Rotulação Automática de *Clusters* Baseados em Análise de Filogenias

Francisco N. C. de Araújo[1], Antonio H. M. Soares[1], Vinicius P. Machado[1], Ricardo de A. L. Rabêlo[1]

[1]Departamento de Computação – Universidade Federal do Piauí (UFPI)
Teresina – PI – Brasil

{netoaraujjo, ahelsonms}@gmail.com, {vinicius,ricardoalr}@ufpi.edu.br

Abstract. *This paper proposes the joint use of unsupervised and supervised Machine Learning methods for data clustering and labeling tasks, respectively. 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.*

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. 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.*

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 [Fayyad et al. 1996].

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*) [Sanches et al. 2011] mostrou ser capaz de encontrar correlações em bases de dados de tipos mistos (registros com 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 [Anaya-Sánches et al. 2008]. 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. [Lopes et al. 2016] 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. [Lopes et al. 2016] obteve resultados positivos, conseguindo rotular *clusters* com taxa de acerto média de 85%, aos ser aplicada em agrupamentos realizados pelo algoritmo K-means [MacQueen 1967].

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) [Lopes et al. 2016] para rotular os *clusters* formados pelo DAMICORE, em substituição ao K-means, e compara-se os resultados com os obtidos em Lopes et al. [Lopes et al. 2016]. 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 2 tem-se o referencial teórico com descrição dos métodos utilizados; na Seção 3 é descrita a forma como foram conduzidos os testes e são apresentados os resultados; na Seção 4 é feita uma comparação com os resultados de [Lopes et al. 2016]; por fim, na Seção 5 são apresentadas as conclusões obtidas.

2. Referencial Teórico

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

2.1. DAMICORE

O DAMICORE une algoritmos largamente utilizados na literatura, produzindo resultados eficientes como demonstrado em Sanches et al. [Sanches et al. 2011]. O método faz uso de um conjunto de técnicas de várias áreas do conhecimento (Teoria da Computação, Bioinformática e Física) de forma a extrair informações através de uma métrica universal e robusta. O DAMICORE surge como um método de identificação de correlação entre tipos de dados diversos, procedimento relativamente complexo para a maioria dos algoritmos de agrupamento. Além disso, não é necessário informar ao algoritmo a quantidade de *clusters* na qual os elementos devem ser alocados.

A execução do DAMICORE é iniciada pelo cálculo da Matriz de Distância usando a NCD (*Normalized Compression Distance*) [Cilibrasi e Vitányi 2005], a qual calcula uma razão de distância entre os dados determinando a semelhança entre os valores das variáveis com base nos tamanhos de seus dados compactados. A NCD tem sido aplicada com sucesso em áreas como a genética, literatura, música e astronomia. Essa abordagem não requer nenhum conhecimento específico do domínio da aplicação.

A partir da Matriz de Distância é reconstruída uma Árvore Filogenética [Cancino e Delbem 2007] (que pode representar relações hierárquicas entre os in-

divíduos) usando o algoritmo NJ (*Neighbor Joining*) [Saitou e Nei 1987]. A saída do NJ é, por sua vez, convertida do formato Newick[1] para o formato de Matriz de Adjacências.

Na Matriz de Adjacências obtida é aplicado o FA (*Fast Newman Algorithm*) [Newman e Girvan 2004], um algoritmo de detecção de estruturas de comunidades da área de Redes Complexas [Duch e 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, que representam as variáveis do problema.

2.2. Rotulação Automática de *Clusters* com MRA

Lopes et al. [Lopes et al. 2016] propõe a utilização de Redes Neurais Artificiais (RNA) do tipo Perceptron Multi-Camadas para a obtenção de um rótulo para cada *grupo*, que melhor o define. Para cada atributo dos elementos de um dado *cluster* é criada uma RNA. Estas 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 *cluster* trabalham com os mesmos elementos variando somente a maneira como estes elementos são utilizados – entrada ou saída.

Cada RNA é criada de forma a representar e avaliar a importância de um atributo em relação aos demais, para cada *cluster*. 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 de cada *cluster* com as maiores taxas de acerto são definidos como rótulo.

Um parâmetro de variação v é utilizado para reduzir ambiguidades entre rótulos de diferentes *clusters*, 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 *cluster* tenha sido de 95%, todas as RNAs com taxa de acerto a partir de 90% serão selecionadas para compor o rótulo.

3. Experimentos e Resultados

Para a realização dos experimentos foram utilizados os mesmos *data sets* usados por Lopes et al. em [Lopes et al. 2016] para rotulação de *clusters* criados pelo algoritmo K-means (*Glass*, *Iris* e *Seeds*). A Figura 1 resume as etapas da metodologia utilizada. Antes de serem agrupados os *data sets* foram submetidos a um pré-processamento, composto pelas fases de discretização (I) e codificação (II).

Os atributos cujos valores eram contínuos foram discretizados utilizando dois métodos: EWD (*Equal Width Discretization*), no qual o intervalo de valores assumidos pelo atributo é dividido em faixas de larguras iguais; e o EFD (*Equal Frequency Discretization*), que divide o intervalo de valores do atributo de forma a alocar a mesma quantidade de elementos em cada faixa resultante.

Em seguida, os *data sets* foram submetidos a uma fase de codificação na qual os valores de atributos numéricos são substituídos por códigos alfanuméricos. Essa fase visa

[1] usualmente empregado por ferramentas de bioinformática.

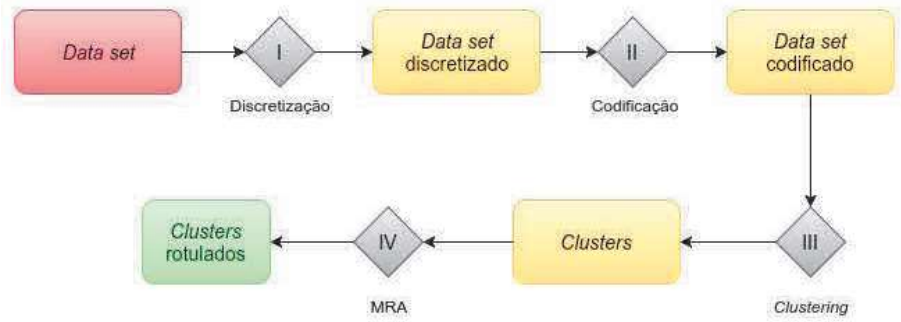

Figura 1. Etapas do método proposto.

reforçar a diferença entre valores como, por exemplo, 1 e 11 – que por vezes são considerados mais próximo que 1 e 2. O *data set* codificado é então submetido ao DAMICORE para realização do agrupamento (III). Obtém-se ao final do processo, uma lista contendo o índice de cada elemento seguido por um número inteiro representando o *cluster* ao qual o elemento foi alocado. Esses valores são adicionados ao *data set* original e então submetidos ao MRA (IV), que fornecendo um rótulo para cada *cluster*.

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. [Lopes et al. 2016].

3.1. *Glass* - Identificação de Vidros

O *data set Glass* se refere à identificação de vidros e pode ser encontrado no repositório de dados *UCI Machine Learning*[2]. O *data set* 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 [Evett e Spiehler 1988].

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.

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.

[2]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		Rel. (%)	Análise	
		Atrib.	Intervalo		# Erros	Acertos (%)
1	15	Ba	$0 \sim 0{,}15$	100	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	9	Ba	$0 \sim 0{,}15$	100	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
14	14	IR	$1{,}5202 \sim 1{,}5339$	66,67	1	92,86
		Si	$71{,}96 \sim 72{,}48$	66,67	1	92,86
		Ba	$0 \sim 0{,}15$	72,22	2	85,71
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	13	Mg	$0 \sim 2{,}39$	100	0	100
22	9	K	$0 \sim 0{,}13$	100	0	100
		Mg	$0 \sim 2{,}39$	100	0	100
		Al	$1{,}94 \sim 3{,}5$	100	0	100

3.2. *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*. É apresentado em Fisher [Fisher 1936], podendo ser encontrado no repositório de dados *UCI Machine Learning*[3].

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.

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.

3.3. *Seeds* - Identificação de Sementes

O último *data set* utilizado nos experimento se refere à identificação de sementes de plantas e – assim como os *data sets Glass e Iris* – pode ser encontrada no repositório de dados *UCI Machine Learning*[4] *Seeds*, tendo sido apresentado em Kulczycki e Charytanowicz [Kulczycki e Charytanowicz 2011].

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 das sementes: área, perímetro, densidade, comprimento da semente (*LK*), largura da semente (*WK*), coeficiente de assimetria (*AC*) e

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

Tabela 2. 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	Acertos (%)
1	12	PW	0,1 ∼ 0,4	100	0	100
		SL	4,0 ∼ 5,6	93,33	1	91,67
2	11	PW	0,1 ∼ 0,4	100	0	100
		SL	4,3 ∼ 4,9	86,67	2	81,82
		PL	1 ∼ 1,7	100	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	8	PW	1,1 ∼ 1,5	100	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
13	8	PW	1,1 ∼ 1,5	83,33	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	7	PW	1,9 ∼ 2,5	100	0	100
		PL	6,3 ∼ 7	100	0	100

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		Rel. (%)	Análise	
		Atrib.	Intervalo		# Erros	Acertos (%)
1	13	WK	3,073 ∼ 3,337	72,22	2	84,62
		Área	13,37 ∼ 15,11	66,67	5	61,54
⋮	⋮	⋮	⋮	⋮	⋮	⋮
12	7	Wk	3,073∼ 3,337	100	0	100
		Área	12,05 ∼ 13,37	100	0	100
		Perímetro	13,31 ∼ 14,02	100	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	13	AC	4,933 ∼ 8,456	72,22	2	84,62
		Área	10,59 ∼ 12,05	66,67	2	84,62
		LKG	4,805 ∼ 5,132	66,67	2	84,62
		Perímetro	12,41 ∼ 13,31	66,67	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.

4. Discussão dos Resultados

A Figura 2 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*.

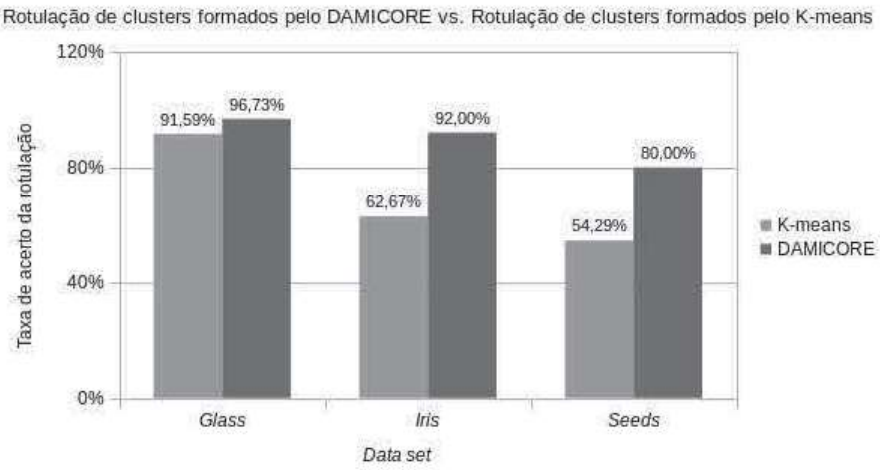

Figura 2. Comparação entre as taxas de acerto do MRA aplicado aos *clusterings* realizados pelo K-means e pelo DAMICORE.

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%.

5. Conclusão

O Método de Rotulação Automática (MRA) foi utilizado para rotular *clusters* formados pelo DAMICORE. Os resultados obtidos foram comparados com os apresentados em Lopes et al. [Lopes et al. 2016], ao se realizar a rotulação automática sobre *clusters* formados 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.

Referências

Anaya-Sánches, H., Pons-Porrata, A., e Berlanga-Llavori, R. (2008). A new document clustering algorithm for topic discovering and labeling. In *13th Iberoamerican Congress on Pattern Recognition - CIARP 2008*, pags 161–168. LNCS.

Cancino, W. e Delbem, A. (2007). Inferring phylogenies by multi-objective evolutionary algorithm. In *International journal of information technology and intelligent computing*, pags 1–26.

Cilibrasi, R. e Vitányi, P. (2005). Clustering by compression. In *IEEE Transactions on Information Theory*, pags 1523–1545. University of California Press.

Duch, J. e Arenas, A. (2005). Community detection in complex networks using extremal optimization. In *Physical Review E*, pags 406–425.

Evett, I. e Spiehler, E. (1988). Rule induction in forensic science. In *Knowledge based systems*, pags 152–160. Halsted Press.

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*, pags 37–54. AAAI Press.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. In *Annals of Eugenics*, pags 17–188.

Kulczycki, P. e Charytanowicz, M. (2011). A complete gradient clustering algorithm. In *Proceedings of the Third International Conference on Artificial Intelligence and Computational Intelligence*, pags 497–504. Springer-Verlag.

Lopes, A., Machado, V., e Rabêlo, R. (2016). Automatic labelling of clusters of discrete and continuous data with supervised machine learning. In *Knowledge-Based Systems*, pags 231–241. LNCS.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probabilitys*, pags 281–297. University of California Press.

Newman, M. e Girvan, M. (2004). Finding and evaluating community structure in networks. In *Physical Review E*, pags 406–425.

Saitou, N. e Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. In *Molecular Biology and Evolution*, pags 406–425.

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