ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

A Systematic Approach on Data Pre-processing In Data Mining

S.S.Baskar[1], Dr. L. Arockiam[2], S.Charles[3]

[1]Research scholar, Department of Computer Science, St. Joseph's College, Trichirappalli, India.

[2]Associate Professor, Department of Computer Science, St. Joseph's College, Trichirappalli, India

[3]Assistant Professor, Department of Computer Science, St. Joseph's College, Trichirappalli, India

Abstract: Data pre-processing is an important and critical step in the data mining process and it has a huge impact on the success of a data mining Soil classification. Data pre-processing is a first step of the Knowledge discovery in databases (KDD) process 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. Data pre-processing is challenging and tedious task 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 for soil data sets are also useful for classification in data mining.

Keywords: KDD, Data mining, association rules, Pre-processing algorithms.

1. INTRODUCTION

Data analysis is now important component of any data mining task. This work involves on the basis for investigations in many areas of knowledge, from science to engineering. This is applied to all the part of discipline. Data and datum on a particular area are collected in the form of symbolic and numeric attributes. Analysis of these data and datum give a better understanding of the fact and interest of the subject. In case of soil classification from soil data , the classification of soil from collected data sets is planned, the data analysis involves discovery of classifier and generation of new soil labels from soil data sets. Data pre-processing is an important issue for data mining work. In general and real-world data soil data tend to be incomplete, noise, and inconsistent. These characteristics of the soil data may not yield the expected result for classification. For any data analysing work, the data pre processing is very essential for good result. Various data pre pre-processing techniques are currently available in the real world. Data pre-processing techniques involves data cleaning, data integration, data transformation, and 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 single point data store and it is named as a data warehouse. Data transformation is one of the pre processing techniques is being applied to data analysis work. These techniques are otherwise known as data normalization. Data reduction is also one of pre processing techniques being applied for reduction of data for data analysis. [2]Data reduction can reduce the data size by aggregation, elimination redundant feature, or clustering, for instance. By using the all data pre processing techniques one can improve the quality of data and consequently results quality mining of data with respect to individual interest and efficiency of mining task is improved.

Data pre processing techniques are helpful in OLTP (online transaction Processing) and OLAP (online analytical processing). It is highly useful for any data mining techniques and methods such as classification and clustering. Data pre-processing is important stage for soil classification or prediction by data mining techniques.

By data pre-processing, one can come to study more about the nature of the data and existing hurdles that may exist in the raw data (e.g. irrelevant or missing attributes in the data sets), change the structure of data (e.g. create levels of granularity). This helps to prepare the data for a more efficient and intelligent data analysis, and solve problems such as the problem of very large data sets. There are several different types of problems, related to data collected from the real world that may have to be solved through data pre-processing. In general the raw data collected from the field or non mining sources tend to be in nature

(i) Data with missing, out of range or corrupt elements,
(ii) Noisy data,
(iii) Data from several levels of granularity,
(iv) Large data sets, data dependency, and irrelevant data
(v) Multiple sources of data.

Soil data sets pre-processing

In the field or soil research station, the collected soil data sets for soil classification 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. If the soil data sets from agriculture research and depart of agriculture are considered to be a dirty, mining result lead to poor and not trust worthy. So such dirty data can cause confusion for the mining procedure, resulting in unreliable output This is shown in figure-3

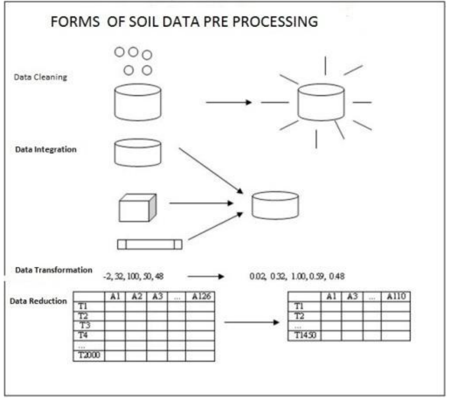

Figure-3

2. NECESSITY FOR SOIL DATA SET PRE-PROCESSING

The soil data sets which collected from the field are very raw and having the tendency of following characteristics. This data has to be processed before analyzing them through data mining techniques.

Incomplete:

When collecting the data of any domain or soil data from the field, there is the possibility of lacking attribute values or certain attributes of interest, or containing only aggregate data. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

Noisy:

Noisy data means that data in the tubles containing errors, or outlier values that deviate from the expected. Incorrect data may also result from inconsistencies in naming conventions or data codes used, or inconsistent formats for input fields, such as date. It is hence necessary to use some techniques to replace the noisy data.

Inconsistent:

Inconsistent means data source containing discrepancies between different data items. Some attributes representing a given concept may have different names in different databases, causing inconsistencies and redundancies. Naming inconsistencies may also occur for attribute values. The inconsistency in data needs to be removed.

Aggregate Information:

It would be useful to obtain aggregate information such as to the soil data sets—something that is not part of any pre-computed data cube in the data warehouse.

Enhancing mining process:

Large number of data sets may make the data mining process slow. Hence, reducing the number of data sets to enhance the performance of the mining process is important.

Improve Data Quality:

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 pre-processing is an important step in the knowledge discovery process, because quality decisions must be based on quality data. Detecting data anomalies and rectifying them can lead to improve the accuracy and efficiency of data analysis.

3. DIFFERENT FORMS OF DATA PROCESSING

Data Cleaning

Data cleaning 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.

a) Missing values: The missing values in the tubles are to be corrected by following measures

i. Ignore the tuple
ii. Fill in the missing value manually
iii. Use a global constant to fill in the missing value
iv. Use the attribute mean to fill in the missing value
v. Use the attribute mean for all samples belonging to the same class.
vi. Use the most probable value to fill in the missing value

Procedure to handle missing data in the soil data sets:

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 to the large soil data sets with many missing values. The missing value in the soil data sets tubles are filled by using a global constant to fill in the missing value and replace all missing attribute values by the same constant. The average or mean value of the attributes in the tubles are used to fill in the missing values One more procedure available is to 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. For example, using the other attributes in the soilr data set, we may construct a decision tree to predict the missing values for income. This is a popular strategy. In comparison to the other methods, it uses the most information from the present data to predict missing values. This procedure is bias the data. The filled-in value may not be correct. It is important to have a good design of databases or data entry procedures which would minimize the number of missing values or errors.

Noisy data

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

i. Binning
ii. Regression
iii. Clustering

Ways to handle Noisy Data

Noise data in the soil data set is considered to be a random error or variance in a measured variable. Data sets having such characteristics lead to poor efficiency and accuracy in the soil classification. This would be corrected by binning method.

Binning: Binning methods have the tendency to smooth a sorted data by consulting its "neighbourhood," that is, the data around it. The sorted soil data 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 means, each value in a bin is replaced by the mean value of the bin.

Smoothing by bin medians can be employed, in which each bin value is replaced by the bin median.

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.

Example

Sorted soil data for Nitrogen content in the soil in Kilograms): 45, 81, 75, 71, 71, 74, 66, 78, 84

Partition into (equal-depth) bins:

Bin 1: 45, 81, 75
Bin 2: 71, 71, 74
Bin 3: 66, 78, 84

Smoothing by bin means:

Bin 1: 67, 67, 67
Bin 2: 72,72, 72
Bin 3: 76, 76, 76

Smoothing by bin boundaries:

Bin 1: 45, 45, 75
Bin 2: 71, 71, 74
Bin 3: 66, 66, 84

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 as shown in the figure-1.

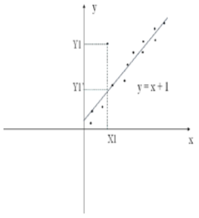

Figure -1

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 outliers as shown in the figure-2

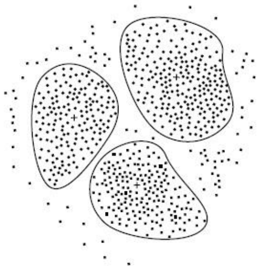

Figure-2

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. For example, the attribute for soil sample number for identification may be referred to as soil id in one data store 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 [2][3]. Additional data cleaning can be performed to detect and remove redundancies that may have resulted from data integration.

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[3]. 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.

Data Reduction:

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

Data aggregation and attribute subset selection

This is removable of irrelevant attributes through correlation analysis and one of the strategy for data reduction.

Dimensionality reduction

This is the reduction of dimension of sets (e.g., using encoding schemes such as minimum length encoding or wavelets)

Numerosity reduction

This is known as "replacing" the data by alternative, smaller representations such as clusters or parametric models and it is considered to be the data reduction techniques[1]. Generalizations with the use of concept hierarchies are one of the data reduction approach by organizing the concepts into varying levels of abstraction. Data discretization is very useful for the automatic generation of concept hierarchies from numerical data.

Data Reduction Techniques

Data cube aggregation

Data cube aggregation operations are applied to the data in the construction of a data cube. For example, data for sales per quarter is known. But, one might be interested in the annual sales, rather than the total per quarter. Thus the data can be aggregated so that the resulting data summarize the total sales per year instead of per quarter. The resulting data set is smaller in volume, without loss of information necessary for the analysis task.

Attribute Subset Selection

Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions). The goal is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. It helps to make the patterns easier to understand.

Dimensionality Reduction

In dimensionality reduction, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data[1]. If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless. If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy.

Wavelet Transforms

It is a lossy compression technique. The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector X, transforms it to a numerically different vector, X', of wavelet coefficients. The wavelet transformed data can be truncated. Thus, a compressed approximation of the data can be retained by storing only a small fraction of the strongest of the wavelet coefficients. An approximation of the original data can be constructed by applying the inverse of the DWT used[1].

The DWT is closely related to the discrete Fourier transform (DFT), a signal processing technique involving sines and cosines. However, the DWT achieves better lossy compression.

Principal Component Analysis:

PCA (also called the Karhunen-Loeve, or K-L, method), searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where k <= n. The original data are thus projected onto a much smaller space, resulting in dimensionality reduction. PCA "combines" the essence of attributes by creating an alternative, smaller set of variables.

Sampling:

Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random sample (or subset) of the data. Suppose that a large data set, D, contains N tuples. Simple random sample without replacement (SRSWOR) of size s: This is created by drawing s of the N tuples from D (s < N), where the probability of drawing any tuple is the same.

Data Discretization and Data summarization

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels reduces and simplifies the original data.

Discretization techniques can be categorized based on how the discretization is performed, such as top-down vs. bottom-up. If the discretization process uses class information, then we say it is supervised discretization, else it is unsupervised.

Top-down discretization or splitting

This process starts by first finding points (split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals.

Bottom-up discretization

This process starts by considering all of the continuous values as potential split-points, removes some by merging neighbourhood values to form intervals, and then recursively applies this process to the resulting intervals.

Concept hierarchies are useful for mining at multiple levels of abstraction. A concept hierarchy for a given numerical attribute defines a discretization of the attribute. They can be used to reduce the data by collecting and replacing low-level concepts (such as age) with higher-level concepts (such as youth, middle-aged, or senior). Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret.

4. OVERVIEW

Data preparation is an important issue for both data warehousing and data mining, as real-world data tends to be incomplete, noisy, and inconsistent. Data preparation includes data cleaning, data integration, data transformation, and data reduction [4] [5].

Data cleaning routines can be used to fill 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 convert the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between small ranges, 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.

Data Discretization and Automatic generation of concept hierarchies For numeric data, techniques such as binning, histogram analysis, and clustering analysis can be used.

5. CONCLUSION

Data preparation is an important issue for data mining, as real world data tends to be incomplete, noisy and inconsistent. Data preparation includes 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 categorise 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 lustering analysis can be used. Although several methods of data preparation have been developed, data preparation remains an active and important area of research

6. REFERENCES

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

[2] Mohd Helmy Abd Wahab, *Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan(1998) Mohamad Mohsin*

[3] Agrawal, Rakesh and Ramakrishnan Srikant, “Fast Algorithms for Mining & Preprocessing Assosiation Rules”, *Proceedings of the 20th VLDB Conference,* Santiago, Chile (1994).

[4] Salleb, Ansaf and Christel Vrain, “An Application of Assosiation *Knowledge Discovery and Data Mining (PKDD) 2000,* LNAI 1910, pp. 613-618, Springer Verlag (2000).

[5] Agarwal,R and Psaila G, Active Data Mining. In Proceedings on Knowledge Discovery and Data Mining (KDD-95), 1995, 3-8 Menl