A Review on Data Mining and Data Preprocessing Techniques in Data Mining

[1] Miss Vaidya Vijayshri Dattatray, [2] Mr Nibe Abhishek Annasaheb

[1] *Lecturer, Department of Computer Technology, P. Dr. V. V. P. Polytechnic Loni Maharashtra India*

[2] *Lecturer, Department of Computer Technology, P. Dr. V. V. P. Polytechnic Loni Maharashtra India*

Abstract

Data mining is the process of searching the knowledge i.e. interesting or useful data from the large amount of data stored in data warehouse. KDD [6] process can be used by experts to find out the useful data for business decision making. Data mining is fully dependent on the quality of data. Data used for data mining is collected from different users and sources, which leads to incomplete data, noisy data or inconsistent data. In this paper we are going to study preprocessing techniques [2] which can be used to preprocess the raw data and convert into quality data for data mining. Data preprocessing is the major step in data mining, which cleans the data that can be used for business decision purpose.

Keywords: *Data Mining, KDD (Knowledge Discovery in Database), Data Warehouse, Data Preprocessing*

1. INTRODUCTION

Data mining [6] refers to extraction of small information from large amount of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data mining is the process of discovering interesting patterns and knowledge from *large* amounts of data. Data mining is used by companies in order to get customer preferences, determine price of their product and services and to analyze the market.

Data mining is also known as knowledge discovery in Database [6] (KDD).

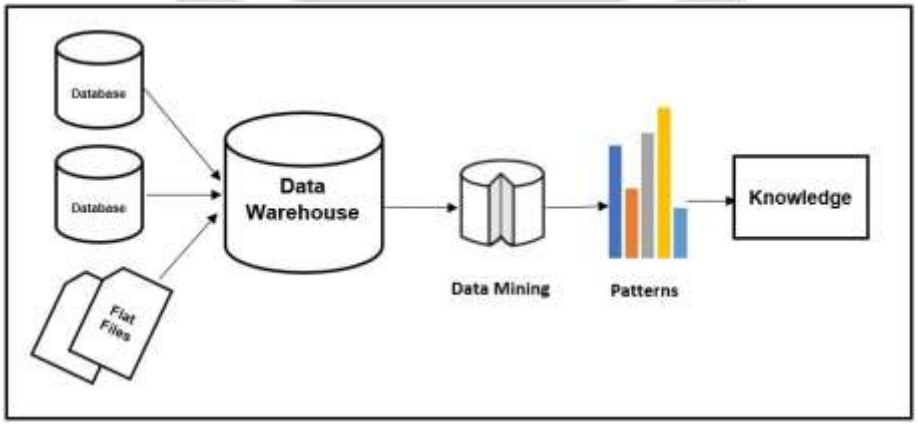

Fig 1: Steps in KDD (Knowledge Discovery in Database)

i. Data cleaning:
In data cleaning it removes the noise or errors and inconsistent data.

ii. Data integration:
In data integration, the data can be collected from varied sources and integrated at single location.

iii. Data selection:
In data selection process, only the data relevant to the analysis task can be retrieved from the database.

iv. Data transformation:
The data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations. i.e. the data from different data sources which is of varied types can be converted into a single standard format.

v. Data mining:
Data mining is the process in which intelligent methods or algorithms are applied on data to extract useful data patterns.

vi. Pattern evaluation:
This process identifies the truly interesting patterns representing actual knowledge based on user requirements for analysis.

vii. Knowledge presentation:
In this process, visualization and knowledge representation techniques are used to present mined knowledge to users for analysis and decision making.

2. DATA PREPROCESSING

Data preprocessing [2] [4] is a data mining technique that involves transforming raw data into an understandable format.
Real-world data is incomplete, inconsistent or contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Data Preprocessing is required because real world data are generally:
1. Incomplete:
When a dataset contains missing values, it is referred to as an incomplete dataset.
i.e. Missing attribute values, missing some important attributes, or having only aggregate data.

2. Noisy:
Noisy data means it contain errors or outliers. Noisy data is meaningless data.
Include any data that cannot be understood and interpreted correctly by machines, such as unstructured data.

3. Inconsistent:
Inconsistent data means data containing discrepancies in codes or names.
Data inconsistency is a situation where there are multiple tables within a database that deal with the same data but may receive it from different inputs.
Data preprocessing includes the tasks such as data cleaning, data integration, data transformation, data reduction, data discretization. [3] [4] [5] [7]

2.1 DATA CLEANING

Quality of the data is important for final analysis. Any data which is incomplete, noisy and inconsistent can affect the results. Data cleaning [1] [5] in data mining is the process of detecting and removing corrupt or inaccurate records from a record set, table or database.

2.1.1. Handle Missing Values:
a. Ignore the tuple:
This is done when class label is missing. Ignore the tuple only if maximum attributes have the missing values.

b. Fill in the missing value manually:
This approach is effective on small data set with some missing values.

c. Replace all missing attribute values with global constant:
We can use any global constants to replace the missing value like "Unknown".

d. Use the attribute mean to fill in the missing value:
Missing value is replaced by the average value of that column or attribute. For example, customer average income is 25000 then you can use this value to replace missing value for income.

e. Use the most probable value to fill in the missing value.
We can replace the missing value by most probable value which is consistent for that attribute.

2.1.2. Regression:
Data can be smoothed by fitting the data into a regression function. Example: If we measured the height of child per year, if child grows 3 inches approximately, then the regression function may be: **child growing 3 inches per year**

2.1.3. Clustering:
Outliers may be detected by clustering, where similar values are organized into groups, or "clusters. Values that fall outside of the set of clusters may be considered outliers. The outliers may be ignored while analysis of data.

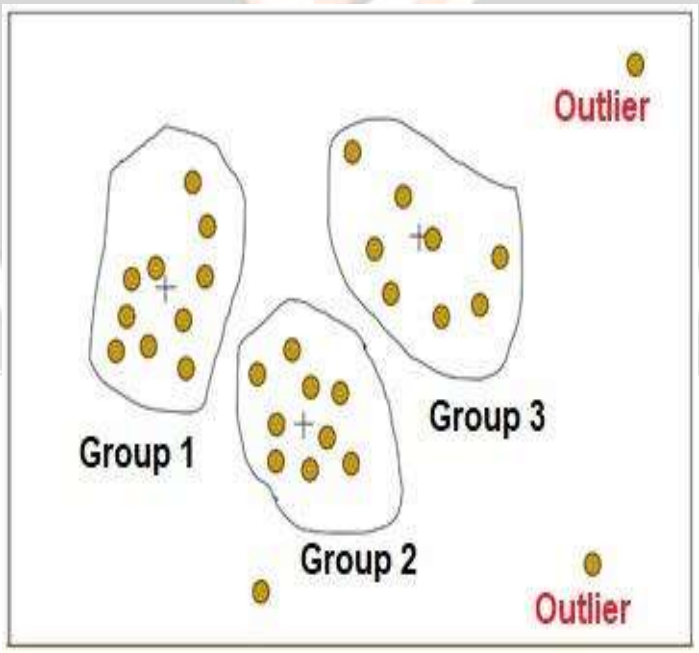

Fig 2: Cluster analysis

2.2 DATA INTEGRATION:

Data Integration [5] [8] is a data preprocessing technique that combines data from multiple data sources and provides a unified view of these data to users.

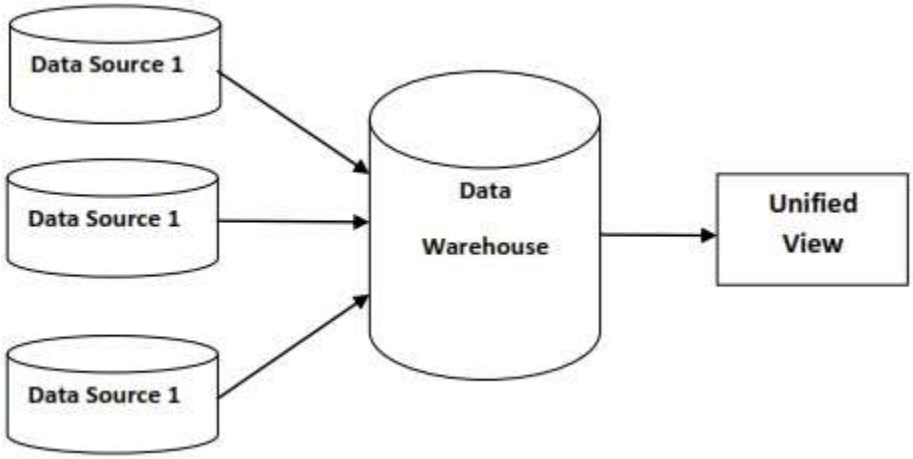

Fig 3: Data integration

These sources may include multiple databases, data cubes, or flat files. One of the most well-known implementation of data integration is building an enterprise's data warehouse. The benefit of a data warehouse enables a business to perform analysis based on the data in the data warehouse. There are mainly 2 major approaches for data integration:

2.2.1. Tight Coupling
In tight coupling data is combined from different sources into a single physical location through the process of ETL - Extraction, Transformation and Loading.

2.2.2. Loose Coupling
In loose coupling data only remains in the actual source databases. In this approach, an interface is provided that takes query from user and then sends the query directly to the source databases to obtain the result.

2.3 DATA TRANSFORMATION

In data transformation [5] [8] process, data are transformed from one format to another format that is more appropriate for data mining.
Ex: Original data: 1.2, 3.2, 4.6, 123
Transformed data: 120, 320, 460, 123

2.3.1. Smoothing:
Smoothing is a process of removing noise from the data. For removing the noise from data, binning method is used which can smoothen the data by removing noise using bin means or bin boundaries.

2.3.2. Aggregation:
Aggregation in data mining is the process of finding, collecting, and presenting the data in a summarized format to perform statistical analysis of business decisions. Aggregated data help in finding useful information about a group after they are written as reports. To represent the aggregate data, multidimensional data cube or OLAP cube is used.

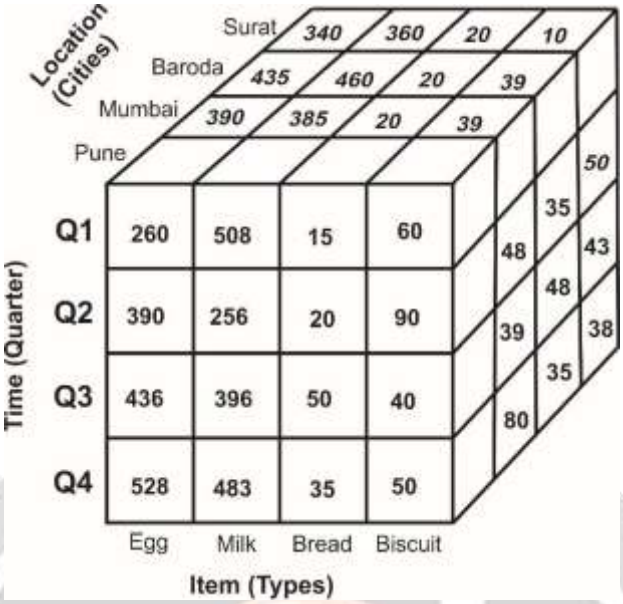

Fig 4: Data cube aggregation

2.4 DATA REDUCTION

A database or date warehouse may store large amount of data. So, it may take very long to perform data analysis and mining on such huge amounts of data. Data reduction [5] [8] techniques can be applied to obtain a reduced representation (without loss of any data) of the dataset that is much smaller in volume but still contain critical information.

2.4.1. Data Cube Aggregation:

Aggregation operations are applied to the data in the construction of a data cube. Cube can represent the aggregate data only, required for analysis purpose. Refer the figure 4.

2.4.2. Dimensionality Reduction:

In dimensionality reduction, redundant attributes are detected and removed, which reduces the dataset size.
Example:

Table 1: Dimensionality reduction

Before Reduction					After Reduction		
A1	A2	**A1**	A3		**A1**	A2	A3
10	11	**11**	21		**21**	11	21

2.4.3. Discretization:

Refer the below section.

2.5 DATA DISCRETIZATION:

Data Discretization [5] [8] techniques can be used to divide the range of continuous attribute into intervals. (Continuous values can be divided into discrete (finite) values) i.e. it divides the large dataset into smaller parts. Numerous continuous attribute values are replaced by small interval labels. This leads to a brief, easy-to-use, knowledge-level representation of mining results. Data mining on a reduced data set means fewer input/output operations and is more efficient than mining on a larger data set. Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining, rather than during mining.

2.5.1. Binning method:
Binning method is used which can divide the continuous values into equal bins and smoothen the data by removing noise using bin means or bin boundaries.

2.5.2. Cluster Analysis:
Cluster analysis is a popular data discretization method. In clustering, a group of different data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data. A clustering algorithm can be applied to discrete a numerical attribute of A by partitioning the values of A into clusters or groups. Each initial cluster or partition may be further decomposed into several subcultures, forming a lower level of the hierarchy. Refer fig 2.

CONCLUSION:

In this paper we studied the steps in KDD (knowledge Discovery in Database). Data preprocessing helps the analyst for better analysis and making good business decisions as it provides the quality data for data mining by using data cleaning, data transformation, data integration, data reduction and data discretization methods. Data preprocessing is needed to handle the raw data that may be incomplete, noisy or inconsistent.

REFERENCES:

[1] R. Gonzalez and A. Kamrani, "A Survey of Methodologies and Techniques for Data Mining and Intelligent Data Discovery," in D.Braha (ed.), Data Mining for Design and Manufacturing, Kluwer Academic Publishers, 2001, pp 41-59.

[2] K. Matousek, Z. Kouba, and P. Miksovsky, "Data Pre-Processing Support for Data Mining," In IEEE International Conference on Systems, Man and Cybernetics, pages 208-212. IEEE, October 2002.

[3] Dorian Pyle, Data Preparation for Data Mining, Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA, 1999.

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

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

[6] Ilan J. and M Karnber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Koufmann Publisher, San Fransisco, USA., ISBN-13: 978-1558609013, Pages:800.

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

[8] Suad A. Alasadi and Wesam S. Bhaya, Review of Data Preprocessing Techniques in Data Mining, J. Eng.Applied Sci. 12(16): 4102-4107, 2017.