INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

AN EFFICIENT PREPROCESSING AND POSTPROCESSING TECHNIQUES IN DATA MINING

R.Tamilselvi[1], B.Sivasakthi[2], R.Kavitha[3]

[1] M.Phil. Research Scholar,tamilselvirps@gmail.com
[2] M.Phil. Research Scholar,sivasakthibaskar91@gmail.com
[3]Assistant Professor,rkavithamscmphil@gmail.com

Department of computer science, Vivekanandha College for women, Unjanai, Tiruchengode, TamilNadu, India

Abstract: - Organizations are maintaining history of data for future analysis. These huge volume of database is analysed to Predict and improve the benefits and profits of the organization and also for the development. By analysing the history of data, strategic decisions can be made to improve the performance of the organizations by the top level peoples. So organizations are interested in analysing the data which will result in valuable insight. The data subjected to mining consists of inconsistent, blank or null and noisy values which have to be cleaned before mining. Usually the Techniques of Mean, Mode, and Median will be used to clean the data which are inefficient methods. Here I am representing the efficient data pre-processing and post-processing which is to be carried out before actual mining process can be performed. The data from different databases, different locations and different formats are considered for pre-processing and post processing. The data pre-processing includes four stages. They are cleaning the Data, integrating the data, Data Reduction and Data Transformation. Cleaning the data: is to fill the empty value of the data and to ignore the noisy data and to correct the inconsistencies data. Integrating the data: merging the data from multiple data source. Data Reduction: a reduced representation of the data set that is much smaller in volume yet maintains the integrity of the original. Data Transformation: the data set are transformed or consolidated into forms appropriate for mining. Post processing includes two stages. They are visualization and summarization. Data visualization: aims to communicate data clearly and effectively through the graphical representation. Data summarization: A collection of patterns can be regarded as a summary of the data.

Keywords: Data Pre-processing, Post Processing, Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Summarization, Data Visualization

1. INTRODUCTION

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprises decision making process. Conventional database systems are often designed for day-to-day running of an organization and are called online Transactions processing (OLTP) systems. Data mining is often a complex process and may require a verity of steps before some useful results are obtained. Data mining comes in two flavours—directed and undirected. Directed data mining attempts to explain or categorize some

particular target field such as income or response. Undirected data mining attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. Data pre-processing is challenging as it involves extensive manual effort and time in developing the data operation scripts. There are a number of different tools and methods used for pre-processing, including: sampling, which selects a representative subset from a large population of data; transformation, which manipulates raw data to produce a single input; denoising, which removes noise from data; normalization, which organizes data for more efficient access; and feature extraction, which pulls out specified data that is significant in some particular context. Pre-processing technique is also useful for association rules algorithms Like-Apriori, Partitioned, Princer-search algorithms and many more algorithms. Data post processing is used to Visualization technologies have recently been used in steering computation, in aiding directed analysis, in query interfaces to complex multimedia databases, and in information presentation and navigation. Summarisation is closely related to compression, machine learning, and data mining.

2. PREPROCESSING TECHNIQUES

Data pre-processing is an often neglected but import step in the data mining process. The phrase "Garbage IN, Garbage Out" is particularly applicable to data mining and machine learning. Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to help improve the quality of data and, consequently of the mining results raw data is pre-processed so as to improve the efficiency and case of the mining process. Data pre-processing is one of the most critical steps in data mining process which deals with the preparation and transformation of the initial dataset. Data pre-processing methods are divided in to four categories:

- **Data Cleaning**
- **Data Integration**
- **Data Transformation**
- **Data Reduction**

2.1 Data Cleaning

Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Fig: Data cleaning

2.1.1 Missing Values

Filling in the missing values for the particular attribute.it consists of following methods:

Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.
Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like *"Unknown"* or □¥. If missing values are replaced by, say, *"Unknown,"* then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of *"Unknown."* Hence, although this method is simple, it is not fool proof.
Use the attribute mean to fill in the missing value: For example, suppose that the average income of *All Electronics* customers is $56,000. Use this value to replace the missing value for *income*.
Use the attribute mean for all samples belonging to the same class as the given tuple: For example, if classifying customers according to *credit risk*, replace the missing value with the average *income* value for customers in the same credit risk category as that of the given tuple.

Use the most probable value to fill in the missing value this may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

2.1.2 Noisy Data

Noise is a random error or variance in a measured variable. The data to remove the Noise use the following data smoothing techniques:

Binning: Binning methods smooth a sorted data value by consulting its "neighbourhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or *bins*. Because binning methods consult the neighbourhood of values, they perform *local* smoothing. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the *bin boundaries*. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be *equal-width*, where the interval range of values in each bin is constant. Binning is also used as a discretization technique.

Regression: Data can be smoothed by fitting the data to a function, such as with regression. *Linear regression* involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other. *Multiple linear regression* is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Clustering: Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outlier analysis.

2.2 DATA INTEGRATION

Data integration, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. There are a number of issues to consider during data integration. *Schema integration* and *object matching. Redundancy* is another important issue. An attribute (such as *annual revenue*, for instance) may be redundant if it can be "derived" from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set. Some redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data. For numerical attributes, we can evaluate the correlation between two attributes, *A* and *B*, by computing the correlation coefficient.

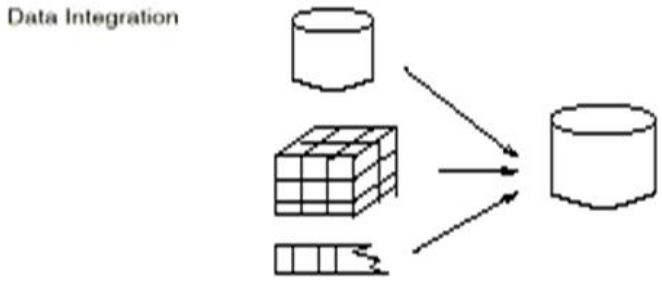

Fig: Data Integration

2.3 DATA TRANSFORMATION

In *data transformation*, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

Smoothing: This works to remove noise from the data. Such techniques include binning, regression, and clustering.

Aggregation: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.

Generalization: The data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like *street*, can be generalized to higher-level concepts, like *city* or *country*. Similarly, values for numerical attributes, like *age*, may be mapped to higher-level concepts, like *youth*, *middle-aged*, and *senior*.
Normalization: where the attribute data are scaled so as to fall within a small specified range, such as 1:0 to 1:0, or 0:0 to 1:0.Attribute construction (or *feature construction*), where new attributes are constructed and added from the given set of attributes to help the mining process. An attribute is normalized by scaling its values so that they fall within a small specified range, such as 0.0 to 1.0.Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest-neighbour classification and clustering. If using the neural network back propagation algorithm for classification mining, normalizing the input values for each attribute measured in the training tuples will help speed up the learning phase.

Fig: Data Transformation

2.4 DATA REDUCTION

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the analytical results.

Strategies for data reduction include the following:

Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.

Attribute subset selection: where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

Dimensionality reduction: where encoding mechanisms are used to reduce the data set size.

Numerosity reduction: where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.

Discretization and concept hierarchy generation: where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction.

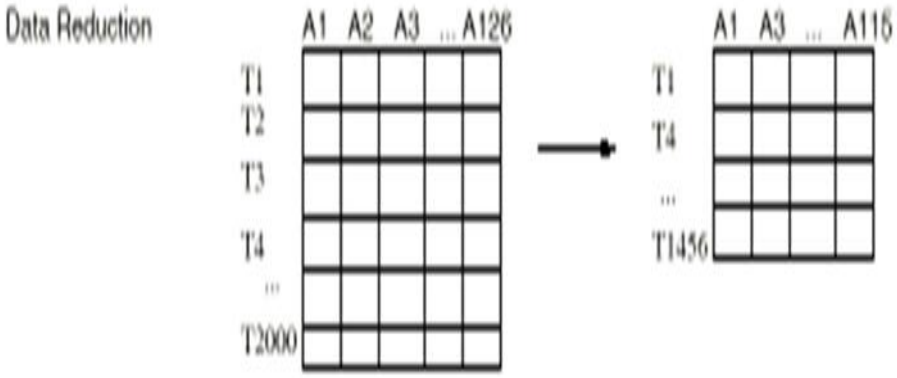

Fig: Data Reduction

3. POST PROCESSING TECHNIQUES

It is essential to visualize the extracted knowledge in such a way that user can interpret the knowledge easily Data post-processing methods are divided in to two categories:

- **Data Visualization**
- **Data summarization**

3.1 DATA VISUALIZATION

Data Mining is useful for extracting the knowledge from large databases. After extraction, it is important to visualize the extracted knowledge in such form so that user can gain insight into data for better decision making. Various techniques are available for information visualization. Such as,

Scatter Plot Matrix Technique: Scatter Plots are organized in matrix form and use Cartesian co-ordinates to plot data points [53-55]. The relationship between two variables, also known as correlation, is represented by scatter plots. The correlation between two variables may be positive or negative. If the data points are distributed uniformly in the scatter plot, then the correlation between two variables is low or zero. High correlation may be positive or negative depending on the relationship between variables. If the value of one variable increases with the increment of the value of another variable and if data points are represented by a straight line in scatter plot, then the correlation is said to be high positive correlation. If the value of one variable decreases with the increment of the value of another variable and data points make a straight line then correlation is called high negative correlation.

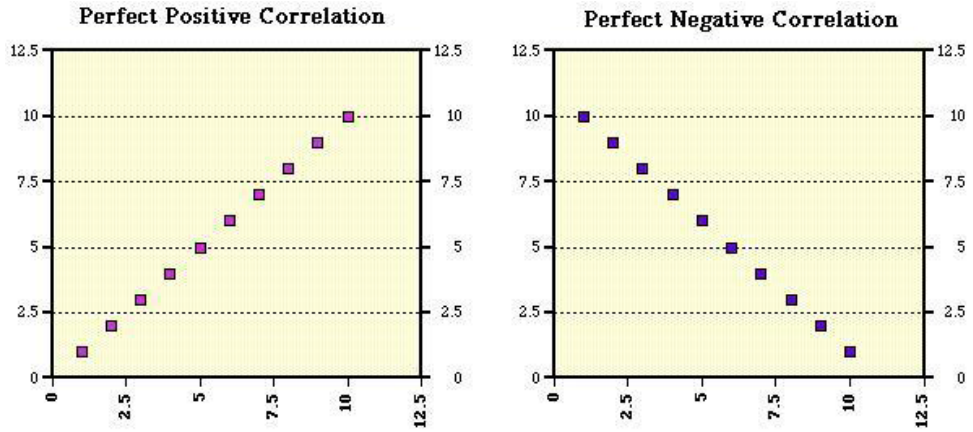

Figure: Perfect Positive and Negative Correlation between Two Variables

Parallel Co-ordinates: This technique maps a multi-dimensional point onto a number of parallel axes. Initially in this method coordinates start mapping with one axis and then gradually more axes may be lined up as per requirement. A line is used to connect the individual coordinate mappings. This method may be extended to n-dimensions but there is a practical limit which depends on the screen display area [53, 58].

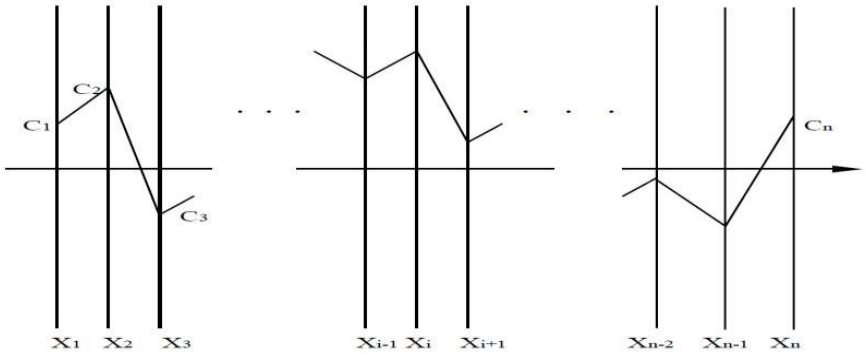

Figure: Parallel Axes

3.2 DATA SUMMARIZATION

Knowledge discovery in both structured and unstructured datasets stored in large repository database systems has always motivated methods for data summarisation. Summarisation is closely related to compression, machine learning, and data mining. The closest connection is to data mining. Data summarisation methods for the unstructured domain usually involve text .Categorisation which groups together documents that share similar characteristics. With the ever growing number of text documents in large database systems, algorithms for text summarisation in the unstructured domain, such as document clustering, are often limited by the dimensionality of the data features.

4. CONCLUSION

This paper shows a brief explanation on pre-processing and post-processing techniques. It starts with pre-processing techniques which includes detailed description of various data cleaning approaches, and ignores the noisy data, imbalanced data handing and dimensionality reduction. And also transformed the data set into appropriate forms for mining. Effective post-processing visualization is the important for a verity of government, corporate as well as industrial applications.

REFERENCES

[1] Data Pre-processing & Mining Algorithm, Knowledge & Data Mining & Preprocessing, 3rdedition, Han & Kamber.
[2] DATA MINING TECHNOLOGY by Jiawei Han Department of Computer Science University of Illinois at Urbana-Champaign
[3] Soukup, T., & Davidson, I. (2002). Visual data mining: Techniques and tools for data visualization and mining, Wiley
[4] J. Wang and G. Karypis. On anciently summarizing categorical databases. Knowledge and Information Systems, 9(1):19{37, January 2006.