An Overview on Data Preprocessing Methods in Data Mining
R. Dharmarajan¹ R.Vijayasanthi²
¹Assitant Professor ²M.Phil Research Scholar
¹,²Department of Computer Science
¹,²Thanthai Hans Roever College, Perambalur

Abstract— Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data pre-processing is an often ignored but important step in the data mining process. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data pre-processing includes cleaning, Integration, Transformation, reduction, etc. This paper analyzes existing work done in the preprocessing steps. A brief overview of various data preprocessing techniques for Data cleaning, Data Integration, Data Transformation, Data reduction, Data Discretization is discussed.

Keywords: Data Preprocessing, Data cleaning, Data Integration, Data Transformation, Data reduction.

I. INTRODUCTION

Data mining refers to extracting on interested information or knowledge from large amounts of data sources. Alternative names are data mining is Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis etc. The real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources. Low-quality data will lead to low-quality mining results (1).

The quality of data affects the data mining results. In order to help improve the quality of the data and, consequently, of the mining results raw data is pre-processed so as to improve the efficiency and ease of the mining process. Data preprocessing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset. Data preprocessing methods are divided into following categories:

- Data Cleaning
- Data Integration
- Data Transformation
- Data reduction

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. 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 (2).
A. Data Cleaning

Data cleaning is a first step of data preprocessing method. Cleaning on the data is one biggest problem in constructing of data warehousing and mining. It routines with respect to soil data work is to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Dirty data can cause confusion for the mining procedure, resulting in unreliable and poor output. There is necessity for useful pre processing step to be used some data-cleaning routines (4).

If mining technique is apply to this type of data then our result is unreliable and poor output. Then be needed to clean on the data before data mining. The missing values in the tables are to be corrected by following measures
- Ignore the tuple
- Fill in the missing value manually
- Use a global constant to fill in the missing value
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class.
- Use the most probable value to fill in the missing value.

Noisy data means that data in the tables containing errors, or outlier values that deviate from the expected. This problem is corrected by following procedures or techniques
- Binning
- Regression
- Clustering

B. Data Integration

Integration of different type of data, attributes and schema are biggest problem in constructing of data warehousing and data mining because real world data are available in a different location. If mining technique or analysis technique is apply to this type of data warehouses then it taken to more time and decision is unreliable and quality less. Then be needed to combines data from multiple sources into a one place or one database form of data integration, attributes integration and schema integration. The sources may be multiple databases, flat files and data cubes (3). Some attributes representing a given concept may have different names in different databases, causing inconsistencies and redundancies. For example, the attribute for soil sample number for identification may be referred to as soil id in one database and soil_sample_id in another. Naming inconsistencies in soil datasets may also occur for attribute values. Having a large amount of redundant data may slow down or confuse the mining process (4).

Additional data cleaning can be performed to detect and remove redundancies that may have resulted from data integration. The following points listed different type of data integration technique and their example
1) Data integration
in this technique combines data from multiple sources as in data warehousing, for example cust_id for one database and cust_number is another database attributed then time data analyst check to value in both database if match then be integrated do not integrated.
2) Schema integration
it technique identify and remove entity identification and Redundancy problem. In this technique combine similar type of attributed as a one attributed. Metadata can be used remove entity identification and correlation analysis used to remove Redundancy. For example cust_id for one database and cust_number is another database attributed then time data analyst check to metadata both attributed if match then be integrated otherwise do not integrated.
3) Detecting and resolving data value conflicts
when attribute values are differing but concept and structure are same then time we use this technique. Functional dependencies, referential constraints and scaling, or encoding are used to remove this type of problem and check lower level of abstraction of database. For example cust_id for one database and cust_number is another database attributed then time data analyst check to metadata both attributed if match then be integrated otherwise do not integrated.

C. Data Transformation

Data transformation operations, such as normalization and aggregation, are additional data pre-processing procedures that would contribute toward the success of the mining process. Normalization: Normalization is scaling the data to be analyzed to a specific range such as \([0.0, 1.0]\) for providing better results in data mining process in ANN classification techniques (2). Aggregation of the data is one of the data transformation task and it would be useful for data analysis to obtain aggregate information such as the Nitrogen content of particular field or location. Transformation of different type of data, schema from one format to another format is largest problem of constructing in a data mining and data warehousing, WWW etc. because.
real world data are available in a different format and language.

If mining technique or analysis technique is apply to this data in data warehouses or WWW then it taken to more time, holding on more space, decision is suspended and quality less. Then be needed to transform the data one format to another format use of Smoothing, aggregation, generalization, normalization technique (6).

1) Aggregation (use for numerical data)
In this technique calculated to summary on data in a specific attributed. This technique is useful in constructing a data cube for analysis of multiple granularities. For example Daily_sales attributed value may be aggregated so as to compute monthly_sales and annual_sales amounts.

2) Generalization
In this technique low-level data are replaced by higher-level data through concept hierarchies’ technique. It technique useful of categorical attributes numerical attributes. For example Address attributed like street data, can be generalized to higher-level concepts, like city data, after that it generalized to state data, after that it generalized to country data.

3) Normalization
Normalization is scaling on data to be analyzed to a specific range such as (0.0, 1.0). It technique useful of ANN classification algorithm. min-max normalization, z-score normalization and decimal scaling are use in normalization technique for example Score_pretences attributed value are normalized as a graded value.

4) Attribute construction
In this technique new attributes are constructed and added from the given set of attributes. It technique useful of feature construction. For example add the attribute area on database then it based (calculated) on the attributes height and width.

D. Data reduction
Imagine that you have selected data from the All Electronics data warehouse for analysis. The data set will likely be huge! Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible. 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 same (or almost the same) 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 (5).

III. CONCLUSION
Data preprocessing is an important issue for both data warehousing and data mining, as real-world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation, and data reduction. This paper gives overview on data preprocessing method and their example. Aim of data preprocessing is given to qualities data for any type of mining like that data mining, text mining and web mining. Data cleaning method are clean the noisy data, completed on uncompleted data and remove unnecessary data. Data integration method is integrated to different source of data in one place. Data transformation method change form of data and data reduction reduces the volume of database by schema integration.

REFERENCE
[5] Data Pre-processing & Mining Algorithm, Knowledge & Data Mining & Pre-processing, 3rd edition, Han & Kamber.