A Study on Data Preprocessing Techniques

T.Preethi[1] E.Manohar[2] K.Chitralakshmi[3]
[1]P.G Scholar [2,3]Assistant Professor
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Francis Xavier Engineering College, Tirunelveli, India

Abstract— Data preprocessing is a major technique used for refining the data. The preprocessed data can be used for different purposes. They are used in the fields of data mining, artificial intelligence, training a network in neural network. The processed data provides different patterns for the set of input data. There are different steps for processing the data. They are highly applicable in the machine learning.
Key words: Data preprocessing, cleaning, Integration, reduction. Transformation, discretion

I. INTRODUCTION

In the real world data is very noisy, incomplete, and inconsistent. The data might lack attributes may contain errors and the quality of the data may be low. The data is of different types and forms. The data can be numeric, state hierarchy, static, temporal. The different kinds can be text, web, metadata, images, audio/video and the data can also be distributed.

The raw data cannot be used for the different purposes such as data mining, machine learning. The data must be gathered and the combinations of the data must be found. The impossible combinations are also found in the data gathering methods. They must be neglected in order to avoid the wrong results. The quality of the data must be the first for the consideration. The quality of the data will influence the results of data mining. For improving the data mining results the raw data is preprocessed and the raw data is improved in efficiency.

Data preprocessing is one of the important steps in data mining but it is often neglected. If the irrelevant and redundant information are not removed .The knowledge discovery during the training is more complex. This process involves more processing time. The data preprocessing involves the following steps

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

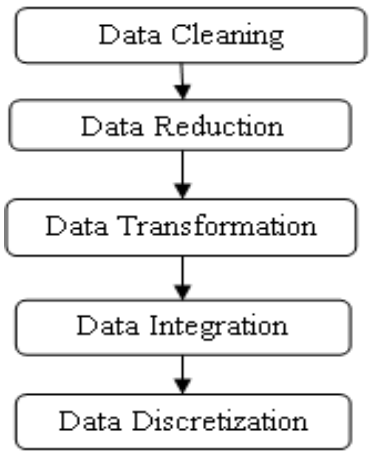

Fig. 1: Data preprocessing steps

II. DATA CLEANING

Data cleaning is the first process in the data preprocessing. Data that is not processed must be checked for the possible combinations and other possibilities for the proper consistency. Incomplete and inconsistent data cannot be used for the purposed of the data mining. The attributes relating to the data is not available. They may not be included to the data size they may be considered as very important during the time of consideration. The data must be processed through some data cleaning steps. The data that are redundant must be removed in the form of data cleaning. The data must be accurate so the data must be consolidated .the consolidation of different data representation and elimination of duplicate information become necessary. Data cleaning needs support from the data warehouses. The decision making is mainly done by data warehousing. The correctness of the data is more important to avoid the wrong conclusions. The data cleaning must satisfy several requirements

First remove the major errors and inconsistencies both in individual data sources and integrating multiple sources. The different papers focus on the problem of the duplicate identification and elimination. The approaches perform the following process First analyze the data for different errors and the inconsistent data. The workflow of the data must be analyzed. The verification for the correctness of the data must be identified. Analysis design and verification steps must be done in multiple iterations. After the errors are removed the cleaned data must replace the unclean raw data.

III. DATA INTEGRATION

Data integration is the process of combing multiple data to the coherent store. Schema integration is the process of intenerating metadata from different sources.

Detecting and resolving the conflicts in the data values for the different reasons, different scales .while data integration of multiple databases redundant data might occur often. The handling of redundant data is more important n the data integration.

IV. DATA TRANSFORMATION

The data transformation involves the process of removing noise from the data and then the summarization of the data and aggregation of data. The generalization concept is used for the hierarchy climbing. The data transformations include the instance related transformation. There takes the generation of transformation code that reduces the amount of self-programming that is necessary to provide the requires transformation in the appropriate language. Various ETL tools are used for supporting rule language. ETL means extraction, transformation, loading tools. They also support the SQL to perform transformation. User defined functions

are implemented in SQL or a general purpose language. They perform many data transformation and they also perform less complicated reuse for various transformation and other query processing tasks. The transformation forms a view and performs mapping operation. They also contain some cleaning logic for the efficient results of cleaning.

V. DATA REDUCTION

Data reduction is the process of transformation of numbers and digital information into an ordered and simplified form. It is the process of reduction of large amount of data into meaningful parts. Data reduction can be done by the clustering techniques. Clustering is the process of grouping elements of similar features. The elements from different clusters are dissimilar. There are different types of clustering. Hierarchical clustering performs the clustering by pairing up the data items and moves up or down by the hierarchy. They are classified as the agglomerative and divisive. This clustering is represented by the dendrogram. K-means clustering is more popular. Unsupervised clustering is faster than the hierarchical clustering. The comparison of the quantitative data is also done. They are designed to explore the different types of data.

More data will take more time to complete the process of data preprocessing. Hence the large amount of data must be reduced. Data reduction is the process of obtaining a reduced data set representation that is much smaller in size and produces the same analytical results. Data reduction can be done by the process of the histograms, clustering, sampling. The clustering is the process of grouping the elements in a group by their similar features.

VI. DATA DISCRETIZATION

Discrete values plays major role in data mining and knowledge discovery. They are the intervals of numbers that are very easy to make use of them, as they are closer to a knowledge based representation than the continuous set of values induction algorithms are used for discrete features. The data discretization has three types of attributes nominal, ordinal, continuous. Discretization is the process where the range of continuous attributes is divided into intervals. The data size must be reduced by discretion. It performs the reduction of the number of values for a given attribute by dividing the range of attributes into the certain intervals. The reduced data sets can be used for further data mining process.

VII. CONCLUSION

This paper describes the different techniques of the data preprocessing. The data preprocessing is the major step that is to be done before the mining process. The different techniques and the algorithms used for the steps of the data preprocessing will be discussed in the future work.

REFERENCES

[1] Rajashree Y.Patil,, Dr. R.V.Kulkarni "A Review of Data Cleaning Algorithms for DataWarehouse Systems"(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012,5212 - 5214.

[2] V.Chitraa, Dr. Antony Selvdoss Davamani ,"A Survey on Pre-processing Methods for Web Usage Data"(IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[3] Navin Kumar Tyagi, A.K. Solanki& Sanjay Tyagi "An algorithmic Approach To Data Preprocessing In Web Usage Mining" International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 279-283.

[4] Ankit R Kharwar1, Chandni A Naik2, Niyanta K Desai," A Complete Pre Processing Method for Web Usage Mining" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013.

[5] Marathe Dagadu Mitharam " Preprocessing in Web Usage mining"International Journal of Scientific & Engineering Research, Volume 3, Issue 2, February -2012 1 ISSN 2229-5518 IJSER © 2012 http://www.ijser.org.

[6] Vahid Nouri, Mohammad-R. Akbarzadeh-T. Alireza Rowhanimanesh "A Hybrid Type-2 Fuzzy Clustering Technique for Input Data Preprocessing of Classification algorithm"2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) July 6-11, 2014, Beijing, China.

[7] IndreZliobaite and Bogdan Gabrys "Adaptive Preprocessing for Streaming Data" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, February 2014

[8] Mark E. Snyder, Ravi Sundaram, Ravi Sundaram," Preprocessing DNS Log Data for Effective Data Mining"978-1-4244-3435-0/09/$25.00 ©2009 IEEE

[9] Salvador Garcı´a, Julia´n Luengo, Jose´ Antonio Sa´ ez, Victoria Lo´ pez, and Francisco Herrera " A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning"Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013.

[10] Parag C. Pendharkar "A Data Envelopment Analysis-Based Approach for Data Preprocessing"Ieee Transactions On Knowledge And Data Engineering, Vol. 17, No. 10, October 2005.

