**21.103 Sistemas de Gestão de Bases de Dados**

**Atividade Formativa**

Leia o capítulo 20 – Data Warehouse and Mining (6ª edição) do livro adotado e responda às seguintes questões:

1) Qual as semelhanças e diferenças entre bases de dados e ‘data mining’?

As questões de Bases de Dados são muito semelhantes às de Data Mining. Neste trabalho vamos ver a diferença entre as duas abordagens.

Em Base de Dados pretende-se por exemplo:

• Identificar os clientes que compraram mais de 1.000 euros;

• Identificar os dois produtos mais vendidos;

• Identificar os 10 clientes com mais reclamações;

Enquanto que, em Data Mining procura-se:

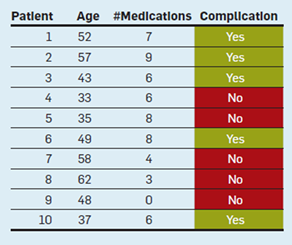
• Identificar os grupos de clientes com hábitos de compra idênticos (‘clustering’)

• Encontrar o produto X que é adquirido com o produto Y (regras associativas)

• Encontrar os atributos que levam os clientes a reclamar (classificação)

Embora as questões sejam semelhantes, nas Bases de Dados é apresentado um padrão (e.g. consulta SQL) e são devolvidos dados, por outro lado, e em Data Mining são fornecidos os dados e pretende-se extrair padrões (e.g. regras associativas). Para exemplificar, vamos considerar a Tabela 1 com dados de um hospital.

Tabela 1 – Dados de complicações num hospital



Uma consulta (‘query’) típica de uma base de dados, será encontrar os doentes que tiveram complicações, sendo devolvido os doentes 1,2, 3, 6 e 10. Em SQL teremos:

**Select patient, age, #medications**

**From patientTable**

**Where complication = ‘yes’**

Por outro lado, em Data Mining são fornecidos os dados e pretende-se extrair padrões. Assim, para a mesma tabela pretendemos saber os atributos que causam as complicações médicas. Com um algoritmo de classificação encontraremos a seguinte regra: **(Age>=37 and #Medications >=6) => Complication =Yes**

Isto é, quem tem mais de 37 anos e toma mais do que 6 medicamento tem complicações médicas. Julgamos ter identificado a diferenças entre Bases de Dados e Data Mining:

• nas Bases de Dados é apresentado um padrão e são devolvidos dados,

• em Data Mining são fornecidos os dados e são devolvidos padrões.

2) Que tipos de algoritmos de ‘data mining’ conhece ?

Resposta: Os tipos podem ser considerados os supervisionados (os preditivos) e os não-supervisionados (ou descritivos). Exemplo dos supervisionados é a Classificação (3.a) e dos não-supervisionados as Regras de Associativas (3.b) e Segmentação (3.c).

2.a) Classificação

Os modelos em árvores são designados por árvores de decisão, no caso de problemas de classificação e árvores de regressão nos problemas de regressão. Uma árvore de decisão usa a estratégia ‘divisão e conquista’ para resolver um problema classificação. Cada é problema decomposto em subproblemas mais simples, aplicando recursivamente a mesma a mesma estratégia a cada subproblema. A capacidade de discriminação de uma árvore provém: da divisão do espaço definido pelos atributos em subespaços, a cada subespaço está associada uma classe do atributo ‘target’ (ou variável dependente).

O problema de construir uma árvore de decisão consistente com um conjunto de exemplos, e com o menor número de nós é um problema NP completo. Todos os algoritmos são heurísticas recursivas e ‘greedy’ (gulosas) que utilizam otimização local. As regras classificatórias são induzidas dos dados por partição recursiva, construindo uma árvore da raiz para as folhas, por divisão das instâncias segundo condições simples sobre as variáveis independentes. Estes algoritmos são especialmente adequados para dados com muitos atributos e com variáveis independentes categóricos ou discretos.

Existem vários algoritmos de classificação que utilizam estratégias diferentes para a subdivisão da árvore.

CART, ‘classification and regression tree’, proposto por Breiman et al. (1984) utiliza o critério do índice Gini. A diferença entre o índice Gini para o nó pai e a soma dos valores para o nó filho (ponderada pela proporção de casos em cada filho) é apresentada na árvore como ‘improvement’.

ID3, C4.5 e C5.0 são versões distintas por J. Ross Quinlan (1986, 1993, 1998) que utilizam o conceito de entropia. Os algoritmos maximizam uma medida de entropia (ou conteúdo em informação) dos nós ‘filhos’ relativamente ao nó ‘pai’.

CHAID, ‘CHi-square Automation Interaction Detection, (Kass 1980) e QUEST, ‘Quick, Unbiased, Efficient Statistical Tree’ (Loh e Shih 1997) utilizam a probabilidade de significância do Qui-Quadrado da variável dependente categórica ou o teste-F da variável dependente quantitativa, para maximiza a distinção estatística das subárvores ‘filhos’.

O sobre-ajustamento (‘overfitting’) ocorre quando o modelo se adapta demasiado bem aos dados de treino e gera demasiados erros nos dados de teste. O sobre-ajustamento tem origem em árvores demasiado extensas difíceis de generalizar e interpretar. Para ultrapassar este problema é utlizada a técnica da poda da árvore. Esta abordagem está relacionada com o princípio da parcimónia, da ‘Occam’s razor’ ou da simplicidade. Existem duas possibilidades: parar o crescimento da árvore mais cedo (pré-poda) ou construir uma árvore completa e podar a árvore (pós-poda). Segundo Quinlan (1988) deixar crescer e podar posteriormente, usando validação cruzada para avaliar o erro resultante da poda, é um processo mais fiável.

A avaliação de algoritmos supervisionados é normalmente realizada por meio de análise do desempenho do preditor na classificação de novas instâncias, ou seja, instâncias que não foram utilizadas no conjunto de treino. Os principais métodos de amostragem existentes são: ‘holdout’, amostragem aleatória, validação cruzada e ‘bootstrap’.

Como métodos de comparação de algoritmos, a partir da Matriz de Confusão, um conjunto vasto de métricas pode ser encontrado, tais como a taxa de acerto, precisão, sensibilidade e especificidade. Uma forma alternativa de avaliar classificadores em problemas binários é baseada na análise das curvas ROC, ‘receiving operating characteristics’. O gráfico bidimensional denominado espaço ROC, representa as medidas de taxa de falsos positivos e a taxa de verdadeiros positivos, nos eixos X e Y.

Para além das árvores de decisão existem outros classificadores. O algoritmo dos K-vizinhos mais próximos K-NN, ‘nearest-neighbor’, (Cover, Hart 1967) consiste em selecionar K instâncias mais próxima de uma instância não classificada e usar a maior frequência das instâncias escolhidas para atribuir a classificação. O algoritmo ‘Naive Bayes’ utiliza o teorema de Bayes com êxito em vários conjuntos de dados, dado que a estimativa das probabilidades não precisa de elevado rigor, basta que a classe modal seja correta. Um outro algoritmo de fácil implementação é o algoritmo 1R descrito num artigo de Holte (1993) testado em 16 conjuntos de dados.

Nas técnicas de classificação para além da qualidade das métricas é também importante a compressão dos resultados, pelo que as técnicas de seleção de atributos se relevam de extrema importância (Liu, Yu 2005). Como exemplo destas técnicas aplicadas a grandes volumes de dados podemos referir (Cavique et al. 2018b).

2.b) Regras Associativas

No tópico de Regras Associativas podemos abordar os seguintes itens:

• Geração itens frequentes: algoritmo ‘Apriori’

• Métricas e Geração de regras: cortes baseados na métrica de confiança

• Métodos alternativos: padrões de sub-grafos, padrões sequenciais

O algoritmo que gera o conjunto de itens mais frequentes utilizando regras associativas, é conhecido pelo algoritmo ‘Apriori’ e foi apresentado por Agrawal et al. (1996). O algoritmo utiliza como ‘input’ a tabela com várias transações de produtos, tendo como ‘output’ um conjunto de regras do tipo E=>D.

O algoritmo ‘Apriori’, num primeiro passo, filtra os produtos com frequência superior ao suporte (ou frequência) mínima estabelecida, sup\_min; como resultado obtém-se P1, que corresponde ao conjunto de produtos com um único produto com frequência superior a sup\_min. Em seguida iterativamente, para cada conjunto de produtos Pk gera todos os candidatos a Pk+1, tal que Pk⊆Pk+1, associando à combinação existente um novo produto frequente. Para cada candidato Pk+1 o algoritmo vai varrer todas as transações na tabela e verificar se a sua frequência é superior à frequência mínima, reduzindo assim a explosão combinatória sem perder qualquer item frequente. O ciclo termina quando são gerados todas as combinações para Pmax\_k. A complexidade do algoritmo é O(Nmax\_k), sendo N o número de produtos.

Na geração de produtos frequentes existem três métricas fundamentais: suporte, confiança e melhoramento (‘lift’).

A medida de suporte, ou frequência relativa, é obtida pela razão da frequência absoluta de E&D pelo número total de transações. Podemos interpretar ainda esta medida como a probabilidade de compra de E e D em conjunto.

A medida de confiança é dada pelo suporte(E&D) / suporte(E) e pode ser vista também como a probabilidade condicionada de comprar D se comprou E.

Finalmente a medida de melhoramento é obtida por suporte(E&D) / [suporte(E) x suporte(D)] e indica qual o desempenho da regra comparando com uma regra “nula” i.e. quando a medida de melhoramento é maior que 1, então a regra é melhor do que uma escolha aleatória.

Depois da geração de um grande conjunto de regras associativas, o utilizador pode criar corte com base nas métricas existentes e escolher as regras que melhor se adaptam à solução pretendida.

A descoberta de padrões sequenciais combina a descoberta de conjuntos de itens frequentes e a ordem em que aparecem. A maioria das técnicas de descoberta de padrões de sequência apresentar algumas desvantagens como a geração de um grande número de regras e a falta de escalabilidade. Em Cavique (2015) é proposto o algoritmo Ramex, que considera a análise do todo em vez das partes, proporcionando assim uma visão holística das sequências. O algoritmo analisa os ‘logs’ de eventos e permite que o utilizador não especialista identifique as sequências usando uma visualização de poli-árvore. A escalabilidade associada a estruturas de dados condensados, que reduzem os dados sem perder informações, permite lidar com o desafio ‘Big Data’. Ramex foi implementado em diferentes cenários por Tiple, Cavique e Marques (2016).

2.c) Segmentação

No tópico de Segmentação podemos aborda os seguintes itens:

• Algoritmo K-médias (‘k-means’)

• Métodos hierárquicos

• Avaliação da segmentação

• Outros assuntos: segmentação usando grafos

O conceito de segmentação, ou análise de agrupamento (‘clustering’), introduz-se como sendo a estruturação de um conjunto de entidades (tipicamente multidimensionais) de modo a constituir grupos de entidades tais que se verifica homogeneidade intra-grupos e heterogeneidade entre-grupos.

Os algoritmos de segmentação dividem-se em dois grandes grupos: os particionais (ou não-hierárquicos) e os hierárquicos.

Os métodos particionais baseiam-se na obtenção de um número predefinido de ‘clusters’, k, que conterão todos os casos observados. O principal representante desta categoria é o algoritmo k-médias. O algoritmo k-médias particiona o conjunto de dados em k ‘clusters’, em que k é fornecido pelo utilizador. Esses ‘clusters’ são formados de acordo com uma medida de similaridade. O algoritmo k-médias utiliza uma técnica de realocação iterativa, que pode convergir para um ótimo local, evitando a solução ótima.

Os métodos hierárquicos geram uma hierarquia de grupos desde 1 até ao número de observações. Distinguem-se os hierárquicos aglomerativos que iniciam com N grupos e vão-se juntando os mais parecidos e os hierárquicos divisivos que iniciam com um grupo e vão particionando os grupos.

Na avaliação de modelos descritivos, compreender e operacionalizar o conceito de boa partição não é uma tarefa fácil (Milligan 1980). No entanto, é comum a referência a algumas propriedades desejáveis de uma partição que se podem associar a bons resultados de agrupamento: i) os grupos são compactos e bem separados; ii) a solução é robusta, resistindo a algumas variações sobre os dados e/ou sobre o processo de agrupamento.

O índice Silhueta (Rousseeuw 1987) apresenta-se como um exemplo clássico que ilustra uma verificação de i) com grupos compactos e bem separados. No que diz respeito à robustez da solução, para avaliar a concordância de duas partições (resultados de replicações sobre uma mesma amostra na avaliação cruzada) são aplicados o Índice de Rand (Rand 1971) e o Índice de Rand modificado (Hubert and Arabie 1985).

A deteção de agrupamentos (‘community detection’) na análise de redes sociais é uma questão muito importante. O algoritmo ComDetection (Cavique et al. 2018a) permite a identificação de comunidades e atores que vinculam dois ou mais grupos. A redução de dados permite uma boa visualização de grafos densos, utilizando um hipergrafo que identifica comunidades e intermediários.

3) Com base nas “Lecture Notes: Ciências dos dados”, desenvolva questões para base de dados e para data mining num exemplo à sua escolha.

Resposta:

Por exemplo, num hotel pretende-se saber de uma base de dados:  
- Qual a taxa de cancelamento do mês anterior?  
- Qual foi a bebida X mais vendida que acompanha o prato de peixe Y?  
- Quantas reclamações de clientes ocorreram no mês anterior?

As consultas de *data mining* procuram obter informações com base em padrões nos dados:

- Quais os clientes que vão cancelar no próximo mês? (classificação)   
- Qual a bebida X para acompanhar o prato de peixe Y? (associação)

- Quais os N grupos de clientes que frequentam o hotel? ('clustering')