Journal of Engineering and Applied Sciences 12 (16): 4102-4107, 2017
ISSN: 1816-949X

Review of Data Preprocessing Techniques in Data Mining

Suad A. Alasadi and Wesam S. Bhaya
College of Information Technology, University of Babylon, Babil, Iraq

Abstract: Data mining is the process of extraction useful patterns and models from a huge dataset. These models and patterns have an effective role in a decision making task. Data mining basically depend on the quality of data. Raw data usually susceptible to missing values, noisy data, incomplete data, inconsistent data and outlier data. So, it is important for these data tobe processed before being mined. Preprocessing data is an essential step to enhance data efficiency. Data preprocessing is one of the most data mining steps which deals with data preparation and transformation of the dataset and seeks at the same time to make knowledge discovery more efficient. Preprocessing include several techniques like cleaning, integration, transformation and reduction. This study shows a detailed description of data preprocessing techniques which are used for data mining.

Key words: Data mining, data preprocessing, data set, KDD (Knowledge Discovery in Databases), dataset, pattein

INTRODUCTION

Knowledge Discovery in Databases (KDD) is a process of extraction valuable information from huge data sources. Data mining is a step of KDD which is performs analysis and models for huge dataset using classification, clustering, association rules and many other techniques (Rajaraman and Ullman, 2011). The raw data are highly vulnerable to missing, noise, outliers and inconsistent because of their huge size, multiple resources and their gathering methods. The poor quality data will effect on data mining results. Therefore, preprocessing technique must be applied on data to improve the efficiency of these data. Figure 1 shows steps of KDD process from dataset (Tomar and Agarwal, 2014; Larose, 2006).

As shown is Fig. 1, the data must be selected to determine the target data then the selected data must be preprocessed to enhance its reliability. After preprocessed the data, it must be transformed to suitable form for data mining process. Then, the mining procedure will be applied such as clustering, classification, regression, etc., in order to extract patterns which is interrupted and evaluated in the final step.

MATERIALS AND METHODS

Preprocessing techniques: Data preprocessing is one of the most data mining tasks which includes preparation and transformation of data into a suitable form to mining procedure. Data preprocessing aims to reduce the data size, find the relations between data, normalize data, remove outliers and extract features for data. It includes several techniques like data cleaning, integration, transformation and reduction. Figure 1 and 2 show data preprocessing categories (Tamilselvi *et al*., 2015).

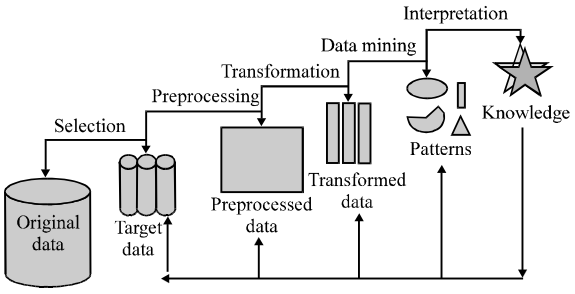

Fig.1: Knowledge discovery steps

Data cleaning: Row data may have incomplete records, noise values, outliers and inconsistent data. Data cleaning is a first step in data preprocessing techniques which is usedto find the missing values, smooth noise data, recognize outliers and correct inconsistent. These dirty data will effects on miming procedure and led to unreliable and poor output. Therefore, it is important for some data-cleaning routines to be used. Table 1 shows an example of dirty data (Maingi, 2015).

Missing values: If there are records with un recorded values for its records then these values may be filled using the following ways.

Corresponding Author: Suad A. Alasadi, College of Information Technology, University of Babylon, Babil, Iraq

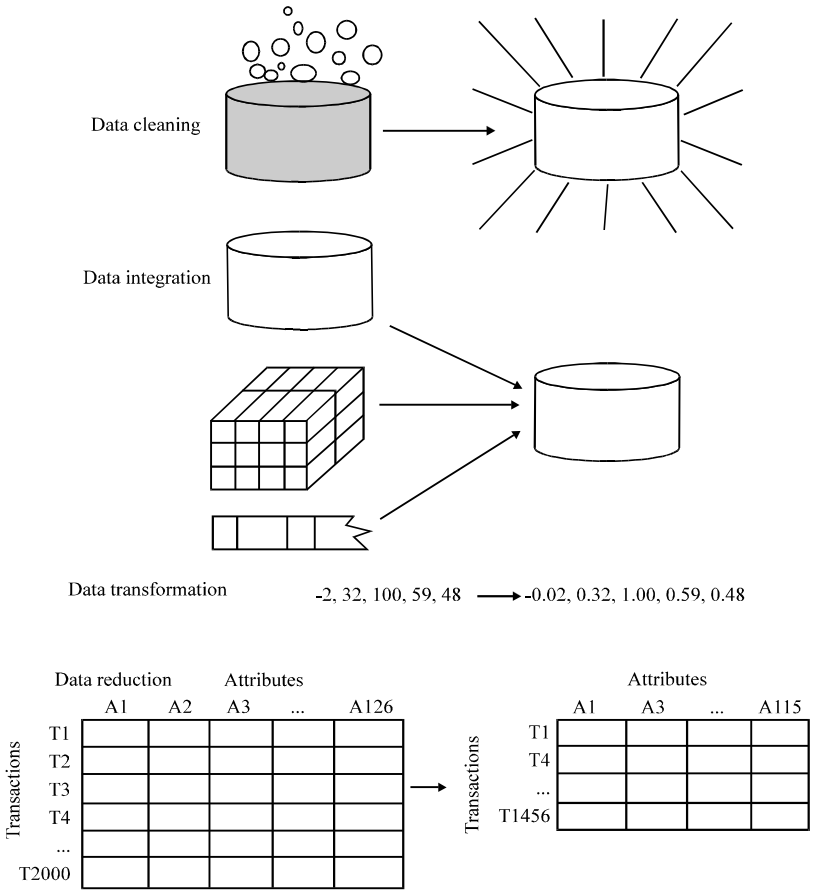

Fig. 2: Preprocessing forms

Table 1: Dirty data

Dirty data	Problems
Gender = S	Wrong value
Address = 00	Incomplete record
C1_name = Rose M	Duplicate record
C2_name = R. Mohan	
Name = Rose 15-10-2015	Multiple values in single column

Ignore the tuple: This choice is selected when the value of class label is not existing (it is used with classification mining task). This method is not effective but it is used when the tuple haveseveral attributes with empty values.

Fill the missing value manually: This approach in general requires human effort and time consuming. It cannot be used with the large size of dataset.

Use a global constant to fill the missing value: This method works by replacing missing values of attribute by a particular constant which is similar for all records for example using "Unknown" as a label. This method have problems because when the missing values are replaced by a specific term "Unknown" as an example, the mining programs may believe that they form an importantconcept, since they have a common value.

Use the attribute mean to fill the missing value: This method works by replacing the missing value for a particular attribute by the average value for that attribute.

Use the attribute mean for all samples belonging to the same class as the given tuple: For example, if we classifyusersdepending on credit risk, the missing value can be replaced by the average value of income for the users which belong to the similar credit risk class for a given tuple.

Use the most probable value to fill the missing value: This approach is used with techniques like inference based regression using a decision tree induction or Bayesian formalism.

Noise data: One of the most problems which effects on mining process is noise. Noise is a random error or variance in a measured variable. Noise data means that there is an error in data or outliers which deviates from the normal. It can be corrected using the following methods (Han and Kamber, 2006):

Binning: This method works on smoothing stored data based on its "neighborhood" which is the values around it. The sorted values are divided into a number of "buckets" or bins. Since, these methods depend on the neighbor's data thus, they perform local smoothing.

In smoothing by bin boundaries, the min and max values in each bin are determined asbin boundaries. Then each value is replaced by the closest boundary value. In general whenever thewidth of bucket is larger, the effect of smoothing is greater. Alternatively, in the case of equal width bucket where the interval range of values in each bucket is the same, binning is used as a discretization technique. Figure 3 illustrates binning techniques.

Binning techniques: # Stored data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34 .#.

Partition into (equal-depth) bins:

- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34

 # Smoothing by bin means:

- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

 # Smoothing by bin boundaries:

- Bin 1: 4, 4, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 25, 34

This figure shows data for price in dollars where the data are sorted and distributed into equal bin width . In case of smoothing by bin means, each value in a bin is replaced by bin mean. For example, the mean values of 4, 8 and 15 in Bin 1 is 9. Thus, each value in the bin is replaced by 9.

Regression: This methods moothie's data by fitting it to a function. Linear regressionas example includes determining the best line to fit two variables (or attributes), so that each attribute can be used to predict the other.

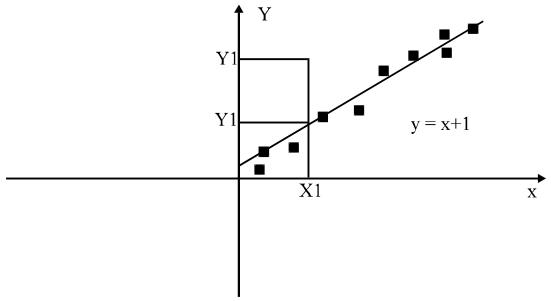

Fig. 3: Regression

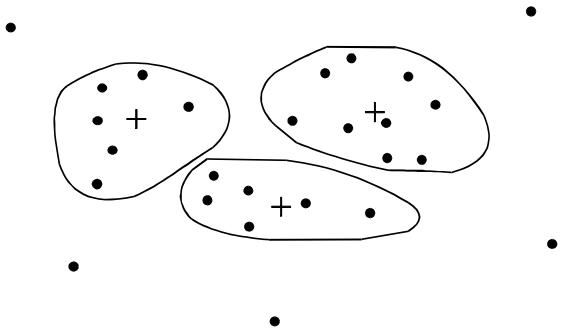

Fig. 4: Clustering

Multi-linear regressionis an expansion to linear regression. It involvestwo or more variables and hence fits data to multi-dimensional domain. Using regression to fit data by finding a mathematical equation may be used to smooth the noise data. Figure 4 explains linear regression.

Clustering: Clustering is defined as grouping set of points into clusters according to a distance measure. The result of clustering is set of clusters each cluster will have set of points with small distance from one another and with large distance from other clusters. This technique can detect outliers, since it grouping similar points into a cluster while points that fall out the clusters are consideredas outlier points. An example of clustering technique. As shown in this figure, there are 3 clusters and the point which are not belong to any clusters are outliers.

RESULTS AND DISCUSSION

Data integration: This technique works by combining data from multi and various resources intoone consistentdata store, like in data warehouse. These resources canhave multi database, files or data cubes. In data integration There are a number of issues for consideration, like Schema integration, object matching

and redundancy which are an important aspects. Each attribute like "annual revenue" is considered as redundant if it "derived" from another attribute or set of attributes. In consistencein attribute or dimension is another form of redundancies. Correlation analysis can be used to detect some redundancies. The correlation between two variables can measure how the attributes can implyone the other strongly. The correlation between (X, Y) attributes can be evaluated by finding the correlation coefficients (Dharmarajan and Vijayasanthi, 2015).

Data transformation: It includes transforming the data to forms suitable for mining process. It involves the following (Baskar *et al*., 2013):

Smoothing: It removes noise from data. It includes techniques such as clustering, regression and binning.

Aggregation: It is the process of applying statistical metrics like means, median and variance which are necessary to summarize the data. The resulted aggregated data are used in data mining algorithms. For example, apply aggregation on the daily sales to compute monthly and annual sales.

Generalization: It includes replacing the lower level data (primitive) by higher level using hierarchical concepts. An example, street which is a type of categorical attributes may be replaced to city or country which ishigh level terms. Another example, age which is a type of numeric concepts can be mapped to senior, younger and youth which are high level concepts.

Normalization: This method works by a adjusting the data values into a specific range such as between 0-1 or -1-1. This method is useful for mining techniques like classification, artificial neural networks and clustering algorithms. Using the normalization to scale the data attributes in tanning face for back propagationneural network algorithm can be used to speed the learning stage. Minimum-maximum, z-score and decimal scaling are popular forms of normalization.

Data reduction: These techniques can be used to reduce the representation of dataset in smaller volume with respect to maintain the integrity of the original dataset. Thus a better data results can be obtained from applying mining techniqueson that reduce data. The following subsection shows data reduction strategies (Garcia *et al*., 2016):

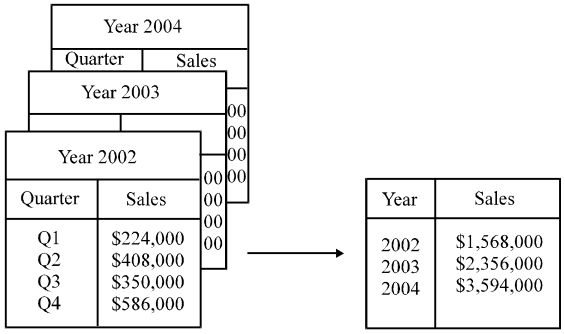

Fig. 5: Data cube aggregation

Data cube aggregation: This approach construct data cube by applying operations of aggregation on data without losing the necessary information for the data analysis (Fig. 5).

As show in Fig. 5, the left side explains sales per quarter for years 2002, 2003, 2004 while the right side, data areaggregated to obtain annual sales.

Attribute subset selection: It reduce the dataset size by removing redundant features or dimensions and irrelevant attributes.

Dimensionality reduction: Also known as (data compression) it uses the mechanisms of encoding to reduce the size of dataset. Reduction can be lossless and lossy based on the retrieved information from decoding. Wavelet Transforms and Principal Component Analysis (PCA) are two effected methods for lossyreduction (Larose, 2006).

Numerosity reduction: It replaces or mapped data to an alternative or smaller representation of data. It consistsof parametricand non-parametric models. The first model needs to store only the parametricof model without storing whole data. While in second model, it includes techniques such as sampling, histogram and clustering.

Data discretization and concept hierarchy generation: These techniques can be used to replace the attributes data values by high level of conceptual or interval ranges. It is a type of numerosity reduction which is very useful for the generation of hierarchal automatically. One of the important tools of data mining is hierarchical and discretization which are perform mining in multi abstraction levels. Data discretization can be classified based on how it performed into supervised or un supervised discretization If it uses class label otherwise, it is top-down or bottom-up discretization. It is consist of the following techniques:

Binning: It is a splitting top to down technique which is depend on the determined bins number. Binning methods which are used for data smoothing are also used in discretization and hierarchy generation. Equal-width or equal-frequency binning can be used to discretize the values of attribute by replacing bin value by mean or median as in smoothing. This process can recursively repeated to produce hierarchy concept. It is unsupervised technique because it don't use class label. Itdepends on the user specification for bin numbers.

Histogram: It is one of the unsupervised techniquesthat does not use class label. It distributes attributes values into ranges (buckets). The values are divided into equal ranges in equal width histogram while in equal frequency histogram, each part has the similar number of data. The algorithm may be repeated reclusively to form multiple level hierarchies. Figure 6 shows an example of histogram. As shown in Fig. 6, each bucket has one pair of price (values, frequency)

Entropy-based: It is one of the popular tools for discretization data. It is presented by Shannon during his study about information theory. It is top down and supervised technique (Zhou and Tao, 2011). It uses class information to reduce the size of data. To discretize a numerical attribute X, it must choose the value of X with minimum entropy as a splitting point this step is repeated recursively to get hierarchical discretization (Holzinger *et al*., 2014).

For example If we have set of tuples with number of attributes and two class C1, C2 for attribute X. To discretize tuples D to an attribute X we hopes that all tuples with class C1 in one part and tuples of C2 class fall in another class. The expected informationto classify a tuple in D based on X is given by Okafor (2005):

$$\text{Info}_A(D) = |D_1|/|D|\ \text{Entropy}(D_1) + |D_2|/|D|\ \text{Entropy}(D_2)$$

D_1 and D_2 are tuples in D, |D|: number of tuples in D. The entropy of D_1 is:

$$\text{Entropy}(D_1) = -\sum_{i=1}^{m} p_i \log_2 (p_i)$$

where, p_i is the probability of C_i in D_1

Clustering: It one of the a mostmethod for data discretization. The algorithms of clustering can be used to discretize a numeric attributes X by dividing the values of attributeto groups or clusters. It produces high results of discretization. It can be classified into either top-down

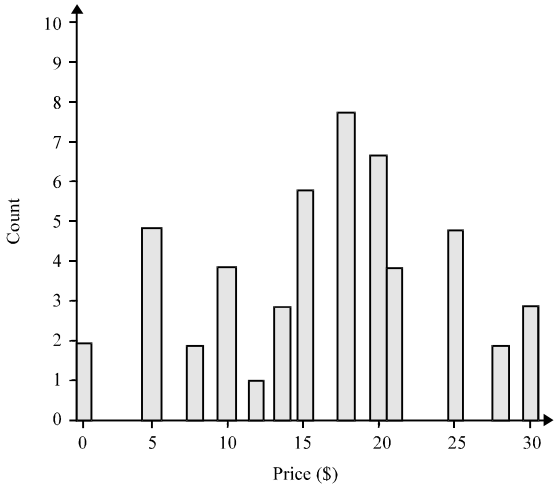

Fig. 6: Histogram

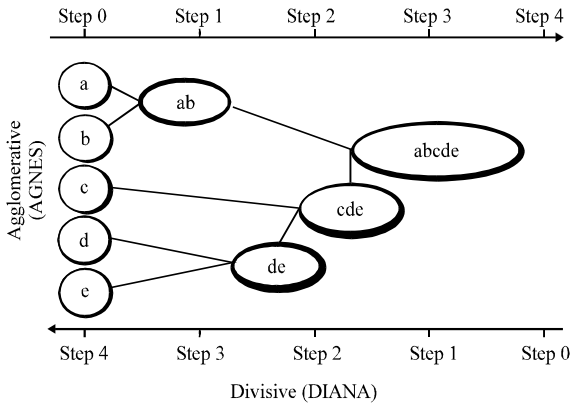

Fig. 7: Clustering strategies

splitting or bottom up merging strategy. In the first type, each cluster which forms anode can be furthersplit into sub clusters, forming low level of hierarchy. While in the second type, clusters are produced by merging the neighbors clusters to form high level concept as shown in Fig. 7 (Rajaraman and Ullman 2011).

CONCLUSION

Real-world data tend to be incomplete, inconsistent, noisy and missing, data preprocessing is one of the important matters for both data warehousing and data mining, data preprocessing includes data cleaning, data integration, data transformation and data reduction. This study explainsan overview on data preprocessing method and their examples. The goal of data preprocessing is given to qualities data for any type ofmining like that data mining, text mining and

web mining. Data cleaning method are used to remove the noisy data, completedon uncompleted data and remove unnecessary data. Datainte gration method is integrated to different source of data inone place. Data transformation method change forms of data and data reduction reduce the volume of database byschema integration. We conclude that data preprocessing techniques have an efficient, effective and important role in preparation, analysis, process large data-scale.

REFERENCES

Baskar, S.S., L. Arockiam and S. Charles, 2013. A systematic approach on data pre-processing in data mining. Compusoft, 2: 335-339.

Dharmarajan, R. and R. Vijayasanthi, 2015. An overview on data preprocessing methods in data mining. Intl. J. Sci. Res. Dev., 3: 3544-3546.

Garcia, S., G.S. Ramirez, J. Luengo, J.M. Benitez and F. Herrera, 2016. Big data preprocessing: Methods and prospects. Big Data Anal., 1: 1-9.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publisher, San Fransisco, USA., ISBN-13: 978-1558609013, Pages: 800.

Holzinger, A., M. Hortenhuber, C. Mayer, M. Bachler and S. Wassertheurer *et al.*, 2014. On Entropy-Based Data Mining. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Holzinger, A. and J. Igor (Eds.). Springer, Berlin, Germany, ISBN:978-3-662-43967-8, pp: 209-226.

Larose, D., 2006. Data Mining Methods and Models. John Wiley and Sons, New York, USA., ISBN-10: 0471666564, pp: 227-229.

Maingi, M.N., 2013. Survey on data preprocessing concept applicable in data mining: Mining. Intl. J. Sci. Res., 4: 1901-1902.

Okafor, A., 2005. Entropy based techniques with applications in data mining. Ph.D Thesis, University of Florida, Gainesville, Florida.

Rajaraman, A. and J.D. Ullman, 2011. Mining of Massive Datasets. Cambridge University Press, UK., ISBN-13: 978-1107015357, Pages: 326.

Tamilselvi, R., B. Sivasakthi and R. Kavitha, 2015. An efficient preprocessing and postprocessing techniques in data mining. Intl. J. Res. Comput. Appl. Rob., 3: 80-85.

Tomar, D. and S. Agarwal, 2014. A survey on pre-processing and post-processing techniques in data mining. Intl. J. Database Theory Appl., 7: 99-128.

Zhou, M.J. and J.C. Tao, 2011. An outlier mining algorithm based on attribute entropy. Procedia Environ. Sci., 11: 132-138.