Extração de conhecimento em banco de dados geográficos

Nara Martini Bigolin*

1. Introdução

O processo de descoberta do conhecimento visa extrair modelos e informações implícitas a partir de grandes bancos de dados. Essa tecnologia é importante devido ao crescimento incessante dos bancos de dados. Entretanto, as técnicas de manipulação desses bancos são ineficazes perante a explosão do tamanho dos bancos de dados. Assim, a união de técnicas multi-disciplinares de diversas áreas como Banco de Dados, Inteligência Artificial e Estatística possibilitam a execução desse processo de descoberta de conhecimento.

A comunidade de Banco de Dados [ABI 95] desenvolveu Sistemas de Gerência de Banco de Dados, que tem por objetivo oferecer ferramentas que possibilitam o armazenamento e a manipulação de grande quantidade de informações estruturadas e um acesso rápido. A preocupação principal no desenvolvimento deste tipo de sistema são os aspectos de modelagem de dados, as linguagens de consultas e a eficiência de recuperação de dados.

A comunidade de Inteligência Artificial [ANA 98] interessou-se pela extração de conhecimento e a aprendizagem a partir de uma quantidade reduzida de informações. A combinação dessas duas abordagens deu origem a uma nova tecnologia chamada Descoberta de Conhecimentos em Banco de Dados (DCBD) [FAY 96].

Os principais desafios deste processo se situam nas etapas de seleção e de pré-tratamento dos dados. A dificuldade na seleção consiste em obter as informações relevantes a serem utilizadas como amostra e a dificuldade no pré-processamento de dados consiste em colocá-las num formato apropriado [BIG 99]. Esses procedimentos são necessários, pois os algoritmos de extração não são capazes de manipular estruturas complexas com grande quantidade de dados.

As técnicas tradicionais de aprendizagem tratam de dados vindos diretamente do mundo real, enquanto que as técnicas de extração de conhecimento utilizam dados vindo de um banco de dados.

No primeiro caso, os dados de entrada são representados em uma estrutura simples apropriada para a aprendizagem. Além disso, as informações relevantes a serem utilizadas pelo processo de extração de conhecimento são definidas com a ajuda de um especialista [BIG 98].

* bigolin@inf.ufrgs.br

No segundo caso, a estrutura de dados é mais complexa, a quantidade de informações é enorme e nem todas as informações são relevantes. Assim, é necessário efetuar dois pré-tratamentos : o primeiro serve para encontrar um subconjunto de informações relevantes e o segundo possibilita a formatação dessas informações numa estrutura apropriada para a aplicação de técnicas de extração de conhecimento.

Nos propomos uma seqüência de técnicas utilizadas para todas as etapas do processo de descoberta de conhecimento, desde a seleção dados até a extração de conhecimento a partir de bancos dados geográficos. Esta seqüência de passos foi implementado através de uma linguagem de consulta com a sintaxe seguinte :

DATAMINING Algoritmo de Extração de Conhecimento

NEBULOSA Função Nebulosa

MEDIDA Funções de Associação

REDUÇÃO Funções de Generalização

WITH

SELECT x.house

FROM x in Database1

WHERE Condicao

onde a cláusula DATAMINING permite a especificação do algoritmo de extração de conhecimento. Nesse exemplo, é utilizado um algoritmo de construção de árvore de decisão.

A cláusula NEBULOSA é utilizada para determinar a função da teoria de conjunto nebuloso. Foi utilizado a função de *pertinência* para eliminar as imprecisões resultantes dos outros processamentos.

A cláusula MEDIDA possibilita a associação dos objetos espaciais. Por exemplo, a função centróide foi utilizada para encontrar o centro de cada objeto.

A cláusula REDUÇÃO permite especificar o algoritmo de generalização que é utilizado para reduzir as informações espaciais.

Cada uma dessas cláusulas serão explicadas dentro das etapas do processo de descoberta (seleção de dados, pré-processamento de dados, extração de conhecimento e interpretação de conhecimento).

2. Processo de descoberta de Conhecimento

2.1 Seleção dos Dados

Inicialmente, uma seleção é feita usando a cláusula SELECT FROM WHERE. O resultado da cláusula SELECT é um conjunto de dados selecionados a partir da banco de dados. Entretanto, nosso banco de dados é geográfico, o que requer uma linguagem de consulta para este tipo de dados [EGE 94]. Esse tipo de linguagem é chamado linguagem de consulta espacial e manipula dados espaciais e não espaciais.

Os dados não espaciais são informações que descrevem características como : nome, população da cidade, etc.

Os dados espaciais especificam as localizações dos dados não espaciais e são representados geralmente por três primitivas espaciais: ponto, linha e área. Um ponto representa o aspecto geométrico de um objeto em sua posição no espaço. Para o exemplo, uma cidade pode ser um ponto ou uma área geográfica dependendo da escala do mapa. Uma região é uma abstração de um objeto num espaço dimensional, por exemplo um país, um lago, um parque nacional ou uma casa representados num mapa da pequena escala. Uma linha é a abstração para facilitar a locomoção no espaço, ou conexões no espaço (estradas, rios, cabos telefônicos, etc.).

Os dados não espaciais podem ser manipulados com uma linguagem de consulta clássica, mas as características de dados espaciais requerem algumas operações específicas que podem processar e manipular dados espaciais representados graficamente. Numa consulta espacial, essas operações são representadas por métodos na cláusula WHERE. Por exemplo,

```
SELECT x.casa
FROM  x in Banco de Dados 1
WHERE  x.casa --> in Area(CoordPtMin, CoordPtMax);
```

permite selecionar o conjunto de casas pertencendo a uma área que é determinada na condição da cláusula WHERE, onde uma função de calculo espacial pode ser utilizado. Por exemplo, o método *inArea(CoordPtMin, CoordPtMax)* determina se um objeto *casa* está numa área definida pelas *CoordPtMin e* CoordPtMax. Esta área pode ser definida pelo usuário com uma interface gráfica.

Uma vez os dados selecionados, os mesmos devem serem preparados para a etapa de extração de conhecimento. Esta preparação é feita na etapa de pré-processamento de dados.

2.2 Pré-processamento de dados

A cláusula REDUÇÃO utilizando funções de generalização reduz a representação em pontos (cordenadas cartesianas) de cada casa a um único ponto, reduzindo assim a quantidade de informações espaciais. Em seguida uma relação entre esses pontos é necessária. Para isso, determina-se a cláusula MEDIDA que determina uma associação entre os objetos espaciais. Por exemplo, a relação pode ser uma função matemática d(p1, p2) que calcula a distância euclidiana entre dois pontos (p1 e p2) que representam casas.

A cláusula NEBULOSA permite tratar as imprecisões geradas nas etapas anteriores Por exemplo, as distâncias podem ser analisadas utilizando as modalidades nebulosas. Assim, dada uma casa c, calcula o número de casas que pertencem a três áreas nebulosas :perto, longe e muito longe, ou seja, para cada casa c' diferente de c, a distancia d(pc, pc') é analisada. A partir destas informações uma base de exemplos é construída e a extração de conhecimento pode ser realizada.

2.3 Extração de Conhecimento

A cláusula DATAMINING possibilita a definição do algoritmo de extração de conhecimento que neste exemplo é o algoritmo de construção de árvores de decisão nebulosa.

A árvore de decisão foi construída a partir de um conjunto de exemplos (a base de aprendizagem). Cada caso é um exemplo conhecido associado a um par (descrição, classe), onde a descrição é um conjunto de atributo/valor, a qual é o conhecimento .

Essas árvores podem ser consideradas como um conjunto de regras nebulosas. Assim, o algoritmo extrai um conjunto de regras nebulosas na forma de árvore a partir do conjunto de dados (base de aprendizagem). Formalmente, a árvore de decisão é equivalente ao conjunto de regras nebulosas RB = R1,...,RN. Cada caminho na árvore da raiz à folha é equivalente a uma regra ***R: Pr--> Co***. A premissa ***Pr*** é composta de testes sobre os valores de atributos e a conclusão ***Co*** é o valor da decisão representado na folha da árvore.

Mais claramente, cada nodo da árvore é associado com um teste sobre os valores nebulosos de um atributo, cada arco a partir dos nodos são etiquetas com valores dos atributos que pertencem a uma partição universal baseada na teoria dos subconjuntos nebulosos e cada folha da árvore é associada com uma classe.

As árvores de decisão nebulosas manipulam valores nebulosos durante a construção da árvore e na classificação de novos casos. O uso da teoria dos subconjuntos nebulosos ajuda na compreensão das árvores de decisão com atributos numéricos.

O algoritmo descobre o número de casas que pertencem ou não a uma zona urbana, considerando um intervalo impreciso definido pela teoria dos subconjuntos nebulosos.

2.4 Validação e interpretação do conhecimento

Uma vez, a base de regras obtida, elas devem ser validadas. Existem vários métodos para essa validação. Foi escolhido a técnica tradicional chamada: base de teste. Isso significa que as regras são aprendidas a partir de uma cidade e após essas mesmas regras são utilizadas em outras cidades comparando a porcentagem de boa classificação. No exemplo, a região estudada é composta de três cidades T1, T2 e T3. A base de exemplos foi gerada a partir da T1 e o conjunto de teste foi gerada a partir das outras duas cidades.

A média de erros na classificação das cidades das cidades T2 e T3 é de 10.1%, ou seja, dadas 413 casas a partir de uma nova zona, 371 casas foram classificadas perfeitamente em urbana ou não urbanas.

3. Conclusão

Neste capítulo foram apresentados sucintamente uma técnica para todas as etapas do processo de descoberta de conhecimento. Esta técnica consiste de uma linguagem de consulta que trata todas as etapas do processo de descoberta de conhecimento. Um exemplo foi desenvolvido para mostrar a validade do processo de descoberta.

Neste contexto, pode-se concluir que a tecnologia de descoberta de conhecimento é uma área de grandes perspectivas, onde o mais importante é a manipulação de técnicas multidisciplinares, para conseguir tratar todo o processo de uma maneira eficaz e correta.

Referências Bibliográficas

[AB 98] ANAND S. and BUECHNER A.. Decision Support Using Data Mining. Financial Times Pitman Publishin, 1998.

[AHV 95] ABITEBOUL S., HULL R., and VIANU V.. Foundations of databases. Addison-Wesley Publishing Company, 1995.

[AIS 93] AGRAWAL R., IMIELINSKI T., and SWAMI A.. Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering, 5(6):914--925, December 1993.

[BFOS 84] BREIMAN L., FRIDMAN J. H., OLSHEN R. A., and STONE C. J.. Classification and regression tree. Chapman and Hall, 1984.

[BIG 99] BIGOLIN N. M.. Méthodes pour la Découverte de Connaissances à partir d'une Base de Données Orientée Objets : Le système LARECOS. PhD thesis, Université Paris VI - Paris, Octobre 1999.

[BIG 98] BIGOLIN N. M. and MARSALA C.. Fuzzy spatial oql for fuzzy knowledge discovery in databases. In J. M. Zytkow and M. Quafafou, editors, 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, pages 246--254, Nantes, France, September 1998. Springer Verlag. LNCS v. 1510.

[CD 97] CHAUDHURI S. and DAYAL U.. An overview of data warehousing and olap technology. Sigmod Record, 26:65--74, 1997.

[FMPS 97] FAYYAD U., MANNILA H., and PIATETSKY-SHAPIRO G., editors. Data Mining and Knowledge Discovery An International Journal. Kluwer Academic Publishers, 1997.

[FPSS 96] FAYYAD U., PIATETSKY-SHAPIRO G., and SMYTH P.. From data mining to discovery knowledge in databases. AI Magazine, 3(17):37--54, 1996.

[HAN 97] HAN J.. Olap mining: An integration of olap with data mining. In Proc. 1997 IFIP Conference on Data Semantics (DS-7), pages 1--11, Leysin, Switzerland, Oct 1997.

[HSK 98] HAN J., STEFANOVIC N., and KOPERSKI K.. Selective materialization: An efficient method for spatial data cube construction. In Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, Australia, 1998.

[JAN 98] JANIKOW C. Z.. Fuzzy decision trees: Issues and methods. IEEE Transactions on Systems, Man, and Cybernetics, 28(1):1--14, 1998.

[LD 95] LAVRAC N. and DZEROSKI S.. Inductive logic programming: Techniques and applications. Ellis Horwood. , 1995.

[LT 92] R. Laurini and D. Thompson. Fundamentals of spatial information systems. Academic Press, 1992.

[MB 98] MARSALA C. and BIGOLIN N. M.. Spatial data mining with fuzzy decision trees. In Nelson F. F. Ebechen, editor, DATA MINING - Proc. International Conference on Data Mining, pages 235--248. WIT Press, Septembre 1998.

[QUI 86] QUINLAN J. R.. Induction of decision trees. Machine Learning, 1:81--106, 1986.