2nd International Conference on Current Research Trends in Engineering and Technology
© 2018 IJSRSET | Volume 4 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Review of Data Pre-processing Techniques for Classification

Trilok Suthar, Digvijaysinh Mahida, Pinkal Shah
Asst. Prof IT Department, Sigma Institute of Engineering, Vadodara, Gujarat, India

ABSTRACT

Data mining is the process of extraction useful patterns from a huge dataset. These models and patterns have an effective role in a decision making task. Data mining basically depends 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 to be pre-processed before being classified using classification algorithms. Data pre-processing is one of the most data mining steps which deals with data preparation and transformation. Preprocessed data make knowledge discovery more efficient. Preprocessing includes several techniques like cleaning, integration, transformation and data reduction. This paper shows the various preprocessing techniques applied to the classification.

Keywords : Data Mining, Noise, Preprocessing Dataset, Skewness

I. INTRODUCTION

Data preprocessing is an often ignored but important step in data mining process. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining. Data collection methods are often loosely controlled, resulting in incorrect values, wrong data combinations, missing values etc. improper analysing of data can produce misleading results. Thus before running an analysis the quality and representation of data is the first step.

If there is noisy, unreliable and irrelevant data then classification may result in poor accuracy. In order to improve the results of classification the quality of row data is improved by preprocessing of data. Data preprocessing is one of the most important step in data mining process which includes preparation and transformation of dataset.

II. DATA PREPROCESSING TECHNIQUES

Raw data is prone to noise, missing value and irrelevant data. The results depend on the quality of data. In order to prepare well quality of data prepressing is required. Data preprocessing techniques are divided into following categories.

- Data cleaning
- Data integration
- Data transformation
- Data reduction

Data cleaning includes missing value replacement, find the outliers and remove the noisy data.

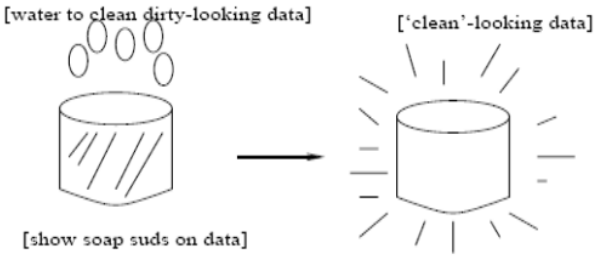

Figure 1: data cleaning

Data reduction involves feature selection. Data reduction techniques helps to reduce the volume or reduce the dimensions without affecting the quality of data. The reduced data set should be more effective and produce the same analytical results.

Figure 2: data reduction

Data integration integrates data from multiple sources like flat files, multiple databases, data cubes. [1] while integration the data no of problems should be addressed. Schema integration can be tricky. Eg. Student_id in one database and student_enrol may refer the same entity.

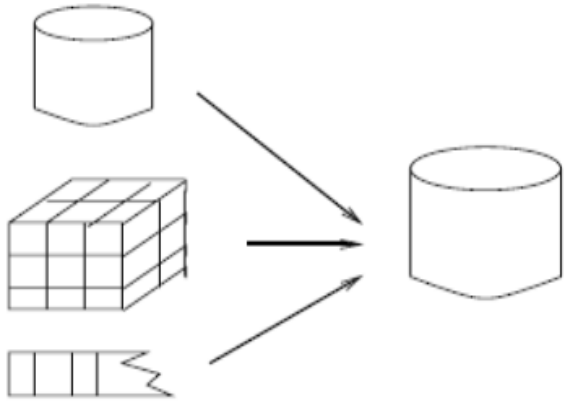

Figure 3: Integration

Data Transformation transforms the data into appropriate form based on the methods. It normally involves normalization, smoothing techniques.

Missing value:

The basic method of deal with missing value is to replace missing value with mean value for numerical data [2]

Sally McClean et al. invented a technique to replace the missing value by making rules establish on background knowledge but still lose some usable rule[3]

Jau ji shen et al. favoured rule recycle bin technique which rehash and compose the rules to receive further outright attributes value association rule which empower the database reborn to prior the veracity and integration rate and progress the validity of missing value completion [4]

Thomas et al. proposed an existing fuzzy rule induction algorithm can consolidate missing values in the training method in avery common way without any need for replacement of missing values [5]

Mie ling shyu et al. designed a framework called F-DCS for replacing missing value which obtains the basic concept of conditional probability approach. This framework can manage both nominal and numeric values with a high certainty [6]

Olga et al. implementd three methods named a singular value decomposition (svd) based method, weighted k-nearest neighbours (K-NN] and row average. K-NN is better than SVD [7]

The missing value in the dataset can influence the performance of the classification process and it become difficult to extract the useful data from datsets.

Anjana sharma et al. proposed three methods named lit wise deletion, K-NN imputation and mean/mode imputation. After applied those methods on classification algorithm and by comparing the K-NN performs better than other two [8]

R malaryuzhi et al. showed K-NN claasifier performs better than K-means clustring in missing value [9]

Phimmarin keerin proposed a CKNN(cluster based K-NN) methods for missing value technique [10]

It is an extension of k nearet neighbour with local clustirng to improve efficiency and proved CKNN perform better than normal K-NN.

Nirmala devi et al. showed missing value technique by mean and median of clusters.

Trilok suthar proposed missing value technique for stream data. The proposed technique is based on the Skewness sensitive. The results proved the better results for stream data. [11]

For categorical data missing value is replaced by highest frequency count.

Clustering techniques are applied for outlier detection.

Data Transformation:

In this technique the data are converted into an appropriate form for mining.

The basic technique is normalization where data are scaled to fall within a small specified range. Generally negative numbers are converted into positive numbers. Such as -10 to 10. Normalization is done by three techniques such as Z-score, by decimal scaling and min max normalization. The normalization results require the range of a sameness or distance measure lies with fixed range.

Data Reduction:

Data Mining techniques may take very long time for complex data or on huge amount of data. This analysis may be impractical or infeasible. Data reduction techniques helps to reduce representation of data without affecting quality of data and produce promising results. methods for data reduction includes data cube aggregation, dimension reduction, data compression and discretization.

In discretization continuous attributes are converted into categorical attributes or into intervals.

Classification:

Classification is an important technique in data mining. The goal of classification is to build a concise model of the distribution of the class label in terms of predictor attributes. The resulting model is used to assign class labels to future records where the values of the predictor attributes are known but the value of the class label is unknown. The preprocessing techniques helps to improve the classification of data. And also improves the classification accuracy.

III. CONCLUSION

In real word data prone to incomplete, inconsistent noisy and missing. Data preprocssing is one of the important technique for getting quality data. Preprocessing includes data cleaning, data integration, data transformation and data reduction. The paper explains the study of various data preprocessing techniques for classification for static and stream data. The study shows various method and its expiation

The preprossed data helps to improve the analysis quality of classification and other data mining technique.

The study concludes that data preprossing methods have effective and important role in preparation, analysis of high dimension data.

IV. REFERENCES

[1] http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining- 3.html

[2] Cw.flek.cvut.cz/lib/exe/fetch.php/cources/ac4m33sad/2_tutorial.pdf.

[3] S. McClean, B. Scotney and M. Shapcott, "UsingBackground Knowledge with Attribute-OrientedData Mining", Knowledge Discovery and Datamining (Digest no, 1998/310), IEEcolloquiumon, pp. 1/1-1/4, 1998.

[4] J. Shena and M. Chen, "A Recycle Technique ofAssociation Rule for Missing Value Completion"in Proc. AINA'03, pp. 526-529, 2003.

[5] Thomas R. Gabriel and Michael R. Berthold,"Missing Values in Fuzzy Rule Induction",Systems, Man and Cybernetics, IEEEInternational Conference on (Volume: 2), 2005.

[6] M. Shyu, I. P. Appuhamilage, S. Chen and L.Chang, "Handling Missing Values viaDecomposition of the Conditioned Set", IEEESystems, Man, and cybernetics society, pp. 199-204, 2005.

[7] Olga Troyanskaya, Michael Cantor, GavinSherlock, Pat Brown, Trevor Hastie, RobertTibshirani, David Botstein and Russ B. Altman,"Missing value estimation methods for DNAmicroarrays", Bioinformatics 17 (6): 520-525,2001.

[8] Anjana Sharma, Naina Mehta, Iti Sharma, "Reasoning with Missing Values in MultiAttribute Datasets" ,International Journal ofAdvanced Research in Computer Science andSoftware Engineering, Volume 3, Issue 5, May2013 .

[9] R. Malarvizhi, A. Thanamani," K-NN

ClassifierPerforms Better Than K-Means Clustering inMissing Value Imputation", IOSR Journal ofComputer Engineering (IOSRJCE), vol. 6, pp.12-15, Nov. - Dec 2012.

[10] Phimmarin Keerin and Werasak Kurutach,Tossapon Boongoen, "Cluster-based KNNMissing Value Imputation for DNA Microarray Data",IEEE International Conference onSystems, Man, and Cybernetics COEX, Seoul,Korea, October 14-17, 2012.

[11] Trilok suthar "novel preprocessing techniques for NID3R" International Journal Of Engineering And Computer Science Volume 6 Issue 5 May 2017.