Applied Artificial Intelligence, 17:375–381, 2003

0883-9514/03 $12.00 + .00
DOI: 10.1080/08839510390219264

Taylor & Francis
Taylor & Francis Group

DATA PREPARATION FOR DATA MINING

SHICHAO ZHANG and CHENGQI ZHANG
Faculty of Information Technology, University of Technology, Sydney, Australia

QIANG YANG
Computer Science Department, Hong Kong University of Science and Technology, Kowloon, Hong Kong, China

Data preparation is a fundamental stage of data analysis. While a lot of low-quality information is available in various data sources and on the Web, many organizations or companies are interested in how to transform the data into cleaned forms which can be used for high-profit purposes. This goal generates an urgent need for data analysis aimed at cleaning the raw data. In this paper, we first show the importance of data preparation in data analysis, then introduce some research achievements in the area of data preparation. Finally, we suggest some future directions of research and development.

INTRODUCTION

In many computer science fields, such as pattern recognition, information retrieval, machine learning, data mining, and Web intelligence, one needs to prepare quality data by pre-processing the raw data. In practice, it has been generally found that data cleaning and preparation takes approximately 80% of the total data engineering effort. Data preparation is, therefore, a crucial research topic. However, much work in the field of data mining was built on the existence of quality data. That is, the input to the data-mining algorithms is assumed to be nicely distributed, containing no missing or incorrect values where all features are important. This leads to: (1) disguising useful patterns that are hidden in the data, (2) low performance, and (3) poor-quality outputs. To start with a focused effort in data preparation, this special issue includes twelve papers selected from the First International Workshop on Data Cleaning and Preprocessing (in conjunction with IEEE International

Address correspondence to Shichao Zhang, Faculty of Information Technology, University of Technology, Sydney, P. O. Box 123, Broadway, Sydney, NSW 2007, Australia. E-mail: zhangsc@it.uts.edu.au

Conference on Data Mining 2002 in Maebashi, Japan). The most important feature of this special issue is that it emphasizes practical techniques and methodologies for data preparation in data-mining applications. We have paid special attention to cover all areas of data preparation in data mining.

The emergence of knowledge discovery in databases (KDD) as a new technology has been brought about with the fast development and broad application of information and database technologies. The process of KDD is defined (Zhang and Zhang 2002) as an iterative sequence of four steps: defining the problem, data pre-processing (data preparation), data mining, and post data mining.

Defining the Problem

The goals of a knowledge discovery project must be identified. The goals must be verified as actionable. For example, if the goals are met, a business organization can then put the newly discovered knowledge to use. The data to be used must also be identified clearly.

Data Pre-processing

Data preparation comprises those techniques concerned with analyzing raw data so as to yield quality data, mainly including data collecting, data integration, data transformation, data cleaning, data reduction, and data discretization.

Data Mining

Given the cleaned data, intelligent methods are applied in order to extract data patterns. Patterns of interest are searched for, including classification rules or trees, regression, clustering, sequence modeling, dependency, and so forth.

Post Data Mining

Post data mining consists of pattern evaluation, deploying the model, maintenance, and knowledge presentation.

The KDD process is iterative. For example, while cleaning and preparing data, you might discover that data from a certain source is unusable, or that data from a previously unidentified source is required to be merged with the other data under consideration. Often, the first time through, the data-mining step will reveal that additional data cleaning is required.

Much effort in research has been devoted to the third step: data mining. However, almost no coordinated effort in the past has been spent on the

second step: data pre-processing. While there have been many achievements at the data-mining step, in this special issue, we focus on the data preparation step. We will highlight the importance of data preparation next. We present a brief introduction to the papers in this special issue to highlight their main contributions. In the last section, we summarize the research area and suggest some future directions.

IMPORTANCE OF DATA PREPARATION

Over the years, there has been significant advancement in data-mining techniques. This advancement has not been matched with similar progress in data preparation. Therefore, there is now a strong need for new techniques and automated tools to be designed that can significantly assist us in preparing quality data. Data preparation can be more time consuming than data mining, and can present equal, if not more, challenges than data mining (Yan et al. 2003). In this section, we argue for the importance of data preparation at three aspects: (1) real-world data is impure; (2) high-performance mining systems require quality data; and (3) quality data yields high-quality patterns.

1. Real-world data may be incomplete, noisy, and inconsistent, which can disguise useful patterns. This is due to:
   - Incomplete data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.
   - Noisy data: containing errors or outliers.
   - Inconsistent data: containing discrepancies in codes or names.
2. Data preparation generates a dataset smaller than the original one, which can significantly improve the efficiency of data mining. This task includes:
   - Selecting relevant data: attribute selection (filtering and wrapper methods), removing anomalies, or eliminating duplicate records.
   - Reducing data: sampling or instance selection.
3. Data preparation generates quality data, which leads to quality patterns. For example, we can:
   - Recover incomplete data: filling the values missed, or reducing ambiguity.
   - Purify data: correcting errors, or removing outliers (unusual or exceptional values).
   - Resolve data conflicts: using domain knowledge or expert decision to settle discrepancy.

From the above three observations, it can be understood that data pre-processing, cleaning, and preparation is not a small task. Researchers and practitioners must intensify efforts to develop appropriate techniques for

efficiently utilizing and managing the data. While data-mining technology can support the data-analysis applications within these organizations, it must be possible to prepare quality data from the raw data to enable efficient and quality knowledge discovery from the data given. Thus, the development of data-preparation technologies and methodologies is both a challenging and critical task.

DESIRABLE CONTRIBUTIONS

The papers in this special issue can be categorized into six categories: hybrid mining systems for data cleaning, data clustering, Web intelligence, feature selection, missing values, and multiple data sources.

Part I designs hybrid mining systems to integrate techniques for each step in the KDD process. As described previously, the KDD process is iterative. While a significant amount of research aims at one step in the KDD process, it is important to study how to integrate several techniques into hybrid systems for data-mining applications. Zhang et al. (2003) propose a new strategy for integrating different diverse techniques for mining databases, which is particularly designed as a hybrid intelligent system using multi-agent techniques. The approach has two distinct characteristics below that differentiate this work from existing ones.

- New KDD techniques can be added to the system and out-of-date techniques can be deleted from the system dynamically.
- KDD technique agents can interact at run-time under this framework, but in other non-agent based systems, these interactions must be decided at design-time.

The paper by Abdullah et al. (2003) presents a strategy for covering the entire KDD process for extracting structural rules (paths or trees) from structural patterns (graphs) represented by Galois Lattice. In this approach, symbolic learning in feature extraction is designed as a pre-processing (data preparation) and a sub-symbolic learning as a post-processing (post-data mining). The most important contribution of this strategy is that it provides a solution in capturing the data semantics by encoding trees and graphs in the chromosomes.

Part II introduces techniques for data clustering. Tuv and Runger (2003) describe a statistical technique for clustering the value-groups for high-cardinality predictors such as decision trees. In this work, a frequency table is first generated for the categorical predictor and the categorical response. And then each row in the table is transformed to a vector appropriate for clustering. Finally, the vectors are clustered by a distance-based clustering algorithm. The clusters provide the groups of categories for the predictor

attribute. After grouping, subsequent analysis can be applied to the predictor with fewer categories and, therefore, lower dimension or complexity.

Part III deals with techniques for Web intelligence. While a great deal of information is available on the Web, a corporation can benefit from intranets and the Internet to gather, manage, distribute, and share data inside and outside the corporation. This generates an urgent need for effective techniques and strategies that assist in collecting and processing useful information on the World Wide Web. Yang et al. (2003) advocate that in Web-log mining, although it is important to clean the raw data, it is equally important to clean the mined association rules. This is because these rules are often very large in number and difficulty to apply. In their paper, a new technique is introduced, which refers to sequential classifiers for predicting the search behaviors of users. Their tree-based prediction algorithm leads to efficient as well as accurate prediction on users' next Web page accesses. While the sequential classifiers predict the users' next visits based on their current actions using association analysis, a pessimistic selection is also presented for choosing among alternative predictions. Li et al. (2003) design a Web data mining and cleaning strategy for information gathering, which combines distributed agent computation with information retrieval. The goal of this paper is to eliminate most of the "dirty data" and the irrelevant information from the documents retrieved. Saravanan et al. (2003) design a high-level data cleaning framework for building the relevance between text categorization and summarization. The main contribution of this framework is that it effectively applies Katz's K-mixture model of term distribution to the summarization tasks.

Part IV presents two feature selection methods. Ratanamahatana and Gunopulos (2003) gives an algorithm that uses C4.5 decision trees to select features. The purpose is to improve Naive Bayesian learning method. The algorithm uses 10% of a training set to build an initial decision tree. Hruschka et al. (2003) introduce a Bayesian approach for features selection, where a clustering genetic algorithm is used to find the optimal number of classification rules. In this approach, a Bayesian network is first generated from a dataset. And then the Markov blanket of the class feature is used for the feature subset selection. Finally, a genetic algorithm for clustering is used to extract classification rules. Zhang and Ling's work (2003) addresses the theoretical issue of mapping nominal attributes into numerical ones when applying the Naive Bayes learning algorithm. Such mappings can be viewed as a part of discretization in data preparation. They show that the learnability of Naive Bayes is affected by such mappings applied. Their work helps us to understand the influence of numeric mappings on the property of nominal functions and on the learnability of Naive Bayes.

Part V includes methods for filling missing values. Batista and Monard (2003) propose a new approach to deal with missing values. They analyze the behavior of three methods for missing data treatment: the 10-NNI method

using a k-nearest neighbor for imputation; the mean or mode imputation; and the internal algorithms used by C4.5 and CN2 to treat missing data. They conclude that the 10-NNI method provides very good results, even for training sets having a large amount of missing data. Tseng et al. (2003) present a new strategy, referred to as Regression and Clustering, for constructing minimum error-based intra-feature metric matrices to provide distance measure for nominal data so that metric or kernel algorithms can be effectively used, which is very important for data-mining algorithms to deal with all possible applications. The approach can improve the accuracy of the predicted values for the missing data.

Part VI tackles multiple databases. It is necessary to collect external data for some organizations, such as nuclear power plants and earthquake bureaus, which have very small databases. For example, because accidents in nuclear power plants cause many environmental disasters and create economical and ecological damage as well as endangering people's lives, automatic surveillance and early nuclear accident detection have received much attention. To reduce nuclear accidents, we need trusty knowledge for controlling nuclear accidents. However, a nuclear accident database often contains too little data to form trustful patterns. Thus, mining the accident database in the nuclear power plant must depend on external data. Yan et al. (2003) draw a new means of separating external and internal knowledge of different data sources and use relevant belief knowledge to rank the potential facts. Instead of common data cleaning works, such as removing errors and filling missing values, they use a pre- or post-analysis to evaluate the relevance of identified external data sources to the data-mining problem. This paper brings out a novel way to consider data pre-processing work in data mining.

CONCLUSION AND FUTURE WORK

Data preparation is employed today by data analysts to direct their quality knowledge discovery and to assist in the development of effective and high-performance data analysis application systems. In data mining, the data preparation is responsible for identifying quality data from the data provided by data pre-processing systems. Indeed, data preparation is very important because: (1) real-world data is impure; (2) high-performance mining systems require quality data; and (3) quality data yields concentrative patterns.

In this paper, we have argued for the importance of data preparation and briefly introduced the research into data preparation, where the details of each achievement can be found in this special issue. By way of summary, we now discuss the possible directions of data preparation.

The diversity of data and data-mining tasks deliver many challenging research issues for data preparation. Below we would like to suggest some future directions for data preparation:

- Constructing of interactive and integrated data mining environments.
- Establishing data preparation theories.
- Developing efficient and effective data-preparation algorithms and systems for single and multiple data sources while considering both internal and external data.
- Exploring efficient data-preparation techniques for Web intelligence.

REFERENCES

Abdullah, N., M. Liquière, and S. A. Cerri. 2003. GAsRule for knowledge discovery. *Applied Artificial Intelligence* 17(5–6):399–417.

Batista, G., and M. Monard. 2003. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence* 17(5–6):519–533.

Hruschka, E., Jr., E. Hruschka, and N. Ebecken. 2003. A feature selection Bayesian approach for extracting classification rules with a clustering genetic algorithm. *Applied Artificial Intelligence* 17(5–6):489–506.

Li, Y., C. Zhang, and S. Zhang. 2003. Cooperative strategy for Web data mining and cleaning. *Applied Artificial Intelligence* 17(5–6):443–460.

Ratanamahatana, C., and D. Gunopulos. 2003. Feature selection for the Naive Bayesian classifier using decision trees. *Applied Artificial Intelligence* 17(5–6):475–487.

Saravanan, M., P. Reghu Raj, and S. Raman. 2003. Summarization and categorization of text data in high level data cleaning for information retrieval. *Applied Artificial Intelligence* 17(5–6):461–474.

Tseng, S., K. Wang, and C. Lee. 2003. A pre-processing method to deal with missing values by integrating clustering and regression techniques. *Applied Artificial Intelligence* 17(5–6):535–544.

Tuv, E., and G. Runger. 2003. Pre-processing of high-dimensional categorical predictors in classification settings. *Applied Artificial Intelligence* 17(5–6):419–429.

Yan, X., C. Zhang, and S. Zhang. 2003. Towards databases mining: Pre-processing collected data. *Applied Artificial Intelligence* 17(5–6):545–561.

Yang, Q., T. Li, and K. Wang. 2003. Web-log cleaning for constructing sequential classifiers. *Applied Artificial Intelligence* 17(5–6):431–441.

Zhang, C., and S. Zhang. 2002. Association Rules Mining: Models and Algorithms. In *Lecture Notes in Artificial Intelligence*, volume 2307, page 243, Springer-Verlag.

Zhang, H., and C. Ling. 2003. Numeric mapping and learnability of Naïve Bayes. *Applied Artificial Intelligence* 17(5–6):507–518.

Zhang, Z., C. Zhang, and S. Zhang. 2003. An agent-based hybrid framework for database mining. *Applied Artificial Intelligence* 17(5–6):383–398.