Impact of Using Preprocessing in Data Mining and Knowledge Discovery Process

[1] Abdelrahman Elsharif Karrar, [2] Nafeesa Hassan Mohammed, [3] Moez Mutasim Ali

[1] College of Computer Science and Engineering, Taibah University
Al Madina, Saudi Arabia

[2] College of Computer Science and Information Technology, Al-Neelain University
Khartoum, Sudan

[3] College of Computer Science and Information Technology, University of Science and Technology
Omdurman, Sudan

Abstract - Data mining works to extract information known in advance from the enormous quantities of data which can lead to knowledge. It provides information that helps to make good decisions. The effectiveness of data mining in access to knowledge to achieve the goal of which is the discovery of the hidden facts contained in databases and through the use of multiple technologies. Unfortunately, real-world databases are highly influenced by negative factors such the presence of noise, inconsistent and superfluous data and huge sizes in dimensions, examples and features. Thus, low-quality data will lead to low-quality Data Mining performance. Data pre-processing is a first step of Data Mining in Knowledge discovery process (KDD) that reduces the complexity of the data and offers better analysis and ANN training. Based on the collected data from the field as well soil testing laboratory, data analysis is performed more accurately and efficiently. This paper study the huge impact of preprocessing in data mining by prepare the data (clean it, transform it, integrate it) to produce a good data that leads to high quality data mining performance.

***Keywords* - Preprocessing, Data Mining, Knowledge Discovery, Data Preparation**.

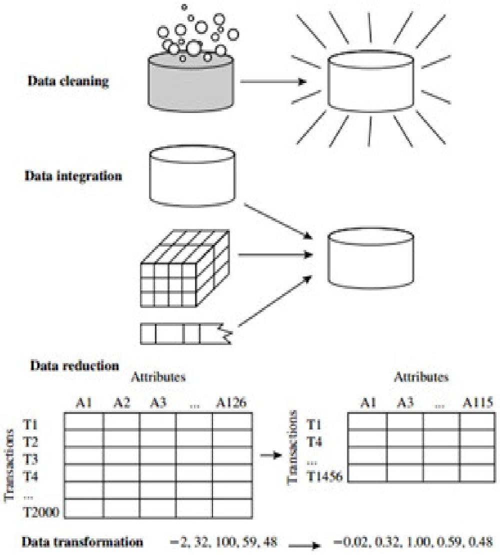
1. Introduction

We live in information age and computerization world where lots amount of data are generated daily and growth explosively, all these data must be collected to be analyzed which is an important need. To analyze this huge data and get knowledge from it the new technology moved toward the development for advanced data mining. Scientists have known the term data mining as "part of the process of knowledge discovery in databases, which are made using multiple methods its goal configure models of data". [1] To get knowledge from data we must go throw preprocessing which is an important step in data mining process. Preprocessing help us to solve the major problems (Noisy, missing, redundant and inconsistent data) that came from many possible reasons like data transmission errors, human or computer errors in data entry, etc.

2. Data Pre-processing

Our real world today generates huge amount of data especially in the age of “Internet of things”. All these data must be preprocessed to get knowledge from it, there are several data preprocessing techniques.

- Data cleaning can be applied to remove noise and correct inconsistencies in data.
- Data integration merges data from multiple sources into a coherent data store such as a data warehouse.
- Data reduction can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.
- Data transformations (e.g., normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements. [2]



Figure(1): Major steps in preprocessing

These techniques are not mutually exclusive they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

2.1 Data Cleaning

Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) can "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

• Missing values: The missing values in the tubles are to be corrected by following measures.

• Ignore the tuple: This is usually done when the class label is missing. 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: 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". If this happened the mining program may mistakenly think that they form an interesting concept, since they all have a value in common. Although this method is simple, it is not recommended.

• 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: This may be determined with inference-based tools using a Bayesian formalism or decision tree induction. [3]

• Noisy data: means that data in the tubles containing errors, or outlier values that deviate from the expected. This problem is corrected by following procedures or techniques.

• Binning: Binning methods smooth a sorted data value by consulting the "neighborhood", or values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighborhood of values, they perform local smoothing values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighborhood of values, they perform local smoothing.

• 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 variables, so that one variable can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two variables are involved and the data are fit to a multidimensional surface. Using regression to find a mathematical equation to fit the data helps smooth out the noise.

• Clustering: Outliers may be detected by clustering, where similar values are organized into groups or "clusters". [2]

2.2 Data Integration

Data integration involves integrating data from multiple databases, data cubes, or files. Some attributes representing a given concept may have different names in different databases, causing inconsistencies and redundancies. Additional data cleaning can be performed to detect and remove redundancies that may have resulted from data integration. [2]

2.3 Data Reduction

Data reduction obtains a reduced representation of the data set that is much smaller in volume. The reduced datasets produces the more or less same analytical results as that of original volume. There are a number of strategies for data reduction:

2.3.1 Data Cube Aggregation

Where aggregation operations are applied to the data in the construction of a data cube.

2.3.2 Dimensionality Reduction

Where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.

2.3.3 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 e.g. regression and log-linear models), or nonparametric methods such as clustering, sampling, and the use of histograms.

2.3.4 Data Discretization and Data Summarization

where raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are a powerful tool for data mining. [4]

2.4 Data Transformation

Data transformation has two main operations which are additional data preprocessing procedures that would contribute toward the success of the mining process.

These operations are:

2.4.1 Normalization

It 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.

2.4.2 Aggregation

It 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. [5]3. Tables, Figures and Equations

3. Impact of Preprocessing in Data Mining

Many factors affect the success of Machine Learning (ML) on a given task. The representation and quality of the instance data is first and foremost. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult.

When collecting the data from the field, there is the possibility of missing data or Noisy data that in the tubles containing errors or incorrect data which can be found in the data source that containing discrepancies between different data items.

When the users know that the data is “dirty” and have all these errors, they will not trust the results of any data mining process that has been applied. All these errors in the data can cause confusion for the mining procedure, resulting in unreliable output.

These days on our developing world towered Machine learning in the general different fields of our life, the entire Machine learning and data mining algorithms automatically extract knowledge from machine-readable information. Analyzing the data with one of data mining techniques without applying the preprocessing can result in the specific data field and guide it to unreliable output, for examples:

- In the healthcare field, it leads to wrong diagnosis.
- In the marketing field, it leads the retailers to misunderstand their buyer’s needs.
- In the education field, it leads to wrong decisions and wrong predict of the students result.
- In the manufacturing field, it leads to wrong patterns discovery in complex manufacturing process.
- In the CRM field, it leads to weak relationship with customers.
- In the financial banking field, it leads to huge business and finance problems.
- In the criminal investigate field, it leads to wrong identification of crime characteristics and wrong criminals detecting.

All the previous problems happened because the collected data in any fields have corrupted items, and all these items didn’t go throw the preprocessing process before analyze it with one of data mining techniques.

Data preprocessing occupies about 70% of time spent on knowledge discovery project. If data have impurities, for example missing or duplicate data, data mining tools may be misled and even give wrong results.

Based on wrong results, companies may make fatal decisions. Besides, preparing data is an integral part of building a data warehouse so that it integrates data of uniform quality.

There are a wide range of research issues to be addressed in this project. These can be summarized at a high level as the following set of over-arching questions:

- How can data pre-processing be improved for data mining?
- What forms of techniques are useful for determining data cleansing, feature reduction and classification?

And we set our future specific aims which can be:

- Investigation of systematic data preparation techniques for data cleansing and feature reduction.
- Investigation and Development of metrics for underpinning missing value.
- Investigation and Development of transforming data and integrate it between different environments.
- Compare the performance of different preprocessing algorithms which may be applied on medical data.

So the main objective here is to help improve the quality of the data and, consequently, of the mining results, and see the impact of preprocessed data on improving the efficiency and ease of the data mining process.

4. Conclusion

Machine learning and data mining algorithms automatically extract knowledge from machine-readable information. Unfortunately, their success is usually dependant on the quality of the data that they operate on. If the data is inadequate, or contains extraneous and irrelevant information, machine learning and data mining algorithms may produce less accurate and less understandable results, or may fail to discover anything of use at all.

Thus, Data pre-processing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process.

Data preparation is the most known preprocessing techniques which include data cleaning, data integration, data transformation and data reduction.
Data cleaning routines can be used to filling in missing values, smooth noisy data, identify outliers and correct data inconsistencies.

Data integration combines data from multiples sources to form a coherent data store. Metadata correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute towards smooth data integration.
Data transformation routines conform the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between a small range, such as 0 to 1.0 .

Data reduction techniques such as data cube aggregation, dimension reduction, data compression, numerosity reduction and discretization can be used to obtain a reduced representation of the data, while minimizing the loss of information content. Concept hierarchies organize the values of attributes or dimensions into gradual levels of abstraction. They are a form a discretization that is particularly useful in multilevel mining. Automatic generation of concept hierarchies for categories data may be based on the number of distinct values of the attributes defining the hierarchy. For numeric data techniques such as data segmentation by partition rules, histogram analysis and clustering analysis can be used. Although several methods of data preparation have been developed, data preparation remains an active and important area of research in Data preprocessing of the data mining knowledge discovery process, because quality decisions must be based on quality data.

References

[1] Gregory Piatetsky, “From Data Mining to Knowledge Discovery:An Introduction”, 2012.

[2] Jiawei Han MK, Jian Pei. “Data Mining - Concepts and Techniques”, 2012.

[3] C. Lemnaru,“Strategies for Dealing with Real World Classification Problems”, 2012.

[4] J. Laurikkala, “Instance-based data reduction for improved identification of difficult small classes”, 2002.

[5] R. Kumar, V.K. Jayaraman, B.D. Kulkarni, “An SVM classifier incorporating simultaneous noise reduction and feature selection”, 2005.