Algoritmos Aglomerativos de Agrupamento Baseados em Teoria de Matrizes

Rodrigo Costa Camargos, Maria do Carmo Nicoletti
Faculdade de Campo Limpo Paulista - FACCAMP
Campo Limpo Paulista - SP

{rodrigocamargos@hotmail.com,carmo@cc.faccamp.br}

Abstract. *Hierarchical clustering algorithms (HC) produce a hierarchy of nested clustering, organized as a hierarchical tree. The so called agglomerative clustering algorithms (AC) can be approached as a particular category of HC, where the clustering process operates bottom-up. The main goal of this research is to identify data characteristics which promote a good performance of AC algorithms based on the matrix theory. The paper describes the ongoing work and main steps aiming at achieving such goal.*

Resumo. *Algoritmos de agrupamento hierárquicos (AH) produzem uma hierarquia de agrupamentos aninhados, organizados como uma árvore hierárquica. Os chamados algoritmos de agrupamento aglomerativos (AA) podem ser abordados como uma categoria particular de AH, em que o processo de agrupamento é feito de maneira bottom-up. O principal objetivo da pesquisa é identificar características de padrões de dados que promovem um bom desempenho de algoritmos AA baseados em teoria de matrizes. O artigo descreve o trabalho sendo conduzido e os principais passos com vistas a atingir o objetivo proposto.*

1. Introdução

Um dos primeiros requisitos para a utilização de um sistema computacional que realiza aprendizado indutivo é dispor de um conjunto de dados, conhecido como *conjunto de treinamento*, que representa o conceito a ser aprendido. Cada dado (ou padrão) de um conjunto de treinamento é, geralmente, descrito por um vetor de atributos (i.e., um vetor de valores associados a atributos) e, dependendo da situação, de uma classe associada (que indica qual conceito o dado em questão representa). A classe de cada padrão do conjunto de treinamento é, na maioria dos casos, determinada por um especialista humano da área de conhecimento à qual pertencem os dados. O fato de a classe participar da descrição do padrão e do algoritmo de aprendizado fazer uso dessa informação caracteriza a técnica de aprendizado automático como de *aprendizado supervisionado* (ver [Mitchell 1997], [Witten *et al*., 2011]).

Em muitas situações do mundo real, entretanto, a classe à qual cada dado pertence é desconhecida e/ou não existe um especialista humano que, com base na descrição dos valores de atributos que descrevem os dados, seja capaz de determiná-la. Técnicas de aprendizado indutivo de máquina que lidam com conjuntos de dados que não têm uma classe associada são conhecidas como técnicas de *aprendizado não-supervisionado*. Uma das técnicas de aprendizado não-supervisionado mais populares é chamada de agrupamento (*clustering*). O objetivo principal de algoritmos de agrupamento é particionar o conjunto de padrões disponível em grupos, de acordo com as similaridades e dissimilaridades entre tais padrões. Como pode ser confirmado em pes-

quisa bibliográfica associada especificamente à agrupamentos, o número de algoritmos propostos na literatura é considerável (ver, por exemplo, [Theodoridis & Koutroumbas 2009], [Duda *et al.*, 2001], [Jain & Dubes 1988], [Jain 2010]).

Este artigo descreve os passos iniciais e as escolhas feitas até o momento, relativas ao projeto de pesquisa em desenvolvimento cujo foco é a investigação empírica de algoritmos de agrupamento caracterizados como hierárquicos, particularmente aqueles que se enquadram na subcategoria de *hierárquicos aglomerativos*. Algoritmos hierárquicos produzem uma hierarquia de agrupamentos aninhados; via de regra esses algoritmos envolvem N passos ou seja, tantos quantos forem os padrões disponibilizados. A cada passo t um novo agrupamento é produzido usando, para isso, o agrupamento produzido no passo anterior i.e., no passo t−1.

De interesse particular neste trabalho são os algoritmos de agrupamento aglomerativos (AA) baseados em conceitos da Teoria de Matrizes, com o objetivo (1) identificação das principais características de domínio de dados que promovem um bom desempenho deste tipo de algoritmo; (2) estudo entre as várias alternativas para o cálculo de distâncias entre conjuntos de padrões. Além desta seção inicial de introdução o artigo descreve, na Seção 2, alguns conceitos e suas formalizações, necessários para a compreensão do que segue. A Seção 3 apresenta brevemente o Esquema Aglomerativo Generalizado (EAG), algoritmo fundamental ao prosseguimento da pesquisa, discutindo sua adequação para o uso da teoria de matrizes. A Seção 4 finaliza esse artigo comentando sobre as próximas etapas do trabalho.

2. Principais Conceitos e Notação Empregada

Considere que o conjunto de N padrões a serem agrupados seja $X = \{P_1, P_2, ..., P_N\}$, em que cada padrão P_i, $1 \leq i \leq N$ é descrito por M atributos, $A_1, A_2, ..., A_M$. Um K-agrupamento de X é uma partição de X em K conjuntos (grupos), $G_1, G_2, ..., G_K$, ou seja, K-Agrupamento = $\{G_1, G_2, ..., G_K\}$. As três condições a seguir devem ser verificadas:

(1) $G_i \neq \emptyset$, i = 1, ..., K (cada um dos grupos é não-vazio)
(2) $\bigcup_{i=1}^{K} G_i = X$ (a união de todos os grupos recompõe o conjunto X original)
(3) $G_i \cap G_j = \emptyset$, $i \neq j$ e i, j = 1, ..., K (os grupos são dois-a-dois disjuntos)

Assume-se que os padrões que pertencem a cada um dos grupos G_i ($1 \leq i \leq K$), quando comparados entre si, são "mais semelhantes" do que quando comparados com padrões que pertencem a um outro grupo, que não o G_i [Jain *et al*. 1999]. O conceito de similaridade adotado para a implementação de algoritmos de agrupamento desempenha um papel altamente relevante no resultado obtido. Uma maneira de implementar o conceito de similaridade é por meio do uso de uma medida de distância definida no espaço de atributos (que descrevem os padrões); dois padrões P_i e P_j são considerados similares se estiverem 'perto' um do outro, em que 'estar perto' precisa também ser quantificado. Na presente etapa do projeto o 'estar perto' é quantificado por meio de uma medida de distância, no caso, a distância euclidiana.

Seja um espaço M-dimensional definido por M atributos (de dados) e dois padrões desse espaço, P_i e P_j, representados por $P_i = (P_{i_1}, P_{i_2}, ..., P_{i_M})$ e $P_j = (P_{j_1}, P_{j_2}, ...,$

P_{jM}) respectivamente. A distância euclidiana entre os dois padrões, P_i e P_j, é calculada pela Eq. (1).

$$d(P_i, P_j) = \sqrt{\sum_{k=1}^{M} \left(P_{i_k} - P_{j_k}\right)^2} \quad (1)$$

Considere novamente o conjunto de N padrões $X = \{P_1, P_2, ..., P_N\}$ e considere dois agrupamentos dos padrões de X, identificados por AG_1 e AG_2, respectivamente. O agrupamento AG_1, contendo k grupos, está aninhado no agrupamento AG_2 que contém r (< k) grupos, (notado por $AG_1 \langle AG_2$) se cada grupo em AG_1 for subconjunto de um conjunto de AG_2 e, pelo menos um grupo de AG_1 for um subconjunto próprio de um elemento de AG_2. Seja $X = \{P_1, P_2, P_3, P_4, P_5\}$. O agrupamento $A_1 = \{\{P_1,P_3\}, \{P_4\}, \{P_2,P_5\}\}$ está aninhado em $A_2 = \{\{P_1,P_3,P_4\}, \{P_2,P_5\}\}$. Entretanto, A_1 não está aninhado nem em $A_3 = \{\{P_1,P_4\}, \{P_3\}, \{P_2,P_5\}\}$ ou tampouco em $A_4 = \{\{P_1,P_2,P_4\}, \{P_3,P_5\}\}$.

A um conjunto de N padrões M-dimensionais $X = \{P_1, P_2, ..., P_N\}$, pode ser associada uma matriz de dimensão N×M, chamada *matriz de padrões*, cuja i-ésima linha representa o i-ésimo vetor de X. Uma outra matriz associada a X é chamada de *matriz de similaridade*, notada por *MS(X)*, que é uma matriz N×N em que seu elemento ms_{ij} representa a similaridade entre os padrões P_i e P_j, para i,j = 1, ..., N.

3. O Esquema Aglomerativo Generalizado (EAG)

Seja o conjunto de N padrões M-dimensionais $X = \{P_1, P_2, ..., P_N\}$ e considere todos os possíveis subconjuntos dois-a-dois disjuntos de X, $SC(X) = \{G_1, G_2, ..., G_h\}$. Seja $g(G_i,G_j)$ (i,j = 1, ..., h) uma função definida em SC(X)×SC(X), com valores reais, que mede a proximidade entre os subconjuntos G_i e G_j (i,j = 1, ..., h) e seja t o nível corrente do processo de obtenção de agrupamentos. A Figura 1 apresenta o pseudocódigo em alto nível do algoritmo EAG (em inglês *GAS – Generalized Agglomerative Scheme*), que cria uma hierarquia de N agrupamentos, de maneira que cada um está aninhado em todos os agrupamentos sucessivos ou seja, $AG_t \langle AG_s$, para t < s, s = 1, ..., N–1.

Algoritmos de AA baseados em Teoria de Matriz podem ser abordados como casos particulares do algoritmo EAG, tendo como input a matriz de dissimilaridade, $MD_0 = MD(X)$, construída a partir de X. Na posição (i, j) da matriz de dissimilaridade está o valor da dissimilaridade (distância euclidiana, por exemplo) entre os padrões P_i e P_j do conjunto X, para i, j = 1, ..., N. A cada nível t, quando dois grupos são unidos em apenas um, o tamanho da matriz de dissimilaridade MD_t se torna (N – t) × (N – t). MD_t é construída a partir de $MD_{t–1}$ por meio da (1) deleção das duas linhas e das duas colunas que correspondem aos grupos que foram unidos e (2) adição de uma nova linha e uma nova coluna, contendo as distâncias do novo grupo formado e os outros grupos do agrupamento. A distância entre o novo grupo formado pela união de dois grupos G_i e G_j i.e., $G_q = G_i \cup G_j$ e um grupo antigo G_s é uma função (f) como representada em Eq. (2), ou seja, depende das três distâncias entre os grupos: G_i e G_s, G_j e G_s e entre G_i e G_j. Para sua implementação é preciso primeiro fazer uma escolha de qual função utilizar para calcular a distância entre dois grupos de padrões (ver, as mais populares, em [Theodoridis & Koutroumbas 2009]).

$$d(G_q,G_s) = f(d(G_i,G_s), d(G_j,G_s), d(G_i,G_j)) \quad (2)$$

procedure EAG (X, AG_t)
Input: X = {P_1, P_2,...,P_N}
Output: AG_t
1. **begin**
2. $AG_0 \leftarrow \{\{P_1\}, \{P_2\},...,\{P_N\}\}$ % agrupamento inicial
3. Nro_G $\leftarrow$ N
4. t $\leftarrow$ 0
5. **repeat**
6. t $\leftarrow$ t + 1
7. entre todos os possíveis pares de grupos (G_r,G_s) em AG_{t-1}, encontrar (G_i,G_j) tal que:

$$g(G_i,G_j) = \begin{cases} \min_{r,s} g(G_r,G_s) \text{ se g for função de dissimilaridade} \\ \max_{r,s} g(G_r,G_s) \text{ se g for função de similaridade} \end{cases}$$

8. New_G $\leftarrow G_i \cup G_j$
9. Nro_G $\leftarrow$ Nro_G – 1
10. $G_{Nro\_G} \leftarrow$ New_G
11. $AG_t \leftarrow (AG_{t-1} - \{G_i,G_j\}) \cup \{G_{Nro\_G}\}$
12. **until** todos os padrões pertençam a um único grupo.
13. **end**
return AG_t
end procedure

Figura 1. Pseudocódigo do Esquema Aglomerativo Generalizado (EAG) de Agrupamento.

4. Comentários Finais e Próximas Etapas

Na continuação do trabalho o EAG adaptado ao uso de teoria de Matrizes será implementado e, em seguida, experimentos iniciais, envolvendo dados artificiais serão conduzidos para uma validação empírica do sistema computacional desenvolvido. Como brevemente discutido anteriormente, o cálculo da distância entre dois grupos (de um agrupamento) é crítico; a adoção de diferentes definições de distância entre grupos provoca diferentes versões do algoritmo. Na sequência de atividades serão investigadas abordagens para o cálculo da distância, particularmente: *single-link*, *complete-link*, *average* e centroide, procurando, sempre que possível, evidenciar vantagens e limitações no uso de cada uma delas relacionadas a determinados tipos de conjuntos de dados (que, particularmente possam ter diferentes formas de grupos e, também, diferente densidades de grupos) (ver, por exemplo, [Webb & Copsey 2011], [Witten *et al*., 2011] e [Kuncheva 2014]). Algumas propostas evidenciadas durante o levantamento bibliográfico serão estudadas em paralelo às atividades anteriores ([Kurita 1991], [Eriksson *et al.* 2001], [Müller 2011] e [Krishnamurthy *et al.* 2012]) com vistas a ampliar o escopo de conhecimento sobre métodos de agrupamento aglomerativos. Ao final do trabalho de pesquisa o projeto contempla a disponibilização de um ambiente computacional para a experimentação dos algoritmos implementados em domínios de dados com diferentes caracterizações (volume, número, formato, densidades, etc.).

Referências

Duda, R. O.; Hart, P. F.; Stork, D. G. (2001) *Pattern Classification*, USA: John Wiley & Sons, Inc.

Eriksson, B.; Dasarathy, G.; Singh, A.; Nowak, R. (2001) Active clustering: robust and efficient hierarchical clustering using adaptively selected similarities, In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), v. 15 of JMLR: W&CP 15, pp. 260-268.

Jain, A.K., Dubes, R.C. (1988) *Algorithms for Clustering Data,* Prentice Hall.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, (31), n.3, pp. 264-323.

Jain, A.K. (2010) Data clustering: 50 years beyond K-Means, Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666.

Krishnamurthy, A.; Balakrishnan, S.; Xu, M.; Singh, A. (2012) Efficient active algorithms for hierarchical clustering, In: Proceedings of the 29th International Conference on Machine Learning, Scotland, UK, 2012.

Kuncheva L. I. (2014) *Combining Pattern Classifiers, Methods and Algorithms,* 2nd ed. USA: John Wiley & Sons, Inc.

Kurita, T. (1991) An efficient agglomerative clustering algorithm using a heap, Pattern Recognition, v. 24, pp. 777-783.

Mitchell, T. M (1997*) Machine Learning*, USA: McGraw-Hill.

Müller, D. (2011) Modern hierarchical, agglomerative clustering algorithms, arXiv:1109.2378v1.

Theodoridis, S.; Koutroumbas, K. (2009) *Pattern Recognition,* 4th ed. USA: Academic Press.

Webb, A. R.; Copsey, K. D. (2011) *Statistical Pattern Recognition*, 3th ed. USA: John Wiley & Sons, Inc.

Witten, I. H.; Frank E.; Hall, M. A. (2011) *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann.