Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 17, Iss. 1, pp. 4-17, 2019

ROTULAÇÃO AUTOMÁTICA DE *CLUSTERS* BASEADA EM ANÁLISE DE FILOGRAMAS

Francisco N. C. de Araújo, Antonio H. M. Soares, Vinicius P. Machado, Ricardo de A. L. Rabêlo
Universidade Federal do Piauí
{netoaraujjo, ahelsonms}@gmail.com, {vinicius,ricardoalr}@ufpi.edu.b

Resumo – 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. 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). A tarefa de rotulação consiste em identificar os *clusters* através de suas características mais relevantes. Os algoritmos utilizados são reconhecidamente eficientes, obtendo resultados satisfatórios nas definições dos *clusters* formados, frequentemente superando taxas de acerto de 90% nos experimentos realizados.

Palavras-chave – Aprendizado de máquina, agrupamento, reconhecimento de padrões.

Abstract – This paper proposes the joint use of unsupervised and supervised Machine Learning methods for data clustering and labeling tasks, respectively. 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. The labeling task consists in identifying the clusters through their most relevant characteristics. The algorithms used are known to be efficient, obtaining satisfactory results in the definitions of the clusters formed, frequently exceeding 90% accuracy in the done experiments.

Keywords – Machine learning, clustering, pattern recognition.

1. 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 [1].

Uma das principais técnicas de reconhecimento de padrões é o agrupamento (*clustering*), o qual visa organizar os dados em grupos (*clusters*). 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*) [2] mostrou ser capaz de encontrar correlações em bases de dados de tipos mistos (registros que possuem diferentes tipos de dados).

Embora tenha sido um dos focos principais dos pesquisadores, o processo de *clustering* não fornece informações que permitam inferir de forma clara as características de cada *cluster* formado, o que se deve a limitações das métricas de distância utilizadas [3]. A rotulação de dados visa identificar essas características e permitir então que se tenha a plena compreensão dos *clusters* resultantes.

A rotulação de um *cluster* busca resumir sua definição, ou seja, descrevê-lo em função de seus atributos mais relevantes, e suas respectivas faixas de valores, a fim de melhor compreendê-lo. Assim, este 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.

Em Lopes et al. [4] é proposta a utilização de Redes Neurais Artificiais para identificar quais os atributos relevantes, e suas respectivas faixas de valores, que juntos formam o rótulo de um determinado *cluster*, ou seja, determinam as características predominantes pelas quais os elementos foram alocados em um mesmo *cluster*. A abordagem proposta por Lopes et al. [4] obteve resultados positivos, conseguindo rotular *clusters* com taxa de acerto média de 85%, aos ser aplicada em agrupamentos realizados pelo algoritmo *K-means* [5].

O método de *clustering* utilizado é um dos fatores de maior influência sobre a acurácia da rotulação. Assim, quanto melhor o agrupamento realizado, maior será a capacidade dos rótulos encontrados definirem os *clusters*. Neste artigo apresenta-se a utilização do Método de Rotulação Automática (MRA) [4] para rotular os *clusters* formados pelo DAMICORE, em substituição ao *K-means*, e compara-se os resultados com os obtidos em Lopes et al. [4]. Com isso aferiu-se a eficiência do MRA em rotular agrupamento por filogenias, alcançando acurácia superior a 90%.

O texto a seguir está organizado da seguinte maneira: na Seção 3 tem-se o referencial teórico com descrição dos métodos utilizados; na Seção 4 é descrita a forma como foram conduzidos os testes e são apresentados os resultados; na Seção 5 é feita uma comparação com os resultados de Lopes et al. [4]; por fim, na Seção 6 são apresentadas as conclusões obtidas.

2. TRABALHOS RELACIONADOS

Propostas de rotulação de dados baseadas em árvores de decisão como, por exemplo, $C4.5$ e $ID3$ [6], 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 [7] 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. [8]; Chuang e Chien [9]; Maqbool e Babri [10], caracterizam-se por tratar exclusivamente com informações textuais. O trabalho [11] 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 [11] 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 [3] 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. [12] 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 [13] também tem como objetivo a rotulação de textos, ao aplicar técnicas de Aprendizagem de Máquina (AM) para rotular siglas presentes no contexto da literatura biomédica. Outro exemplo é a abordagem proposta por Cuayáhuitl, Dethlefs e Hastie [14], 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 et al. [4], envolvendo a rotulação de *clusters* no que diz respeito a apresentar uma definição para se obter conhecimento em relação a atributos numéricos relevantes.

3 REFERENCIAL TEÓRICO

A seguir apresenta-se o funcionamento básico dos métodos DAMICORE e MRA, utilizados para *clustering* e rotulação automática de dados, respectivamente.

3.1 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 [15]. 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.

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 [2]. 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*) [16], 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

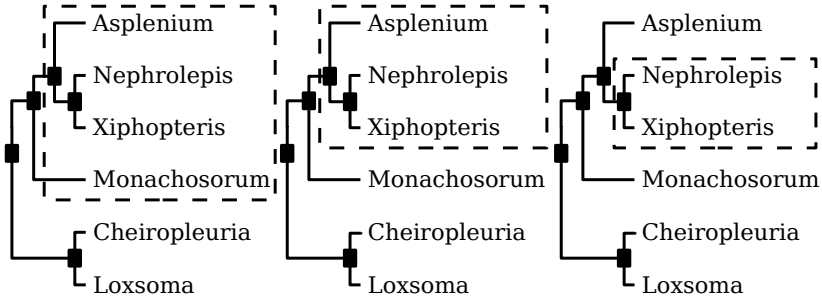


Figura 1: Possíveis clados (linhas tracejadas) que podem ser obtidos de uma mesma filogenia.

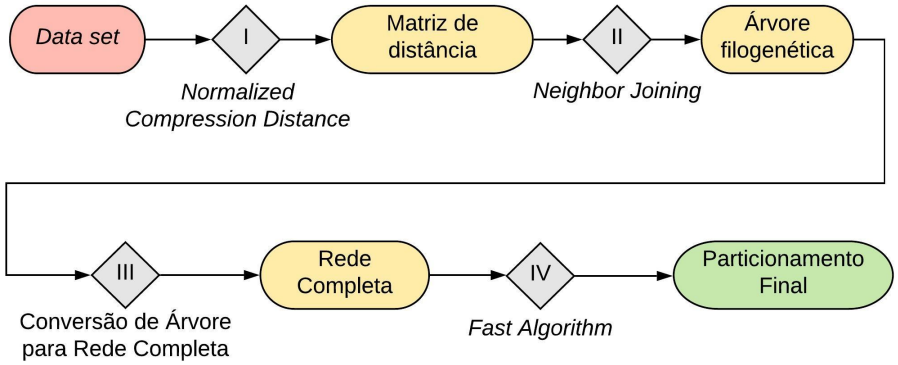


Figura 2: Diagrama resumindo o funcionamento do DAMICORE.

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*) [17], que considera a semelhança entre as variáveis de acordo com a característica dominante que elas compartilham. No entanto, a NID utiliza diretamente o conceito de complexidade de Kolmogorov [18] 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 ε é um limitante superior para o erro do compressor usado. O valor da distância entre X e Y é dado pela Equação 1:

$$D_{NCD}(X,Y) = \frac{C(XY) - min\{C(X), C(Y)\}}{max\{C(X), C(Y)\}}, \tag{1}$$

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. 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) [19] (que representa relações hierárquicas entre os indivíduos, uma vez que a estrutura de uma árvore é intrinsecamente hierárquica) usando o algoritmo NJ (do inglês, *Neighbor Joining*) [20]. 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.

Sobre a Matriz de Adjacências obtida é aplicado o FA (do inglês, *Fast Newman Algorithm*) [21], um algoritmo de detecção de estruturas de comunidades da área de Redes Complexas [22]. 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, que

[1] usualmente empregado por ferramentas de bioinformática.

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.

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

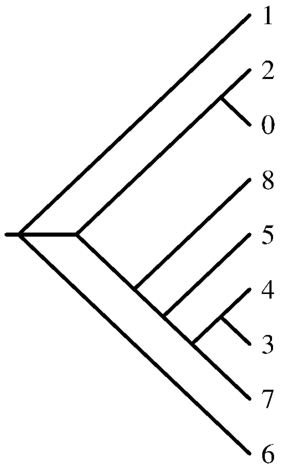


Figura 4: Árvore Filogenética reconstruída pelo NJ.

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.

3.2 ROTULAÇÃO AUTOMÁTICA DE *CLUSTERS*

Existem diversas pesquisas acerca do problema de clusterização, entretanto, poucas são as que focam em rotular os *clusters* resultantes. [23] aponta que, em 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 [23], 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. [24] e [25] 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. [4] define formalmente o problema de rotulação como segue.

Problema da rotulação. *Sendo C um conjunto de clusters definido como $C = \{c_1, \ldots, c_K | K \geqslant 1\}$, de tal modo que cada cluster contenha um conjunto de elementos $c_i = \{\vec{e}_1, \ldots, \vec{e}_{n^{(c_i)}} | n^{(c_i)} \geqslant 1\}$ 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;

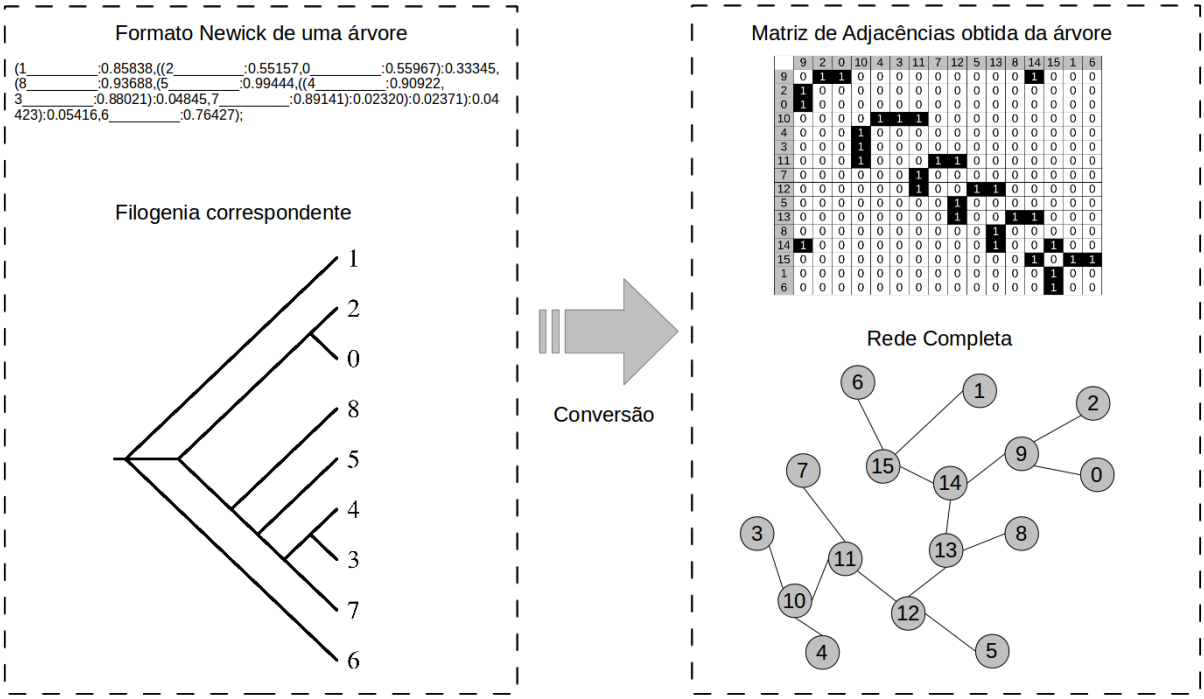


Figura 5: Conversão de filogenia no formato *Newick* para Matriz de Adjacências. Neste exemplo, nós com índices maiores que 8 são nós internos da filogenia.

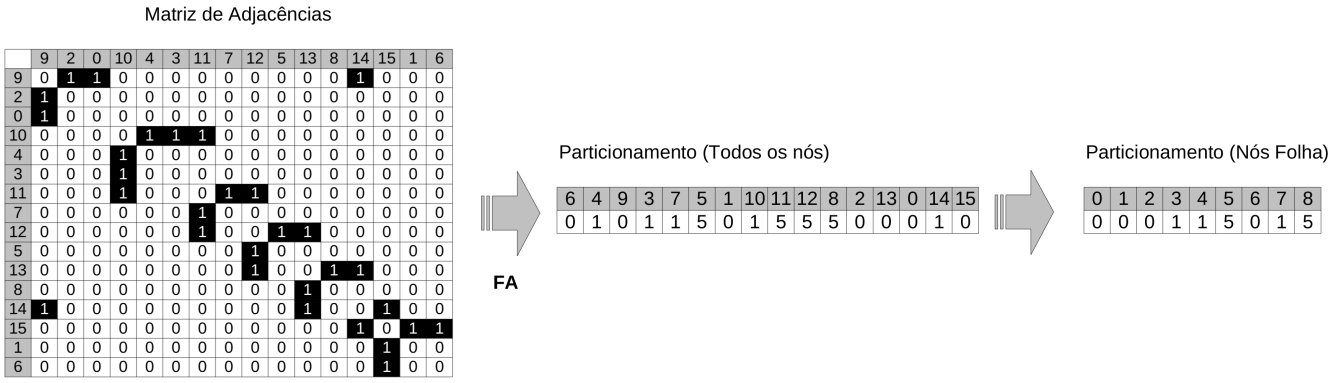


Figura 6: Particionamento Final gerado a partir da Matriz de Adjacências (Figura 5) usando o FA.

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.

3.3 ROTULAÇÃO AUTOMÁTICA DE *CLUSTERS* COM MRA

Lopes, Machado e Rabêlo [4] 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 7 mostra todas as etapas do método proposto.

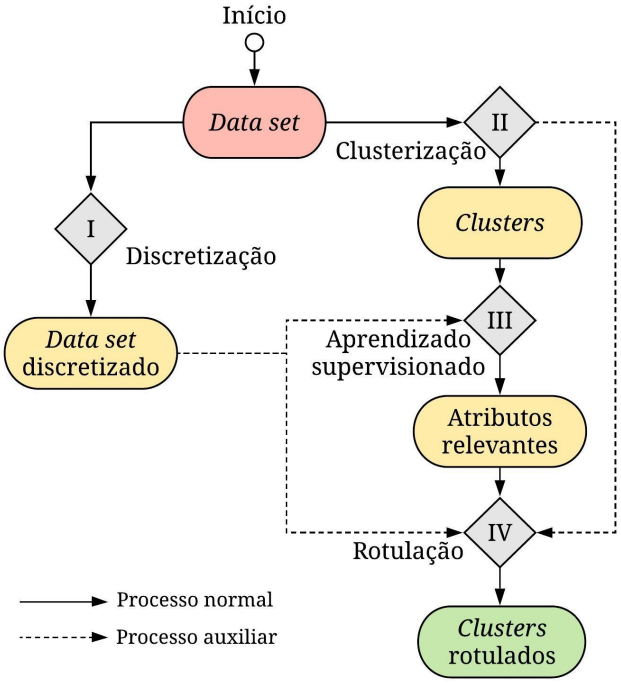


Figura 7: Modelo proposto por [4].

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 7. 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 8.

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

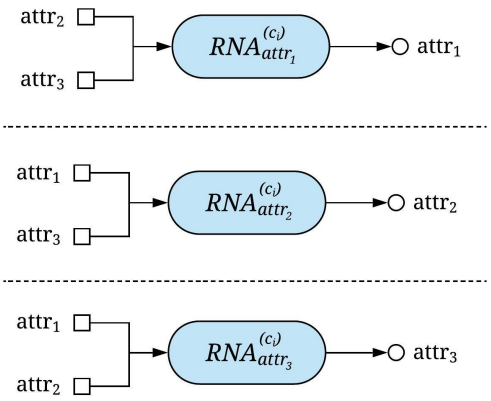


Figura 8: RNAs para seleção de atributos de um *cluster*.

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.

4. 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 em conjunto com o MRA para rotular *clusters*, o que permite analisar as relações hierárquicas existentes entre esses. A Figura 9 resume as etapas 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 [26], [27] e [28]: 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.

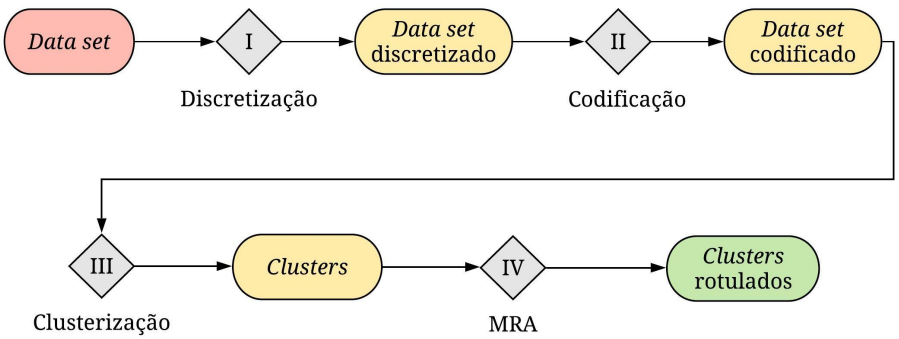
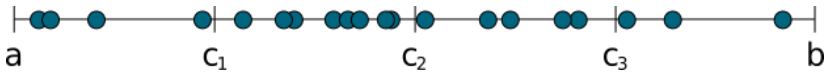


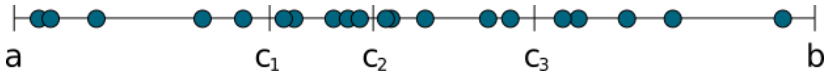
Figura 9: Fluxograma do método proposto.

As Figuras 10(a) e 10(b) 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 10(a), representa os pontos de corte calculados de forma que todas as faixas de valores tenham a mesma largura. Já a Figura 10(b) 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).



(a) Discretização EWD.



(b) Discretização EFD.

Figura 10: Exemplificação do uso dos métodos de discretização EWD e EFD sobre um mesmo conjunto de dados.

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 previamente conhecidas. A Figura 11 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 5.2.

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 11: 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 medir a similaridade dos elementos com maior precisão, contribuindo para a obtenção de agrupamentos mais significantes. A Figura 12 ilustra o resultado dessa Etapa II.

O *data set* codificado é então submetido ao DAMICORE para realização da clusterização, que corresponde à Etapa III da Figura 9. É obtida ao final 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 13.

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

Codificação →

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;PIIT	.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 12: Resultado da aplicação da etapa de codificação sobre subconjunto do *data set Iris* discretizado.

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;PIIT	.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

Clusterização →

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 13: Aplicação da etapa de agrupamento utilizando DAMICORE ou DAMICORE-2 aos dados codificados.

Finalmente, os *clusters* são submetidos ao MRA, que fornece um rótulo para cada um deles, etapa cujo resultado pode ser conferido na Figura 14.

A1	A2	A3	A4	*Cluster*
5,1	3,5	1,4	0,2	0
4,9	3,0	1,4	0,2	0
4,7	3,2	1,3	0,2	0
4,6	3,1	1,5	0,2	0
5,0	3,6	1,4	0,2	0
5,4	3,9	1,7	0,4	0
4,6	3,4	1,4	0,3	0
5,0	3,4	1,5	0,2	0
4,4	2,9	1,4	0,2	0
4,9	3,1	1,5	0,1	0
7,0	3,2	4,7	1,4	4
6,4	3,2	4,5	1,5	4
6,9	3,1	4,9	1,5	4
5,5	2,3	4,0	1,3	4
6,5	2,8	4,6	1,5	4
5,7	2,8	4,5	1,3	4
6,3	3,3	4,7	1,6	4
4,9	2,4	3,3	1,0	5
6,6	2,9	4,6	1,3	5
5,2	2,7	3,9	1,4	5
6,3	3,3	6,0	2,5	10
5,8	2,7	5,1	1,9	10
7,1	3,0	5,9	2,1	10
6,3	2,9	5,6	1,8	10
6,5	3,0	5,8	2,2	11
7,6	3,0	6,6	2,1	11
4,9	2,5	4,5	1,7	11
7,3	2,9	6,3	1,8	11
6,7	2,5	5,8	1,8	11
7,2	3,6	6,1	2,5	11

Rotulação (MRA)

Cluster	Rótulo	
	Atributo	Intervalo de valores
0	A3	1 ~ 1.7
	A1	2.8 ~ 3.2
4	A1	5.6 ~ 6.3
5	A3	5.5 ~ 6.9
10	A3	1 ~ 1.7
11	A4	1.1 ~ 1.5
	A3	3.9 ~ 4.7

Figura 14: Resultado final do MRA com os rótulos definidos para os *clusters* resultantes do agrupamento.

5. RESULTADOS

Para a realização dos experimentos foram utilizados cinco *data sets* obtidos do *UCI Machine Learning*[2]. 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.

As Tabelas 1 a 3 apresentam os resultados obtidos pela aplicação do MRA sobre os *clusters* formados pelo DAMICORE para os 3 *data sets* utilizados. Ressalta-se que são apresentados somente os melhores resultados em relação ao método de discretização (EWD ou EFD), sendo a quantidade de faixas utilizada a mesma apontada pela literatura para cada *data set*. Além disso os valores atribuídos à variação v em cada experimento foram os mesmos utilizados por Lopes et al. [4].

A leitura das Tabelas 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 taxa de acerto (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.

5.1 *Glass* - Identificação de Vidros

O *data set Glass*[3] se refere à identificação de vidros. Ele 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*, *Kg*, *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 processamento [29].

A Tabela 1 apresenta os resultados da rotulação obtidos em relação ao *data set Glass* utilizando o método de discretização EFD. O DAMICORE organizou os elementos em 22 *clusters*. Entretanto é possível visualizar que os *clusters* 1 e 4 possuem o mesmo rótulo, que se repete ainda em outros *clusters* não visíveis na Tabela 1. Isto ocorre pelo fato de o DAMICORE subdividir um *cluster* maior em *clusters* menores, devido à natureza hierárquica da reconstrução de filogenias inerente à estrutura de árvore.

[2]http://archive.ics.uci.edu/ml/index.php
[3]http://archive.ics.uci.edu/ml/datasets/Glass+Identification

Tabela 1: Rótulos e taxas de acerto obtidos pelo MRA aplicado aos *clusters* formados pelo DAMICORE no *data set Glass*.

Cluster	**# Elem.**	**Rótulos**			**Análise**	
		Atrib.	**Rel. (%)**	**Intervalo**	**# Erros**	**Acertos (%)**
1	15	Ba	100	$0 \sim 0{,}15$	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	9	Ba	100	$0 \sim 0{,}15$	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
14	14	IR	66,67	$1{,}5202 \sim 1{,}5339$	1	92,86
		Si	66,67	$71{,}96 \sim 72{,}48$	1	92,86
		Ba	72,22	$0 \sim 0{,}15$	2	85,71
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	13	Mg	100	$0 \sim 2{,}39$	0	100
22	9	K	100	$0 \sim 0{,}13$	0	100
		Mg	100	$0 \sim 2{,}39$	0	100
		Al	100	$1{,}94 \sim 3.5$	0	100

Em vários dos demais grupos vê-se o atributo *Ba* com a mesma faixa de valores ($0 \sim 0{,}15$). Porém, quando combinado com outros atributos considerados relevantes (e suas faixas de valores), novos rótulos são formados, evidenciando características diferentes entre *clusters*, de fato, distintos.

5.2 *Iris* - 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*[4], e foi apresentado em Fisher [30]. Os 150 elementos do *data set* são descritos por 4 características cujos valores são contínuos: comprimento da pétala (*PL*), largura da pétala (*PW*), comprimento da sépala (*SL*) e largura da sépala (*SW*). Os resultados obtidos são apresentados na Tabela 2. Neste caso o método de discretização EWD apresentou os melhores resultados.

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

Cluster	**# Elem.**	**Rótulos**			**Análise**	
		Atrib.	**Rel. (%)**	**Intervalo**	**# Erros**	**Acertos (%)**
1	12	PW	100	$0.1 \sim 0.4$	0	100
		SL	93,33	$4.9 \sim 5.6$	1	91,67
2	11	PW	100	$0.1 \sim 0.4$	0	100
		SL	86,67	$4.3 \sim 4.9$	2	81,82
		PL	100	$1 \sim 1.7$	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	8	PW	100	$1.1 \sim 1.5$	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
13	8	PW	83,33	$1.1 \sim 1.5$	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	7	PW	100	$1.9 \sim 2.5$	0	100
		PL	100	$6.3 \sim 7$	0	100

Para este *data set* o DAMICORE definiu um total de 18 *clusters*. Mais uma vez observa-se a presença de *clusters* com rótulos iguais (grupos 10 e 13), além de outros que não estão visíveis na Tabela 2, reafirmando a natureza hierárquica do agrupamento. Os demais *clusters* exibidos apresentam rótulos distintos, caracterizando-os individualmente.

5.3 *Seeds* - Identificação de Sementes

O último *data set* utilizado nos experimento é o *Seeds*[5], o qual se refere à identificação de sementes de plantas, tendo sido apresentado em Kulczycki e Charytanowicz [31]. Este conjunto de dados é composto por 210 amostras de 3 tipos de sementes de trigo, sendo 70 amostras de cada tipo. Os elementos são descritos por 7 atributos que representam as características geométricas

[4]http://archive.ics.uci.edu/ml/datasets/Iris
[5]http://archive.ics.uci.edu/ml/datasets/seeds

das sementes: área, perímetro, densidade, comprimento da semente (*LK*), largura da semente (*WK*), coeficiente de assimetria (*AC*) e comprimento do sulco da semente (*LKG*). Os resultados obtidos para este *data set*, utilizando o método de discretização EFD (que forneceu os melhores resultados), são exibidos na Tabela 3.

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

Cluster	**# Elem.**	**Rótulos**			**Análise**	
		Atrib.	**Rel. (%)**	**Intervalo**	**# Erros**	**Acertos (%)**
1	13	WK	72,22	3,073 ∼ 3,337	2	84,62
		Área	66,67	13,37 ∼ 15,11	5	61,54
⋮	⋮	⋮	⋮	⋮	⋮	⋮
12	7	WK	100	3,073 ∼ 3,337	0	100
		Área	100	12,05 ∼ 13,37	0	100
		Perímetro	100	13,31 ∼ 14,02	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	13	AC	72,22	4,933 ∼ 8,456	2	84,62
		Área	66,67	10,59 ∼ 12,05	2	84,62
		LKG	66,67	4,805 ∼ 5132	2	84,62
		Perímetro	66,67	12,41 ∼ 13,31	3	76,92

O DAMICORE determinou um total de 23 *clusters* para este *data set*. O número de *clusters* formado foi mais de 7 vezes superior ao apontado pela literatura (3). Novamente houve repetição nos rótulos atribuídos aos *clusters*, o que mostra que na verdade são *subclusters* de um *cluster* maior. A taxa de acerto dos rótulos definidos novamente foi superior à obtida com o agrupamento do *K-means*.

A Figura 15 apresenta um gráfico comparativo entre as taxas de acerto da rotulação resultantes da aplicação do MRA aos agrupamentos realizados pelo *K-means* e pelo DAMICORE sobre os 3 *data sets* citados anteriormente. Os valores fazem referência à taxa de acerto total, em que o número de acertos corresponde ao total de elementos que se enquadram nas faixas de valores de todos os atributos apontados como relevantes para determinado *cluster*.

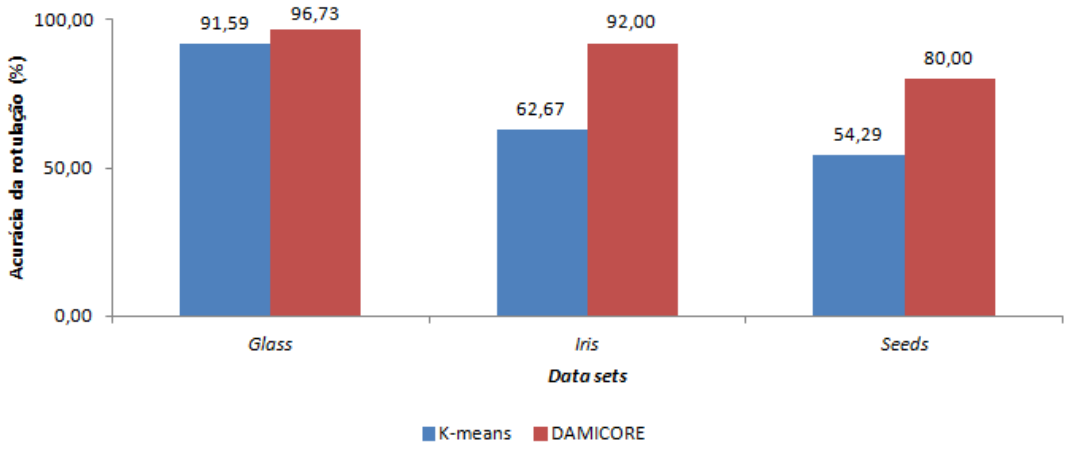


Figura 15: Comparação entre as taxas de acerto do MRA aplicado aos *clustering*s realizados pelo *K-means* e pelo DAMICORE.

A taxa de acerto do MRA para os agrupamentos do DAMICORE foi superior nos 3 casos, pois o método consegue expressar melhor a similaridade entre os elementos de cada *cluster*. Para o *data sets Glass* a diferença em relação à rotulação dos *clusters* formados pelo *K-means* foi de apenas 5,14%.

Para a rotulação do agrupamento do *data set Seeds*, embora a taxa de acerto tenha sido de apenas 80,00%, ela está 25,71 pontos percentuais acima da obtida ao se rotular o agrupamento do *K-means* – que foi apenas 54,29%. O resultado que mais se destacou foi o obtido na rotulação do *data set Iris*, onde a taxa de acerto com o agrupamento do DAMICORE (92,00%) ficou 29,33% acima da obtida com o agrupamento do *K-means*, que foi de apenas 62,67%.

Como dito anteriormente, o MRA alcançou acurácia superior para os 3 testes realizados, usando os *clusters* obtidos pelo DAMICORE, se destacando uma diferença mais significante nas acurácias com os *data sets* com menor número de elementos, o que acarreta em uma maior uniformidade dos valores. Além disso, os *data sets Iris* e *Seeds* têm como característica, a distinção acentuada entre os valores dos atributos para elementos em classes diferentes. Ainda a obtenção de rótulos idênticos para *clusters* distintos revela a existência de sub-*clusters* na árvore filogenética utilizada pelo DAMICORE.

6. CONCLUSÃO

O Método de Rotulação Automática (MRA) foi utilizado para rotular *clusters* obtidos pelo DAMICORE. Os resultados foram comparados com os apresentados em Lopes et al. [4], ao se realizar 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 possuem maior acurácia em comparação à alcançada pela aplicação sobre o agrupamento do *K-means*.

A eficácia da rotulação obtida para o agrupamento do DAMICORE é atribuída à quantidade de *clusters* resultantes. Com um maior número de *clusters* ocorre uma maior especificidade das características de seus respectivos elementos, devido ao menor grau de generalização. Reforça-se ainda o fato de que a técnica de agrupamento é determinante para a qualidade dos rótulos atribuídos pelo MRA, pois quanto maior for a semelhança intra-grupo maior será a acurácia dos rótulos encontrados.

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.

REFERÊNCIAS

[1] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. "From data mining to knowledge discovery in databases". In *Advances in Knowledge Discovery and Data Mining*, pp. 37–54. AAAI Press, 1996.

[2] A. Sanches, J. Cardoso and A. Delbem. "Identifying merge-beneficial software kernels for hardware implementation". In *Reconfigurable Computing and FPGAs (ReConFig)*, pp. 74–79. AAAI Press, 2011.

[3] H. Anaya-Sánches, A. Pons-Porrata and R. Berlanga-Llavori. "A New Document Clustering Algorithm for Topic Discovering and Labeling". In *13th Iberoamerican Congress on Pattern Recognition - CIARP 2008*, pp. 161–168. LNCS, 2008.

[4] A. Lopes, V. Machado and R. Rabêlo. "Automatic labelling of clusters of discrete and continuous data with supervised machine learning". In *Knowledge-Based Systems*, pp. 231–241. LNCS, 2016.

[5] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probabilitys*, pp. 281–297. University of California Press, 1967.

[6] J. R. Quinlan. "Induction of decision trees". , no. 81-106, 1986.

[7] M. Sanderson and B. Croft. "Deriving Concept Hierarchies from Text". In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pp. 206–213, New York, NY, USA, 1999. ACM.

[8] E. Glover, D. M. Pennock, S. Lawrence and R. Krovetz. "Inferring hierarchical descriptions". *Proceedings of the Eleventh International Conference on Information and Knowledge Management - CIKM*, , no. 507-514, 2002.

[9] S.-L. Chuang and L.-F. Chien. "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*, , no. 127-136, 2004.

[10] O. Maqbool and H. Babri. "Interpreting clustering results through cluster labeling". *Proceedings of the IEEE Symposium on Emerging Technologies*, , no. 429-434, 2005.

[11] T. Treeratpituk and J. Callan. "Automatically labeling hierarchical clusters". *Proceedings of the 2006 International Conference on Digital Government Research*, , no. 167-176, 2006.

[12] C. J. Solana-Cipres, J. Albusac, J. J. Castro-Schez and L. Rodriguez-Benitez. "Automatic Object Labelling for Monitored Environments Using Clustering Techniques". *3rd International Conference on Crime Detection and Prevention (ICDP)*, 2009.

[13] L. Yeganova, D. C. Comeau and W. J. Wilbur. "Identifying Abbreviation Definitions". *Ninth International Conference on Machine Learning and Applications*, , no. 500-505, 2010.

[14] H. Cuayáhuitl, N. Dethlefs and H. Hastie. "A Semi-Supervised Clustering Approach for Semantic Slot Labelling". *13th International Conference on Machine Learning and Applications*, , no. 500-505, 2014.

[15] A. Soares, R. Rabêlo and A. Delbem. "Optimization based on phylogram analysis". *Expert Systems with Applications*, vol. 78, pp. 32 – 50, 2017.

[16] R. Cilibrasi and P. Vitányi. "Clustering by compression". In *IEEE Transactions on Information Theory*, pp. 1523–1545. University of California Press, 2005.

[17] J. Lillo-Castellano, I. Mora-Jiménez, R. Santiago-Mozos, J. Rojo-Álvarez, J. R.-B. no and A. Algora-Weber. "Weaning outcome prediction from heterogeneous time series using Normalized Compression Distance and Multidimensional Scaling". *Expert Systems with Applications*, vol. 40, no. 5, pp. 1737 – 1747, 2013.

[18] M. Li and P. M. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag New York, Inc., second edition, 1997.

[19] W. Cancino and A. Delbem. "Inferring phylogenies by multi-objective evolutionary algorithm". In *International journal of information technology and intelligent computing*, pp. 1–26, 2007.

[20] N. Saitou and M. Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees". In *Molecular Biology and Evolution*, pp. 406–425, 1987.

[21] M. Newman and M. Girvan. "Finding and evaluating community structure in networks". In *Physical Review E*, pp. 406–425, 2004.

[22] J. Duch and A. Arenas. "Community detection in complex networks using extremal optimization". In *Physical Review E*, pp. 406–425, 2005.

[23] V. Tzerpos. "Comprehension-Drive Software Clustering", 2001.

[24] H.-L. Chen, K.-T. Chuang and M.-S. Chen. "On data labeling for clustering categorical data". *Knowledge and Data Engineering, IEEE Transactions*, , no. 1458-1472, 2008.

[25] T. Eltoft and R. deFigueiredo. "A self-organizing neural network for cluster detection and labeling". *IEEE International Joint Conference on Neural Networks Proceedings*, vol. 1, no. 408-412, 1998.

[26] S. Kotsiantis and D. Kanellopoulos. "Discretization techniques: A recent survey". In *GESTS International Transactions on Computer Science and Engineering*, pp. 47–58, 2006.

[27] J. Cerquides and R. L. de Màntaras. "Proposal and empirical comparison of a parallelizable distance-based discretization method". In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 139–142, 1997.

[28] J. Dougherty, R. Kohavi and M. Sahami. "Supervised and unsupervised discretization of continuous features". In *12nd International Conference on Machine Learning - ICML*, pp. 194–202. Morgan Kaufmann, 1995.

[29] I. Evett and E. Spiehler. "Rule induction in forensic science". In *Knowledge Based Systems*, pp. 152–160. Halsted Press, 1988.

[30] R. Fisher. "The use of multiple measurements in taxonomic problems". In *Annals of Eugenics*, pp. 17–188, 1936.

[31] P. Kulczycki and M. Charytanowicz. "A complete gradient clustering algorithm". In *Proceedings of the Third International Conference on Artificial Intelligence and Computational Intelligence*, pp. 497–504. Springer-Verlag, 2011.