [47,0;69,75] | 99 | 2 | 99,20 |
| | | CMPT | [2,0;15,25] | 100 | 1 | 99,60 |
| | | TD | [112,0;150,25] | 99 | 32 | 87,15 |
| | | REP | [17,0;20,0] | 99 | 3 | 98,80 |
| | | VEEM | [130,0;177,5] | 100 | 21 | 91,57 |

O DAMICORE-2, usando o método EWD na etapa de discretização, agrupou os elementos formando 7 *clusters*, posteriormente mesclados até resultar em 4 *clusters*, como exibido pela Tabela 13. Todos os *clusters* foram rotulados de maneira distinta pelo MRA. É possível perceber que, com excessão do atributo DEM do *cluster* 4, todos os demais obtiveram acurácia parcial superior a 90%. Com isso o MRA atingiu acurácia média superior ao rotular o agrupamento obtido pelo DAMICORE-2 em relação ao DAMICORE, embora tenha sido superado quando comparado àquele alcançado ao se rotular o agrupamento fornecido pelo K-means.

Tabela 13 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE-2 no *data set Vehicle Silhouetes*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 114 | PTEP | [47,0;69,75] | 100 | 1 | 99,12 |
| | | CMPT | [2,0;15,25] | 100 | 0 | 100 |
| | | DEM | [59,0;78,0] | 100 | 2 | 98,25 |
| 2 | 234 | CMPT | [2,0;15,25] | 100 | 0 | 100 |
| 3 | 139 | CMPT | [2,0;15,25] | 100 | 1 | 99,28 |
| | | TD | [112,0;150,25] | 100 | 0 | 100 |
| | | REP | [17,0;20,0] | 100 | 0 | 100 |
| | | VEEM | [130,0;177,5] | 100 | 0 | 100 |
| | | VEEMe | [184,0;392,50] | 100 | 0 | 100 |
| 4 | 359 | CMPT | [2,0;15,25] | 97,22 | 11 | 96,94 |
| | | DEM | [59,0;78,0] | 97,22 | 78 | 78,27 |

## 3.5 Glass

Rotulando-se os *clusters* resultantes da execução do K-means, foram obtidos os resultados apresentados pela Tabela 14. O tipo de discretização utilizado foi o EWD. O elemento *Ba* é o mais frequente, estando presente em vários dos *clusters*. Em todas os rótulos nos quais aparece, o *Ba* se encontra na mesma faixa de valores. Entretanto, está sempre acompanhado de outros elementos que tornam os rótulos distintos. A acurácia parcial alcançou valores satisfatórios em quase todos os casos, tendo como exceção o atributo *RI* do *cluster* 3, com 12 elementos que não satisfazem o intervalo definido e um valor de 65,71%. No entanto os valores dos demais atributos fizeram com que a acurácia média da rotulação fosse de 87,24%.

A Tabela 15 apresenta os resultados da rotulação obtidos ao se rotular o *data set Glass* utilizando o DAMICORE e o método de discretização EWD. O DAMICORE organizou os elementos em 21 *clusters* que, após o processo de mesclagem, foram reduzidos a 6. É possível visualizar que os *clusters* 1 e 4 possuem o mesmo rótulo, além disso, outro fato fácil de ser notado é a presença do atributo *Ba* na maioria dos *clusters*. Com relação aos demais *clusters* não são identificadas repetições de rótulos, tendo a maioria dos atributos atingido valores satisfatórios de acurácia parcial, embora a acurácia média tenha sido de 85,37%, inferior à obtida no experimento anterior, porém ainda satisfatória.

Foram obtidos 8 *clusters* após a execução do DAMICORE-2. Com a etapa de mesclagem resultaram 6 *clusters*, valor apontado pela literatura para o *data set* em questão. De acordo com a Tabela 16, de maneira similar ao que ocorreu no experimento anterior, os *clusters* 1 e 4 foram rotulados de maneira idêntica, porém também os *clusters* 2 e 3 receberam o mesmo rótulo, indicando a existência de *super-clusters* ainda maiores. Isso é compatível com uma segunda possível maneira de se agrupar os elementos, uma vez

Tabela 14 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo K-means no *data set Glass*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 124 | Ba | [0,0;0,15] | 100 | 0 | 100 |
| 2 | 7 | RI | ]1,52;1,53] | 100 | 0 | 100 |
| | | Mg | [0,0;2,24] | 100 | 0 | 100 |
| | | Ca | ]9,85;16,19] | 100 | 0 | 100 |
| | | Ba | [0,0;0,15] | 100 | 1 | 85,71 |
| 3 | 26 | Na | ]14,20;17,38] | 90 | 4 | 84,62 |
| | | Mg | [0,0;2,24] | 100 | 0 | 100 |
| | | Fe | [0,0;0,07] | 90 | 3 | 88,46 |
| 4 | 5 | K | ]0,68;6,21] | 100 | 0 | 100 |
| | | Ca | [5,43;8,07] | 100 | 0 | 100 |
| | | Fe | [0,0;0,07] | 100 | 0 | 100 |
| 5 | 35 | RI | ]1,52;1,53] | 71,43 | 12 | 65,71 |
| | | Ba | [0,0;0,15] | 85,71 | 3 | 91,43 |
| 6 | 17 | Mg | [0,0;2,24] | 100 | 0 | 100 |
| | | Ba | [0,0;0,15] | 85,71 | 1 | 94,12 |

que podem ser classificados quanto à presença, ou não, de tratamento térmico. A menor acurácia parcial se deu no *cluster* 5, com o valor de 65,52%, no entanto os altos valores obtidos pelos demais permitiram que se alcançasse a acurácia média de 92,25%, a maior para o *data set* Glass. O tipo de discretização utilizado neste experimento foi o EFD.

## 3.6 Considerações Finais

A Figura 25 apresenta um gráfico comparativo entre as taxas de acerto da rotulação resultantes da aplicação do MRA aos agrupamentos obtidos pelo K-means, pelo DAMICORE e pelo DAMICORE-2 sobre os cinco *data sets* citados anteriormente. Os valores fazem referência à acurácia média, em que o número de acertos corresponde ao total de elementos que se enquadram nas faixas de valores de todos os atributos que compõem o rótulo de determinado *cluster*. Já os valores entre parenteses correspondem ao desvio padrão obtido em cada caso.

A acurácia média do MRA para os agrupamentos do DAMICORE e do DAMICORE-2, em relação à alcançada utilizando os agrupamentos do K-means, foi superior na maioria dos casos, tendo como única exceção o *data set* Vehicles, no qual a rotulação dos *clusters* obtidos pelo K-means atingiu acurácia média superior aos demais.

Para o *data set* Iris, a acurácia média com o agrupamento do DAMICORE-2 (90,26%) ficou 27,39% acima da obtida para o agrupamento do DAMICORE, que foi de apenas 62,87%. Além disso, a média obtida pelo DAMICORE-2 também teve o menor

Tabela 15 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE no *data set Glass*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 165 | K | [0,0;1,04] | 95,45 | 4 | 97,58 |
| | | Ba | [0,0;0,53] | 96,97 | 2 | 98,79 |
| 2 | 12 | RI | ]1,51;1,52] | 80 | 5 | 58,33 |
| | | Fe | [0,0;0,09] | 80 | 2 | 83,33 |
| 3 | 8 | K | [0,0;1,04] | 100 | 0 | 100 |
| | | Ca | ]9,02;10,81] | 100 | 3 | 62,5 |
| | | Ba | [0,0;0,53] | 100 | 0 | 100 |
| | | Fe | [0,0;0,09] | 100 | 1 | 87,50 |
| 4 | 12 | K | [0,0;1,04] | 100 | 0 | 100 |
| | | Ba | [0,0;0,53] | 100 | 0 | 100 |
| 5 | 10 | Na | ]14,06;15,16] | 100 | 1 | 90 |
| | | Mg | [0,0;0,75] | 100 | 1 | 90 |
| | | K | [0,0;1,04] | 100 | 0 | 100 |
| | | Ca | ]7,22;9,02] | 100 | 0 | 100 |
| | | Fe | [0,0;0,09] | 100 | 0 | 100 |
| 6 | 7 | RI | ]1.51;1,52] | 100 | 0 | 100 |
| | | Na | ]14,06;15,16] | 100 | 0 | 100 |
| | | Mg | [0,0;0,75] | 100 | 0 | 100 |
| | | K | [0,0;1,04] | 100 | 0 | 100 |
| | | Ca | ]7,22;9,02] | 100 | 2 | 71,43 |
| | | Ba | ]0.53;1,05] | 100 | 0 | 100 |
| | | Fe | [0,0;0,09] | 100 | 1 | 85,71 |

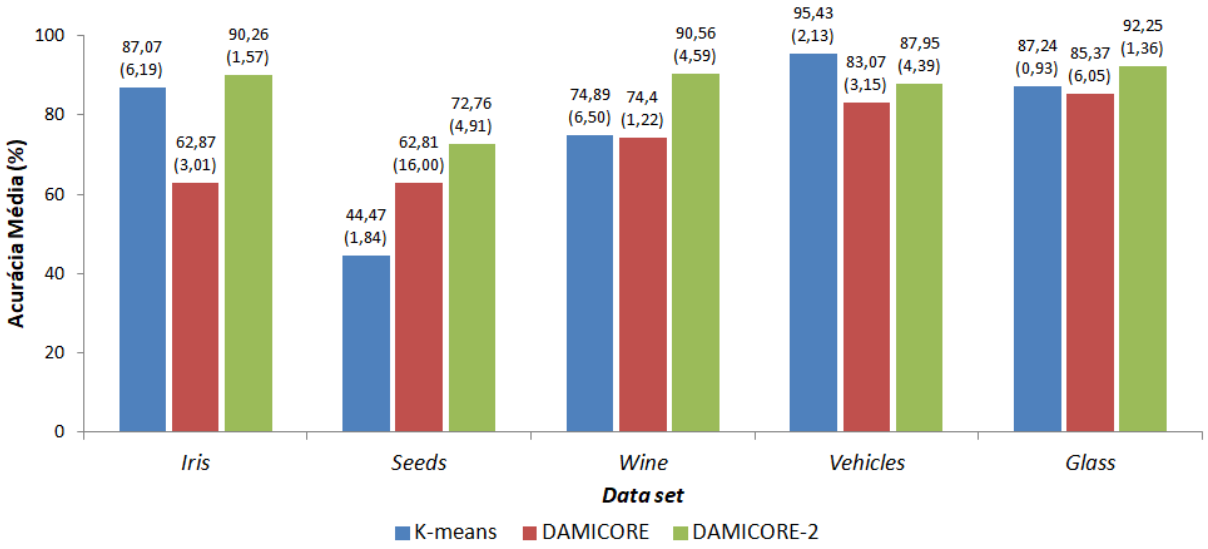



Figura 25 – Gráfico comparativo entre as taxas de acerto da rotulação dos grupos obtidos pelo K-means, pelo DAMICORE e pelo DAMICORE-2 para os *data sets* testados.

desvio padrão. A rotulação do agrupamento do K-means obteve acurácia média de 87,07%,

Tabela 16 – Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE-2 no *data set Glass*.

| ***Cluster*** | **# Elem.** | **Rótulos** | | **Rel. (%)** | **Análise** | |
|---|---|---|---|---|---|---|
| | | **Atrib.** | **Intervalo** | | **# Erros** | **AP (%)** |
| 1 | 56 | Ba | [0,0;0,15] | 90,91 | 2 | 96,43 |
| 2 | 54 | Ba | [0,0;0,15] | 100 | 0 | 100 |
| 3 | 26 | Ba | [0,0;0,15] | 100 | 0 | 100 |
| 4 | 27 | Ba | [0,0;0,15] | 100 | 0 | 100 |
| 5 | 29 | Mg | [0,0;2,24] | 100 | 3 | 89,66 |
| | | Al | ]2,02;3,5] | 91,67 | 10 | 65,52 |
| | | Fe | [0,0;0,07] | 91,67 | 4 | 86,21 |
| 6 | 22 | Mg | [0,0;2,24] | 88,89 | 2 | 90,91 |
| | | Ba | [0,0;0,15] | 88,89 | 2 | 90,91 |

apenas 3,19% abaixo do alcançado com o DAMICORE-2, porém, com desvio padrão de 6,19. Quanto à rotulação dos *clusters* do *data set* Seeds, embora a acurácia média tenha sido de apenas 72,76% (a menor dentre as obtidas pelo DAMICORE-2), ela está 28,29 pontos percentuais acima da obtida ao se rotular o agrupamento do K-means – que foi de apenas 44,47% – e 9,95% superior à alcançada pelo DAMICORE (62,81%), sendo esta a média com menor desvio padrão.

Com relação ao *data set* Wine, as acurácias médias da rotulação dos *clusters* obtidos pelo K-means e pelo DAMICORE atingiram valores muito próximos, inclusive levando em consideração o desvio padrão. Enquanto o primeiro obteve como resultado 74,89%, o segundo alcançou 74,40%, resultando em uma diferença de apenas 0,59%. Os *clusters* formados pelo DAMICORE-2 obtiveram acurácia média de 90,56%, sendo a maior obtida para o *data set* Wine, com desvio padrão de 4,59.

O *data set* Vehicle Silhouetes protagonizou o único caso onde a acurácia média da rotulação dos *clusters* fornecidos pelo K-means foi superior aos demais, sendo o valor de 95,43% com apenas 2,13 de desvio padrão. Enquanto isso, o DAMICORE alcançou 83,07% de acurácia média, sendo superado também pelo DAMICORE-2, cuja acurácia média foi de 87,95%. Último testado, no *data set* Glass se deu o maior equilíbrio entre os valores da acurácia média. A diferença entre o maior e o menor valor obtidos foi de apenas 5,01%. Novamente o DAMICORE-2 atingiu o maior valor (92,25%). No entanto, foi seguido de perto pelo K-means e pelo DAMICORE, com valores respectivamente iguais a 87,24% e 85,37%, porém este com desvio padrão de 6,05, enquanto aquele apresentou média com desvio padrão de apenas 0,93.

# 4 Conclusões e Trabalhos Futuros

Neste Capítulo são apresentados os principais pontos, tanto positivos quanto negativos, da utilização dos métodos de DAMICORE e DAMICORE-2 para construção de *clusters* com a finalidade de rotulação automática pelo MRA. Adicionalmente, são listados alguns itens passíveis de execução em trabalhos futuros.

## 4.1 Conclusões

Foi apresentada uma abordagem para utilização do Método de Rotulação Automática (MRA) para rotular *clusters* formados pelo DAMICORE e pelo DAMICORE-2. Os resultados obtidos foram comparados com os apresentados em Lopes, Machado e Rabêlo (2016), que realizou a rotulação automática sobre *clusters* obtidos pelo *K-means*. A análise dos resultados mostrou que os rótulos obtidos da aplicação do MRA sobre os *clusters* formados pelo DAMICORE-2 possuem maior acurácia em comparação à alcançada pela aplicação sobre o agrupamento do K-means e do DAMICORE.

A eficácia da rotulação obtida para o agrupamento realizado pelo DAMICORE-2 é atribuída à capacidade de representar a relação entre os elementos de um mesmo *cluster*. Ao fazer uso de várias árvores filogenéticas combinadas na construção de suas redes, o DAMICORE-2 leva em consideração o consenso entre elas, reforçando as relações hierárquicas existentes. Reforça-se ainda o fato de que a técnica de clusterização é determinante para a qualidade dos rótulos atribuídos pelo MRA, pois quanto maior for a semelhança intra-*cluster* maior será a acurácia dos rótulos encontrados. Assim, quanto maior for a capacidade de representação das relações entre os elementos e entre os próprios *clusters*, maior será a acurácia resultante.

O atributo variação $v$ também demonstrou ter fundamental importância para a acurácia média do método proposto, permitindo distinguir os rótulos dos *clusters*. Ainda que o aumento da variação possa acarretar uma diminuição na acurácia média, o fato não ocorre para todos os casos, pois depende diretamente das características do próprio *data set* (por exemplo, se a base é balanceada ou não), e de como os atributos se relacionam.

Com isso, o ajuste deve ser feito buscando um equilíbrio, aceitando a utilização de atributos menos relevantes apenas enquanto se mantém um valor aceitável para a acurácia média. Para a determinação do valor da variação a acurácia parcial toma enorme relevância, sendo ela que indica se o valor deverá ser incrementado ou decrementado para que se tenha ganho na acurácia média ou se permita distinguir os *clusters* de forma única dentro de uma faixa de valores aceitável para a acurácia média.

Embora exista uma gama de parâmetros que possam ser ajustados na busca de melhorias, os resultados obtidos foram satisfatórios, alcançando acurácia média acima de 90% em três dos cinco experimentos realizados, rotulando corretamente quase a totalidade dos elementos. Ainda assim, etapas como a codificação e a discretização podem impactar de maneira significativa o resultado final. Um método de codificação que seja capaz de representar da melhor maneira a similaridade entre os valores dos atributos pode possibilitar que as árvores reconstruídas sejam capazes de indicar mais precisamente a relação hierárquica inter-*clusters*.

Outro parâmetro importante é o tipo de discretização. Neste trabalho foram utilizados apenas métodos que dividem um conjunto de valores em um número de faixas pré-determinado e constante para todos os atributos, ou seja, todos os atributos têm seus intervalos de valores divididos em um mesmo número de faixas. Assim, futuramente pretende-se utilizar métodos de discretização que sejam capazes de calcular o número de faixas adequado para cada atributo, podendo potencialmente aumentar a acurácia da rotulação.

## 4.2 Trabalhos Futuros

Dentre os principais trabalhos futuros passíveis de execução estão:

- A utilização de métodos de discretização que calculem a quantidade adequada de faixas para cada atributo;
- A realização de testes com conjuntos de dados com maior quantidade de elementos;
- Aprimorar o método de codificação;

# Referências

AEBERHARD, S.; COOMANS, D.; VEL, O. D. Comparison of classifiers in high dimensional settings. *Department of Mathematics and Statistics, James Cook University, North Queensland, Australia, Technical Report*, n. 92-02, 1992. Citado na página 22.

ANAYA-SÁNCHES, H.; PONS-PORRATA, A.; BERLANGA-LLAVORI, R. A new document clustering algorithm for topic discovering and labeling. In: *13th Iberoamerican Congress on Pattern Recognition - CIARP 2008*. [S.l.]: LNCS, 2008. p. 161–168. Citado 2 vezes nas páginas 1 e 17.

BRAGA, A. P.; CARVALHO, A.; LUDEMIR, T. *Redes Neurais Artificiais*: Teoria e aplcações. 2. ed. Rio de Janeiro: LTC, 2007. Citado na página 6.

CANCINO, W.; DELBEM, A. Inferring phylogenies by multi-objective evolutionary algorithm. In: *International journal of information technology and intelligent computing*. [S.l.: s.n.], 2007. p. 1–26. Citado na página 9.

CERQUIDES, J.; MÀNTARAS, R. L. de. Proposal and empirical comparison of a parallelizable distance-based discretization method. In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 1997. p. 139–142. Citado na página 24.

CHEN, H.-L.; CHUANG, K.-T.; CHEN, M.-S. On data labeling for clustering categorical data. *Knowledge and Data Engineering, IEEE Transactions*, n. 1458-1472, 2008. Citado na página 16.

CHUANG, S.-L.; CHIEN, L.-F. A practical web-based approach to generating topic hierarchy for text segments. *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management - CIKM*, n. 127-136, 2004. Citado na página 17.

CILIBRASI, R.; VITÁNYI, P. Clustering by compression. In: *IEEE Transactions on Information Theory*. [S.l.]: University of California Press, 2005. p. 1523–1545. Citado na página 8.

CUAYÁHUITL, H.; DETHLEFS, N.; HASTIE, H. A semi-supervised clustering approach for semantic slot labelling. *13th International Conference on Machine Learning and Applications*, n. 500-505, 2014. Citado na página 17.

DIESTEL, R. *Graph Theory*. Springer, 2006. (Electronic library of mathematics). ISBN 9783540261834. Disponível em: <http://books.google.com.br/books?id=aR2TMYQr2CMC>. Citado na página 13.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. In: *12nd International Conference on Machine Learning - ICML*. [S.l.]: Morgan Kaufmann, 1995. p. 194–202. Citado na página 24.

DUCH, J.; ARENAS, A. Community detection in complex networks using extremal optimization. In: *Physical Review E*. [S.l.: s.n.], 2005. p. 406–425. Citado na página 10.

ELTOFT, T.; DEFIGUEIREDO, R. A self-organizing neural network for cluster detection and labeling. *IEEE International Joint Conference on Neural Networks Proceedings*, v. 1, n. 408-412, 1998. Citado na página 16.

EVETT, I.; SPIEHLER, E. Rule induction in forensic science. In: *Knowledge Based Systems*. [S.l.]: Halsted Press, 1988. p. 152–160. Citado na página 23.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. In: *Advances in Knowledge Discovery and Data Mining*. [S.l.]: AAAI Press, 1996. p. 37–54. Citado na página 1.

FEIGENBAUM, E. A.; MACCORDUCK, P. *The Fifth Generation*: Artificial intelligence and japan's computer challenge to the world. [S.l.]: Addison-Wesley, 1983. Citado na página 5.

FISHER, R. The use of multiple measurements in taxonomic problems. In: *Annals of Eugenics*. [S.l.: s.n.], 1936. p. 17–188. Citado na página 21.

GLOVER, E. et al. Inferring hierarchical descriptions. *Proceedings of the Eleventh International Conference on Information and Knowledge Management - CIKM*, n. 507-514, 2002. Citado na página 17.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining*: Concepts and techniques. 3. ed. [S.l.]: Morgan Kaufmann Publishers Inc., 2011. Citado na página 1.

KOTSIANTIS, S.; KANELLOPOULOS, D. Discretization techniques: A recent survey. In: *GESTS International Transactions on Computer Science and Engineering*. [S.l.: s.n.], 2006. p. 47–58. Citado na página 24.

KULCZYCKI, P.; CHARYTANOWICZ, M. A complete gradient clustering algorithm. In: *Proceedings of the Third International Conference on Artificial Intelligence and Computational Intelligence*. [S.l.]: Springer-Verlag, 2011. p. 497–504. Citado na página 21.

LI, M.; VITÁNYI, P. M. *An Introduction to Kolmogorov Complexity and Its Applications*. 2. ed. [S.l.]: Springer-Verlag New York, Inc., 1997. Citado na página 9.

LILLO-CASTELLANO, J. et al. Weaning outcome prediction from heterogeneous time series using normalized compression distance and multidimensional scaling. *Expert Systems with Applications*, v. 40, n. 5, p. 1737 – 1747, 2013. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417412010810>. Citado na página 8.

LIPSCHUTZ, S.; LIPSON, M. *Matemática Discreta: Coleção Schaum*. BOOKMAN COMPANHIA ED, 2004. ISBN 9788536303611. Disponível em: <http://books.google.com.br/books?id=2S9bwDmD1P0C>. Citado na página 14.

LOPES, A.; MACHADO, V.; RABÊLO, R. Automatic labelling of clusters of discrete and continuous data with supervised machine learning. In: *Knowledge-Based Systems*. [S.l.]: LNCS, 2016. p. 231–241. Citado 7 vezes nas páginas 15, 2, 16, 17, 18, 31 e 43.

LUGER, G. F. *Inteligência Artificial*. 6. ed. São Paulo: Pearson Education do Brasil, 2013. 614 p. ISBN 978-85-8143-550-3. Citado 2 vezes nas páginas 5 e 6.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probabilitys*. [S.l.]: University of California Press, 1967. p. 281–297. Citado na página 2.

MAQBOOL, O.; BABRI, H. Interpreting clustering results through cluster labeling. *Proceedings of the IEEE Symposium on Emerging Technologies*, n. 429-434, 2005. Citado na página 17.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. Citado na página 5.

NEWMAN, M.; GIRVAN, M. Finding and evaluating community structure in networks. In: *Physical Review E*. [S.l.: s.n.], 2004. p. 406–425. Citado na página 10.

QUINLAN, J. R. Induction of decision trees: Machine learning. n. 81-106, 1986. Citado na página 16.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligende*: A modern approach. 2. ed. [S.l.]: Prentice-Hall, 2003. Citado na página 7.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. In: *Molecular Biology and Evolution*. [S.l.: s.n.], 1987. p. 406–425. Citado na página 10.

SANCHES, A.; CARDOSO, J.; DELBEM, A. Identifying merge-beneficial software kernels for hardware implementation. In: *Reconfigurable Computing and FPGAs (ReConFig)*. [S.l.]: AAAI Press, 2011. p. 74–79. Citado 2 vezes nas páginas 1 e 8.

SANDERSON, M.; CROFT, B. Deriving concept hierarchies from text. In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1999. (SIGIR '99), p. 206–213. ISBN 1-58113-096-1. Citado na página 17.

SIEBERT, J. P. Technical Report, *Vehicle Recognition Using Rule Based Methods*. 1987. Citado na página 22.

SIMON, H. A. *The Sciences of the Artificial*. 2. ed. [S.l.]: MIT Press, 1981. Citado na página 5.

SOARES, A.; RABÊLO, R.; DELBEM, A. Optimization based on phylogram analysis. *Expert Systems with Applications*, v. 78, p. 32 – 50, 2017. ISSN 0957-4174. Citado na página 8.

SOLANA-CIPRES, C. J. et al. Automatic object labelling for monitored environments using clustering techniques. *3rd International Conference on Crime Detection and Prevention (ICDP)*, 2009. Citado na página 17.

TREERATPITUK, T.; CALLAN, J. Automatically labeling hierarchical clusters. *Proceedings of the 2006 International Conference on Digital Government Research*, n. 167-176, 2006. Citado na página 17.

TZERPOS, V. *Comprehension-Drive Software Clustering*. PhD thesis, 2001. Citado 2 vezes nas páginas 14 e 16.

YEGANOVA, L.; COMEAU, D. C.; WILBUR, W. J. Identifying abbreviation definitions: Machine learning with naturally labeled data. *Ninth International Conference on Machine Learning and Applications*, n. 500-505, 2010. Citado na página 17.