TRATAMENTO DE DADOS FALTANTES PARA ANÁLISE EM GRANDE ESCALA

MAC-5701 – Tópicos em Ciência da Computação

Clodis Boscarioli
Orientador: Prof. Dr. Roberto Marcondes César Jr.
Co-orientador: Prof. Dr. João Eduardo Ferreira
IME – Instituto de Matemática e Estatística
USP – Universidade de São Paulo

Resumo

Este documento refere-se à descrição de um plano de estudos enfocado no levantamento bibliográfico sobre a etapa de pré-processamento de dados para descoberta de conhecimento em bancos de dados, visando identificar os desafios que área impõe, mais especificamente, no tratamento de dados faltantes. A Seção 1 traz algumas justificativas que motivaram o interesse para o desenvolvimento deste estudo. A seguir, na Seção 2, os prazos e metas do plano de estudo são apresentados; e, finalmente, na Seção 3, são apresentados comentários finais do plano de trabalho.

1. Introdução

A aplicação de técnicas de Data Mining em Banco de Dados é motivada pela crescente necessidade de descobrir novas informações em bancos de dados existentes, na forma de regras ou padrões. A integração de Banco de Dados e Data Mining não é trivial e várias são as dificuldades inerentes a esse processo, seja na preparação dos dados, na definição dos operadores, na análise do algoritmo a ser adotado ou, mesmo, na interpretação dos conhecimentos gerados. Dentre todas as etapas de Data Mining, é de particular interesse neste estudo a preparação dos dados, na qual são despendidos esforços para garantir a qualidade aos dados, fundamental para uma mineração eficiente.

A fase de pré-processamento de dados [10], [14], [20], [23] inicia-se após a coleta e organização dos dados em uma base de dados. Podem existir diversos objetivos na fase de pré-processamento de dados. Um deles é solucionar problemas nos dados, tais como identificar e tratar dados corrompidos, atributos irrelevantes e valores desconhecidos.

A ausência de dados representa um problema para os algoritmos de mineração, uma vez que pode dificultar sua aplicação a problemas reais com bancos de dados incompletos.

Devem, também, ser adotadas estratégias [2], [3], [4], [9], [13], [16], [22], [24], [25], [26] para resolver problemas de ausência de dados. Uma possível solução seria utilizar técnicas de reconhecimento de padrões em dados espúrios ou faltantes de modo a diminuir o trabalho de "força bruta" costumeiramente utilizado nesta fase de tratamento de dados.

Apesar da freqüência com que dados desconhecidos ocorrem em bases de dados, estes são tratados, muitas vezes de forma simplista. Entretanto, o tratamento destes valores deve ser cuidadosamente planejado, para que não sejam introduzidas distorções no conhecimento induzido.

Pode-se também estar interessado em aprender mais a respeito dos dados, o que pode ser feito, por exemplo, por meio de visualizações desses dados. Ou ainda, pode-se estar interessado em alterar a estrutura dos dados, por exemplo, por meio da alteração do grau de granularidade dos dados.

As ações realizadas na fase de pré-processamento de dados visam preparar os dados para que a fase de extração de conhecimento seja mais efetiva [5].

De maneira mais específica, os objetivos deste plano de estudo são:

(a) delimitar os conceitos e problemas de pré-processamento de dados para mineração;

(b) apresentar estratégias para o tratamento de dados faltantes, assim como as vantagens e limitações das mesmas.

2. Plano de Trabalho e Cronograma

Nesta seção serão destacadas as atividades pert inentes ao desenvolvimento do relatório de estudo dirigido referente à disciplina Tópicos em Ciência da Computação, conforme cronograma abaixo especificado.

PERÍODO	ATIVIDADE
Agosto - Setembro	• Levantamento bibliográfico • Estudo da bibliografia geral sobre a etapa de pré-processamento de dados para a mineração
Agosto - Setembro	• Levantamento, na literatura, das estratégias para tratamento de dados faltantes
Agosto - Outubro	• Estudo e comparação das estratégias anteriormente investigadas
Outubro – Novembro	• Apresentação de Seminários (MAC-5700 – Seminários em Ciência da Computação)
Outubro – Novembro	• Redação do relatório final do estudo realizado

3. Observações Finais

A execução deste Plano de Trabalho visa dar subsídios teóricos para o desenvolvimento de um projeto de doutorado na área de Reconhecimento de padrões em banco de dados, área esta que, visto a diversidade de aplicações, vem crescendo em volume e quantidade.

Pela pesquisa preliminar realizada, vê-se que a preparação dos dados é uma etapa de suma importância para o processo de descoberta de conhecimento em bancos de dados e uma das etapas mais trabalhosas deste processo.

4. Referências

[1] AGRAWAL, R., IMIELINSKI, T., SWAMI, A. Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering. vol. 5, no. 6., dez/1993.

[2] AGGARWAL, C. C., PARTHASARATHY, S. Mining Massively Incomplete Data Sets by Conceptual Reconstruction. KDD 01. São Francisco. USA, 2001.

[3] BATISTA, G. E. A. P. A., MONARD, M. C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning.

[4] BATISTA, G. E. A. P. A., MONARD, M. C. A Study of K-Nearst Neighbour as a Model-Based Method to Treat Missing Data. In Argentine Symposium on Artificial Intelligence, 2001.

[5] BATISTA, G. E. A. P. A. Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de Doutorado. ICMC-USP, São Paulo, 2003.

[6] CABENA, P., HADJINIAN P. O., STADLER, R, VERHEES, J., ZANASI, A. Discovering Data Mining: From Concepts to Implementations. Prentice Hall, USA, 1997.

[7] CESAR JUNIOR, R. M., COSTA, L. F. Shape Analysis and Classification: Theory and Practice. Boca Raton : CRC Press, 2001.

[8] DUNHAM, M. H. Data Mining: Introductory and Advanced Topics. Prentice Hall, New Jersey, 2002.

[9] DUDA, R. O., HART, P. E., STORK, D. G. Pattern Classification. 2 Ed. John Wiley & Sons, Inc., New York, 2001.

[10] FAMILI, A. SHEN, W.-M., WEBER, R., SIMOUDIS, E. Data Processing and Intelligent Data Analysis, 1997.

[11] FAYAAD, U., SHAPIRO, G. P., SMYTH, P. From Data Mining to Knowledge Discovery in Databases. IA Magazine, 1996.

[12] GHAHRAMANI, Z., JORDAN, M. I. Learning from Incomplete Data. MIT, 1994.

[13] GRZYMALA-BUSSE, J. W., HU, M. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In RSCTC'2000.

[14] HAN, J., KAMBER, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, USA, 2000.

[15] KIMBALL, R. Data Warehouse Toolkit. Makron Books, Rio de Janeiro, 1998.

[16] LAKSHMINARAYAN, K, HARP, S. A., SAMAD, T. Imputation of Missing Data in Industrial Databases. Applied Intelligence, 1999.

[17] LITTLE, R. J., RUBIN, D. B. Statistical Analysis with Missing Data. 2ª. Ed. New York: John Wilwy and Sons, 2002.

[18] LAVINGTON, S. DEWHURST, N., WILKINS, E., FREITAS, A. Interfacing Knowledge Discovery Algorithms to Large Database Management Systems. ELSEVIER, Information and Software Technology, 1999.

[19] NAVATHE, S.B., ELMASRI. R. E. Fundamentals of Database Systems. Addison Wesley Pub., USA, 2001.

[20] PYLE, D. Data Preparation for Data Mining. San Francisco, CA: Morgan Kaufmann, 1999.

[21] RAMAKRISHNAN, R., GEHRKE, J. Database Management Systems. McGraw-Hill, USA, 2000.

[22] RAMONI, M., SEBASTIANI, P. Learning Bayesian Networks from Incomplete Databases. Knowledge Media Institute Technical Report, United Kingdom, 1997.

[23] ZHANG, S., YANG, Q., ZHANG, C. Data Preparation for Data Mining.

[24] ZENG, Z., LOW, B. T. Classifying Unseen Cases with Many Missing Values. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1999.

[25] WEISS, S. M., INDURKHYA, N. Decision-rule Solutions for Data Mining with Missing Values.