Expert Systems with Applications 38 (2011) 6618–6626

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data mining and preprocessing application on component reports of an airline company in Turkey

Feyza Gürbüz [a,*], Lale Özbakir [a], Hüseyin Yapici [b]

[a] Department of Industrial Engineering, University of Erciyes in Kayseri, Kayseri 38039, Turkey
[b] Department of Mechanical Engineering, University of Erciyes in Kayseri, Kayseri 38039, Turkey

ARTICLE INFO

Keywords:
Data mining
Preprocessing
Rough sets
Find laws

ABSTRACT

Risk and safety have always been important considerations in aviation. With the rapid growth in air travel, flight delays, cancellations and incidents/accidents have also dramatically increased in recent years (Nazeri & Jianping, 2002). There is a large amount of knowledge and data accumulation in aviation industry. These data could be stored in the form of pilot reports, maintenance reports, incident reports or delay reports. This paper focuses on different preprocessing and feature selection techniques applied on the 15 component reports of an airline company in Turkey to understand and clean the data set. Regression analysis, anomaly detection analysis, find dependencies and rough sets are used in this study in order to reduce the data set. Also the classification techniques of data mining are used to predict the warning level of the component as the class attribute. For this purpose Polyanalyst, SPSS Clementine, Minitab and Rosetta software tools are used. Find laws module of Polyanalyst is used to find the relations and information retrieval about the components warning level.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Data mining methods have been successfully applied to different fields. Aviation industry is one of these fields. With the rapid growth in air travel, flight delays, cancellations and incidents have also dramatically increased in recent years (Nazeri & Jianping, 2002). As a result of this, there is a large amount of knowledge and data accumulation in aviation industry. These data could be stored in the form of pilot reports, maintenance reports, incident reports, component reports or delay reports. This paper explains the preprocessing and data mining application on the component reports of an airline company in Turkey.

Nowadays the analysis of such data is automatically conducted and analysts have difficulties in dealing with the growing data efficiently and on time.

In conclusion, in the automatic and smart analysis of the complexed structure high volume data in aviation industry capable instruments are needed. Data mining-one of those instruments – which was not known before and is potentially useful is an instrument used for to reveal the information hidden in the data (Jiawei & Kamber, 2001).

The science of extracting useful information from large data sets or databases is known as data mining. It is a new discipline, lying at the intersection of statistics, machine learning, data management and databases, pattern recognition/artificial intelligence, and other areas (Hand, Manila, & Smyth, 2001).

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns (Hand et al., 2001).

Data mining, also popularly referred to as knowledge discovery in databases (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories (Jiawei & Kamber, 2001).

The KDD process is outlined in Fig. 1 (Dunham, 2002). It is interactive and iterative involving, more or less, the following steps (Mitra & Acharya, 2003):

1. *Understanding the application domain:* This includes relevant prior knowledge and goals of the application.
2. *Extracting the target data set:* This is nothing but selecting a data set or focusing on a subset of variables, using feature ranking and selection techniques.
3. *Data preprocessing:* This is required to improve the quality of the actual data for mining. This also increases the mining efficiency by reducing the time required for mining the preprocessed data. Data preprocessing involves data cleaning, data transformation,

[*] Corresponding author.
E-mail addresses: feyza@erciyes.edu.tr (F. Gürbüz), lozbakir@erciyes.edu.tr (L. Özbakir), yapici@erciyes.edu.tr (H. Yapici).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.11.076

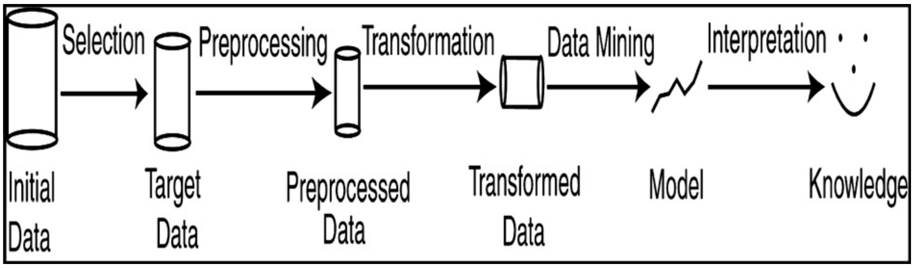

Fig. 1. The kdd process.

data integration, data reduction or data compression for compact representation, etc.

(a) *Data cleaning:* It consists of some basic operations, such as normalization, noise removal and handling of missing data, and reduction of redundancy. Data from real-world sources are often erroneous, incomplete, and inconsistent, perhaps due to operational error or system implementation flaws. Such low-quality data need to be cleaned prior to data mining.

(b) *Data integration:* Integration plays an important role in KDD. This operation includes integrating multiple, heterogeneous data sets generated from different sources.

(c) *Data reduction and projection:* This includes finding useful features to represent the data (depending on the goal of the task) and using dimensionality reduction, feature discretization, and feature extraction (or transformation) methods. Application of the principles of data compression can play an important role in data reduction and is a possible area of future development, particularly in the area of knowledge discovery from multimedia data set.

4. *Data mining:* Data mining constitutes one or more of the following functions, namely, classification, regression, clustering, summarization, image retrieval, discovering association rules and functional dependencies, and rule extraction.
5. *Interpretation:* This includes interpreting the discovered patterns, as well as the possible (low-dimensional) visualization of the extracted patterns. Visualization is an important aid that increases understandability from the perspective of humans. One can evaluate the mined patterns automatically or semiautomatically to identify the truly interesting or useful patterns for the user.
6. *Using discovered knowledge:* It includes incorporating this knowledge into the performance system and taking actions based on the knowledge.

2. Feature selection and preprocessing

Huge data sets have grown increasingly large in terms of number of dimensions and number of instances. Reducing the number of dimensions by selecting variables or features is effective in dealing with high-dimensional data.

Variable and feature selection have became the focus of much research in areas of application for which data sets with tens or hundreds of thousands of variables are available. The objective of variable selection is threefold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data (Guyon & Elisseeff, 2003).

Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points.

As shown in Fig. 1, two of the stages in kdd process are feature selection and preprocessing stages. A feature selection is usually meant as a process of finding a subset of features from the original set of features forming patterns in a given data set, optimal according to the defined goal and criterion of feature selection (Suraj & Delimata, 2006). There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These data processing techniques, when applied prior to mining, can substantially improve the overall data mining results (Jiawei & Kamber, 2001).

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. There are a number of strategies for data reduction. These include data aggregation (e.g., building a data cube), dimension reduction (e.g., removing irrelevant attributes through correlation analysis), data compression (e.g., using encoding schemes such as minimum length encoding or wavelets), and numerosity reduction (e.g., replacing the data by alternative, smaller representations such as clusters, or parametric models). Data can also be reduced by generalization, where low level concepts such as city for customer location are replaced with higher level concepts, such as region or province or state (Jiawei & Kamber, 2001).

In summary, real world data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is therefore an important step in the knowledge discovery process, since quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge pay-offs for decision making (Jiawei & Kamber, 2001).

3. Related work

Data preprocessing or preparation is an important and critical step in the data mining process and it has a huge impact on the success of a data mining project (Hu, 2003). 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 (Zhang, Zhang, & Yang, 2003). In their study they show the importance of data preparation in data analysis, and introduce some research achievements in the area of data preparation. Finally, they suggest some future directions of research and development.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات