Clustering hierárquico: uma metodologia para auxiliar na interpretação dos clusters

Jean Metz*, Maria Carolina Monard

[1]Laboratório de Inteligência Computacional — Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Av. Trabalhador São-carlense, 400 — 13560-970 São Carlos, SP

{metzz, mcmonard}@icmc.usp.br

Abstract. *There is a growing interest in information analysis methods in order to extract knowledge from databases. In this context, hierarchical clustering can be used to analyse data at different levels of details. This work presents a methodology to analyse clusters generated at each level of a hierarchy, aiming to increase the understanding and interpretation of these clusters.*

Resumo. *Existe um crescente interesse em métodos de análise de informações para extração de conhecimento de conjuntos de dados. Nesse contexto, clustering hierárquico pode ser utilizado para analisar dados em diferentes níveis de detalhes. Neste trabalho é apresentada uma metodologia para análise de clusters em cada nível da hierarquia, visando aumentar o entendimento e facilitar a interpretação desses clusters.*

1. Introdução

Motivadas pela grande disponibilidade de recursos computacionais e a facilidade de troca e armazenamento de informações, instituições das mais diversas áreas do conhecimento têm produzido e armazenado eletronicamente uma grande quantidade de dados. Surge, então, a necessidade de técnicas eficientes para análise desses volumes de dados e extração de informações que possam ser úteis ao processo de tomada de decisão. Tais métodos têm contribuído para o surgimento de sistemas computacionais capazes de adquirir novos conhecimentos, novas habilidades e novas maneiras de organizar o conhecimento existente (Mitchell, 1997). Esses sistemas são, na sua maioria, produtos do Aprendizado de Máquina — AM —, cujos algoritmos são amplamente utilizados em tarefas de Descoberta de Conhecimento de Bases de Dados (*Knowledge Discovery from Databases — KDD*) principalmente para automatizar o processo de análise de dados e extração de conhecimento útil e preferencialmente novo (Fayyad et al., 1996).

Para que o conhecimento extraído possa ser melhor aproveitado e utilizado, ele deve estar representado em uma linguagem de fácil interpretação pelos usuários/especialistas. Entretanto, algoritmos de AM não supervisionados[1], em geral, representam os padrões extraídos por meio de agrupamentos de dados (clusters). Sob esse aspecto, o objetivo deste trabalho é aplicar uma metodologia de explicação dos clusters gerados a partir de algoritmos de AM não supervisionado, mais especificamente de *clustering* hierárquico, para obter uma descrição simbólica dos agrupamentos e, conseqüentemente, facilitar a compreensão e utilização do conhecimento extraído.

*Trabalho realizado com o auxílio do CNPq.

[1]Apesar de existirem outras técnicas associadas ao AM não supervisionado, tais como regras de associação e sumarização, neste trabalho o termo aprendizado não supervisionado é utilizado como referência ao agrupamento ou *clustering* de dados.

Este trabalho está organizado da seguinte maneira: Na Seção 2 são aprensentados brevemente alguns conceitos sobre o Aprendizado de Máquina não supervisionado e *clustering* hierárquico. A metodologia de explicação dos clusters e um experimento que ilustra essa metodologia, são descritos na Seção 3. Finalmente, na Seção 4 são apresentadas algumas considerações finais.

2. Aprendizado de máquina não supervisionado

No Aprendizado de Máquina não supervisionado, o conjunto de dados de entrada é composto por exemplos não rotulados, *i.e.*, não existe uma classe associada a cada exemplo. Nesse caso, são utilizados algoritmos de aprendizado para descobrir padrões nos dados a partir de alguma caracterização de regularidade. Assim, a tarefa consiste em agrupar uma coleção de exemplos segundo alguma medida de similaridade de modo que exemplos pertencentes ao mesmo cluster devem ser mais similares entre si e menos similares aos exemplos pertencentes a clusters diferentes (Everitt, 1993).

Existem diversas abordagens de *clustering*, tais como: probabilística, otimização, *clumping* e hierárquica (Jain et al., 1999; Sander et al., 2003). Cada uma utiliza uma maneira diferente para a identificação e representação dos clusters. Neste trabalho, os agrupamentos são obtidos por meio de um algoritmo de *clustering* hierárquico, o qual representa os clusters em uma estrutura conhecida como dendograma — Figura 1 — que consiste de um tipo especial de árvore, na qual os nós pais agrupam os exemplos representados pelos nós filhos. Dessa maneira, um agrupamento hierárquico agrupa os dados de modo que se dois exemplos são agrupados em algum nível, nos níveis mais acima eles continuam fazendo parte do mesmo grupo, construindo uma hierarquia de clusters. Essa técnica permite analisar os clusters em diferentes níveis de granularidade, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos.

Duas abordagens podem ser derivadas do *clustering* hierárquico: aglomerativo (*Botton-up*) e divisivo (*Top-down*). Na primeira abordagem, os dados são inicialmente distribuídos de modo que cada exemplo represente um cluster e, então, esses clusters são recursivamente agrupados considerando alguma medida de similaridade, até que todos os exemplos pertençam a apenas um cluster. Na abordagem divisiva, o processo inicia-se com apenas um agrupamento contendo todos os dados e segue dividindo-o recursivamente segundo alguma métrica até que alcance algum critério de parada, freqüentemente o número de clusters desejados (Berkhin, 2002).

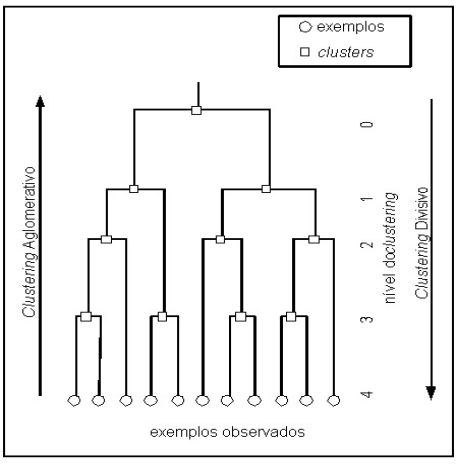

Figura 1: Dendograma.

A tarefa de interpretação dos clusters não é trivial. Em geral, exige a participação de um especialista no domínio da aplicação para análise dos clusters e atribuição de um significado conceitual. Entretanto, essa tarefa manual pode ser muito custosa e complexa. Pensando nessa dificuldade, Martins (2003) propôs uma metodologia, descrita na próxima seção, para interpretação semi-automática de clusters, a qual foi implementada utilizando algoritmos de agrupamento probabilístico.

3. Análise dos agrupamentos

A atribuição de um conceito aos clusters encontrados pelos algoritmos de *clustering*, em geral, é uma tarefa complexa que deve ser realizada pelo especialista do domínio da aplicação. Sob esse aspecto, seria interessante que essa tarefa fosse totalmente automática. Mas, as abordagens de *clustering* tradicionais não possibilitam que essa análise seja feita automaticamente, pois não utilizam conhecimento a priori, mas somente os dados para a extração do conhecimento neles embutido. Como alternativa para facilitar essa tarefa de interpretação dos clusters, Martins (2003) propôs uma metodologia para auxiliar a encontrar uma descrição simbólica dos clusters. Segundo essa metodologia, inicialmente os dados não rotulados são submetidos à um algoritmo de *clustering* para obter um conjunto de agrupamentos. Esse resultado é utilizado como entrada para uma ferramenta que rotula os exemplos, adicionando um atributo cujo valor é o cluster ao qual o exemplo pertence. Esse novo conjunto de dados é, então, utilizado como entrada de algum algoritmo de aprendizado supervisionado, utilizando o novo atributo como atributo classe do conjunto de dados, com o intuito de encontrar uma descrição simbólica para os clusters gerados. Finalmente, com a obtenção dessa representação simbólica, a interpretação do conhecimento extraído torna-se mais simples e menos custosa.

Para ilustrar a utilização dessa metodologia, a ser aplicada ao *clustering* hierárquico, é apresentado um experimento utilizando o conjunto de dados *Iris* da UCI (Blake and Merz, 1998). Esse conjunto de dados é composto de 150 exemplos descritos por quatro atributos contínuos (*comprimento da sépala*, *largura da sépala*, *comprimento da pétala* e *largura da pétala*) e contém informações referentes à três tipos da planta *Iris*. Os 150 exemplos estão igualmente distribuídos entre as três classes: *Iris Setosa*, *Iris Versicolor* e *Iris Virginica*, sendo 50 exemplos em cada classe, das quais somente uma é linearmente separável das demais. O atributo classe do conjunto de dados original não foi considerado durante a realização do *clustering*.

Os clusters foram obtidos por meio do algoritmo *CHAMELEON* (Karypis et al., 1999), utilizando a abordagem aglomerativa com o *co-seno* como medida de similaridade intra-cluster e *Complete Link* para similaridade inter-cluster. Para obtenção das regras foi utilizado o algoritmo *C4.5Rules* (Quinlan, 1993) com parâmetros padrão.

Inicialmente, o conjunto de dados foi submetido ao *clustering* para identificação de dois agrupamentos. Após, o resultado foi utilizado para rotular os exemplos e utilizar esse novo conjunto de dados como entrada para o algoritmo *C4.5Rules* para a geração das regras que descrevem simbolicamente os clusters encontrados. As regras geradas para esse primeiro conjunto de agrupamento são:

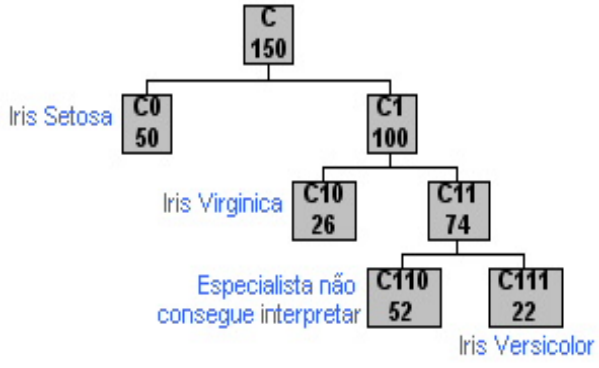

Figura 2: Resultados.

```
Regra 1:
        largura da pétala <= 0.6
         ->  classe C0  [97.3%]
Regra 2:
        largura da pétala > 0.6
        ->  classe C1  [98.6%]
```

Analisando essas regras, suponha que o especialista descobriu que os exemplos no cluster C0 referem-se à classe *Iris Setosa*, mas não consegue interpretar os exemplos no cluster C1. Nesse caso, o especialista pode realizar a análise dos dois clusters que pertencem à C1, os clusters C01 e C11 na Figura 2, rotulando os exemplos contidos em cada um desses clusters, que são utili-

zados para gerar um novo conjunto de regras. Suponha que após analisar essas regras o especialista consegue descobrir que os 26 exemplos contidos no cluster C10 descrevem o conceito *Iris Virginica*. Porém, não encontra uma boa descrição para os 74 exemplos contidos no cluster C11. Então, pode repetir o processo nos clusters C110 e C111. Suponha que a partir das regras obtidas no terceiro estágio, o especialista descobre que os exemplos contidos no cluster C111 referem-se à classe *Iris Versicolor*, mas não consegue interpretar o cluster C110. Esse processo de análise pode ser repetido pelo especialista enquanto houver dificuldade para encontrar o conceito representado pelos exemplos nos clusters.

4. Considerações finais

A utilização da abordagem de *clustering* hierárquico em conjunto com essa metodologia de explicação dos clusters facilita a interpretação dos resultados e a utilização do conhecimento extraído, pois uma das principais características do *clustering* hierárquico é a flexibilidade para a análise dos agrupamentos nos diferentes níveis do dendograma, o que naturalmente sugere um refinamento na análise dos padrões neles descritos. Com isso, a tarefa de atribuição de conceitos aos clusters deixa de ser totalmente manual e passa a ser uma tarefa semi-automática, necessitando ainda da participação do especialista, mas exigindo menor esforço para a compreensão e eventual atribuição de um significado semântico aos clusters.

A proposta aqui apresentada é um dos temas a serem desenvolvidos como parte de um trabalho de mestrado (Metz, 2005). Ela será projetada e implementada como um módulo de um sistema computacional de grande porte para extração e análise de conhecimento, em desenvolvimento no Laboratório de pesquisa LABIC-ICMC, com o objetivo de auxiliar o especialista do domínio na interpretação de clusters.

Referências

Berkhin, P. (2002). Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA.

Blake, C. and Merz, C. (1998). UCI repository of machine learning databases.

Everitt, B. S. (1993). *Cluster Analysis*. Edward Arnold, 3 edition.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 82–88.

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

Karypis, G., Han, E.-H. S., and NEWS, V. K. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75.

Martins, C. A. (2003). *Uma Abordagem para Pré-processamento de Dados Textuais em Algoritmos de Aprendizado*. PhD thesis, ICMC-USP.

Metz, J. (2005). Interpretação de clusters gerados por algoritmos de agrupamento hierárquico. Monografia para Exame de Qualificação de Metrado.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

Sander, J., Qin, X., Lu, Z., Niu, N., and Kovarsky, A. (2003). Automatic extraction of clusters from hierarchical clustering representations. In *PAKDD - Pacific-Asia Knowledge Discovery and Data Mining*, volume 2637 of *LNAI*, pages 75–87. Springer-Verlag.