Data Mining and Knowledge Discovery, 12, 121–125, 2006
© 2006 Springer Science + Business Media, Inc. Manufactured in the United States.
DOI: 10.1007/s10618-006-0041-y

Mining Multiple Data Sources: Local Pattern Analysis

SHICHAO ZHANG* zhangsc@it.uts.edu.au
Department of Automatic Control, Beijing University of Aeronautics and Astronautics, Beijing 100083, China; Faculty of Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia

MOHAMMED J. ZAKI zaki@cs.rpi.edu
Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Published online: 7 April 2006

1. Introduction

Many large organizations process data from multiple data sources, such as the different branches of an interstate or international company. Also the Web has emerged as a large, distributed data repository consisting of a variety of data sources and formats. Although the data collected from the Web or multiple local datasets brings us opportunities in improving the quality of decisions, it generates significant challenges at the same time, for example, how to efficiently discover useful knowledge from different data sources and how to integrate them. We call this the multiple data source (MDS) mining problem, and it has recently been recognized as an important research topic in the data mining community.

This problem is difficult to solve due to the fact that MDS mining involves the discovery of useful patterns in multidimensional spaces across diverse sources; and putting all data together from different sources might amass a huge database for centralized processing and might cause serious problems in data privacy, data inconsistency, data conflict, and data irrelevance.

On the other hand, mining local patterns at different data sources and forwarding the local patterns (rather than the original raw data) to a centralized place for global pattern analysis can provide a feasible way to deal with MDS problems (Zhang et al., 2004). Local knowledge/pattern sharing can alleviate the challenges of a centralized processing approach, and is an attractive approach since the local patterns may in any case be mined for knowledge discovery at each data source independently to discover local trends and for making local decisions.

The above observations encourage the development of pattern discovery algorithms based on local patterns. *Local pattern analysis* is an in-place strategy specifically designed for mining multiple data sources, providing a feasible way to generate globally interesting models from data in multidimensional multi-databases. With local pattern analysis, one can better understand the distribution and inconsistency of local data

*Corresponding author.

patterns, and develop high-performance data mining systems to deal with multiple data sources in which local patterns are fused to make global patterns.

2. Papers in this special issue

This special issue represents the recent advances in local pattern analysis applied to the mining of multiple data sources. The call for papers attracted 20 papers, from which we selected 8 submissions after a thorough review.

2.1. Local pattern analysis

Kum, Chang and Wang (in this issue) present an alternative local mining approach for finding sequential patterns in the local databases within a multi-database. Their main contributions include: (1) the notion of approximate sequential (or consensus) pattern mining is defined as identifying patterns approximately shared by many sequences; such patterns can effectively summarize and represent the local databases by identifying the underlying trends in the data; (2) a novel algorithm, ApproxMAP, is designed for mining approximate sequential patterns, from large sequence databases in two steps: sequence clustering and consensus pattern discovery directly from each cluster through multiple alignments; and (3) a model is proposed to identify both high vote sequential patterns and exceptional sequential patterns from the collection of these consensus patterns from each local database.

Zhu and Wu (in this issue) develop a solution to bridge the local and global analysis for noise cleansing. More specifically, the proposed effort tries to identify and eliminate mislabeled data items from large or distributed datasets through local analysis and global incorporation. For this purpose, they make use of distributed datasets or partition a large dataset into subsets, each of which is regarded as a local subset and is small enough to be processed by an induction algorithm at one time to construct a local model for noise identification. Good rules are identified from each subset, and the good rules are used to evaluate the whole dataset. For a given instance, error count variables are used to count the number of times it has been identified as noise by all data subsets. The instance with higher error values will have a higher probability of being a mislabeled example. Two threshold schemes, majority and non-objection, are used to identify and eliminate the noisy examples.

Ratio rules (RR) are a form of quantitative association mining. Traditional techniques used for ratio rule mining is an eigen-system analysis which can often fall victim to noise. This has limited the application of ratio rule mining greatly. Also, the traditional batch methods for ratio rule mining cannot cope with dynamic data. Yan et al. (in this issue) design an integrated method for mining ratio rules from distributed and dynamic data sources. Specifically, it first mines the ratio rules from each data source separately through a novel robust and adaptive one-pass algorithm (which is called Robust and Adaptive Ratio Rule (RARR)), and then integrates the rules of each data source in a simple probabilistic model. This allows the global rules to be acquired from all the local information sources adaptively.

A peculiarity pattern is a model that is hidden in a small subset of peculiar data. Mining peculiarity patterns can be much more interesting than frequent pattern discovery

in real-world applications. Previous methods for finding peculiar patterns are based on attributes. Zhong, Yao and Liu (in this issue) propose a new strategy for mining peculiarity patterns in multiple relational databases. Peculiar data are first identified on the record level, and then peculiar rules are identified and explained in a relational mining framework.

2.2. Learning from multiple sources

Ling and Yang (in this issue) design a novel method that classifies data from multiple sources without class labels in each source, called CMS (Classification from Multiple Sources). CMS consists of two major steps. In the first step, it partitions the whole training set in each source into clusters such that examples in any cluster belong to the same class. In the second step, CMS generates consistent and more succinct class labeling by a merging algorithm. The main contributions of their work include: (1) CMS can improve the classification accuracy; and (2) from the machine learning perspective, the method removes the fundamental assumption of providing class labels in supervised learning, and bridges the gap between supervised and unsupervised learning.

Efficiently detecting outliers or anomalies is an important problem in many areas of science, medicine and information technology. Applications range from data cleaning to clinical diagnosis, from detecting anomalous defects in materials to fraud and intrusion detection. Over the past decade, researchers in data mining and statistics have addressed the problem of outlier detection using both parametric and non-parametric approaches in a centralized setting. Otey, Ghoting and Parthasarathy (in this issue) propose a tunable algorithm for distributed outlier detection in dynamic mixed-attribute data sets. The main contributions of their work include: (1) defining an anomaly score wherein one can effectively identify outliers in a mixed attribute space by considering the dependencies between attributes of different types; (2) designing a two-pass distributed algorithm for outlier detection based on the anomaly score; and (3) scaling up the two-pass distributed algorithm by approximating and extending it to handle dynamic datasets.

2.3. Web mining

Hu and Zhong (in this issue) advocate a conceptual model with dynamic multi-level workflows corresponding to a mining-grid centric multi-layer grid architecture, for multi-aspect analysis in building an e-business portal on the Wisdom Web. The integrated model assists in dynamically organizing status-based business processes that govern enterprise application integration. Two case studies are presented for demonstrating the effectiveness of the proposed model in the real world. The first case study is about how to organize and mine multiple data sources for behavior-based online customer segmentation, which is the first crucial step for personalization and one-to-one marketing. The second case study is about how to evaluate and monitor data quality, which in return can optimize the knowledge discovery process for intelligent decision making. The proposed methodology attempts to orchestrate various mining agents on the mining-grid for integrating data and knowledge in a unified portal utilizing a service-oriented architecture.

Yang et al. (in this issue) propose a novel categorization algorithm, called the Iterative Reinforcement Categorization Algorithm (IRC), to exploit the full inter-relationships between different types of objects on the Web, including Web pages and queries. Their main contributions include: (1) traditional classification methods are extended to multi-type interrelated data objects; and (2) a reinforcement algorithm is designed for classifying inter-related Web data objects by iteratively reinforcing the individual classification results of different types of objects via their inter-relationships. Experiments on a click through-log dataset from the MSN search engine show the effectiveness of their approach.

References

Hu, J. and Zhong, N. Organizing multiple data sources for developing intelligent e-business portals. Data Mining and Knowledge Discovery, in this issue.

Kum, H., Chang, J., and Wang, W. sequential pattern mining in multi-databases via multiple alignment. Data Mining and Knowledge Discovery, in this issue.

Ling, C. and Yang, Q. Discovering classification from data of multiple sources. Data Mining and Knowledge Discovery, in this issue.

Otey, M., Ghoting, A., and Parthasarathy, S. Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knowledge Discovery, in this issue.

Yang, Q., Xue, G., Shen, D., Zeng, H., Yu, Y., and Chen, Z. Reinforcing web-object categorization through interrelationships. Data Mining and Knowledge Discovery, in this issue.

Yan, J., Liu, N., Yang, Q., Zhang, B., Chen, Q., and Chen, Z. Mining adaptive ratio rules from distributed data sources, in this issue.

Zhang,S., Zhang, C., and Wu, X. 2004. Knowledge discovery in multiple databases. Springer, ISBN: 1-85233-703-6, p. 233.

Zhong, N., Yao, Y., and Liu, C. Relational Peculiarity Oriented Mining, in this issue.

Zhu, X. and Wu, X. Bridging local and global data cleansing: Identifying class noise in large, distributed data datasets. Data Mining and Knowledge Discovery, in this issue.

Shichao Zhang: is a senior research fellow in the Faculty of Information Technology at UTS, Australia, and a chair professor of Automatic Control at BUAA, China. He received his PhD degree in computer science from Deakin University, Australia. His research interests include data analysis and smart pattern discovery. He has published about 35 international journal papers, including 6 in IEEE/ACM Transactions, 2 in Information Systems, 6 in IEEE magazines; and over 40 international conference papers, including 2 ICML papers and 3 FUZZ-IEEE/AAMAS papers. He has won 4 China NSF/863 grants, 2 Australian large ARC grants and 2 Australian small ARC grants. He is a senior member of the IEEE, a member of the ACM, and serving as an associate editor for Knowledge and Information Systems, and The IEEE Intelligent Informatics Bulletin.

Mohammed J. Zaki: is an Associate Professor of Computer Science at RPI. He received his Ph.D. degree in computer science from the University of Rochester in 1998. His research interests focus on developing novel data mining techniques for bioinformatics, performance mining, web mining, and so on. He has published over 100 papers on data mining, co-edited 11 books (including "Data Mining in Bioinformatics, Springer-London, 2005), served as guest-editor for several journals, served on the

program committees of major international conferences, and co-chaired many workshops (BIOKDD, HPDM, DMKD) in data mining. He is currently an associate editor for IEEE Transactions on Knowledge and Data Engineering, action editor for Data Mining and Knowledge Discovery: An Int'l Journal, and editor for Scientific Programming, Int'l Journal of Data Mining and Bioinformatics, Int'l Journal of Data Warehousing and Mining, Int'l Journal of Computational Intelligence, and ACM SIGMOD Digital Symposium Collection. He received the National Science Foundation CAREER Award in 2001 and the Department of Energy Early Career Principal Investigator Award in 2002. He also received the ACM Recognition of Service Award in 2003, and the IEEE Certificate of Appreciation in 2005.