Pre-Processing Concept in Data Mining

C.Ranjith[1] Dr.M.Praveena[2]
[1]M.Sc. Student [2]Assistant Professor (MCA, M.Phil., Ph.D.)
[1,2]Department of Computer Science Engineering
[1,2]Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore, India

Abstract— Data pre-processing is an essential and critical step in the data mining process and it has a huge impact on the success of a data excavating project. Data pre-processing is a first step of the Knowledge discovery in databases (KDD) process that reduces the involvement of the data and offers better analysis and ANN training. There are a number of different tools and methods used for pre-processing, including: sampling, which selects a characteristic subset from a large population of data; transformation, which handle raw data to produce a single input; denoising , which removes noise from data; control, which organizes data for more efficient access. Data pre-processing includes cleaning, Integration, Transformation, reduction. Pre-processing technique for soil data sets are also useful for classification in data mining.

Keywords: Data cleaning, Data Integration, Data transformation, Data reduction

I. INTRODUCTION

Data mining refers to extracting on involved information or knowledge from large amounts of data foundation Data pre-processing means perform certain tasks before the data to get process. . Data pre-processing is important stage for Data warehousing and Data mining The data pre-processing step begins with a step review of the structure of the data and quality declaration. The seven steps with specific method. Maintain four steps in data pre-processing involves data Cleaning, data Integration, data Transformation and data reduction. Data pre-processing is an important issue for both data warehousing and data mining, as real-world data tend to be imperfect, noise, and inconsistent. Data pre-processing method/techniques are helpful in OLTP (online transaction Processing), OLAP (online analytical processing) and any data mining techniques and methods such as classification and clustering. The quality of data affects the data mining results.

Data integration merges data from multiple sources into a coherent data store. Data cleaning mores can be used to fill in missing values, smooth blaring data, establish outliers, and correct data variation. Data integration merges data from multiple sources into a single point data store and it is formed as a data warehouse. Data transformation is one of the pre-processing techniques is being implanted to data resolution work. These techniques are otherwise known as data normalization. Data reduction can reduce the data size by aggregation, elimination redundant feature, or clustering, for instance.

A. Supervised Learning

In this training data includes both the input and the desired results. The correct results are known and are given in inputs to the model during the learning processes. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables the as it is done in regression analysis.

Knowledge discovery, in accepted with many powerful technologies, lends itself both to abuse and to great prosperity. Moreover, like many technologies, the capacity to loss or to cause offense can often be inadvertent. The broadcast of a rule which finally has a negative brunt on the community bears significant risks, through litigation, adverse attention, loss of standing and so on. However, the number and complexity of rules engender from many data mining systems means that the human post-processing of a data mining run can be long and likely convoluted, leading to suspect rules being unnoticed.

Organizations save data in orders of weight greater than ever before. Data mining techniques such as Classification mining, Association rules, Functional dependency are usable for efficient analysis of sensitive data. These techniques disclose relationships or associations between exact values of categorical variables in extensive data sets. This is a everyday task in many data mining projects. These forceful exploratory techniques have a deep range of applications in many areas of business method and also research. These techniques setup analysts and researchers to discover hidden patterns in large data sets. Sensitive information must be protected against unapproved access. Hence, there is also a need for understanding compromise between disclosed information and enforced needs of the data customer. Sensitive data are inferred from non-sensitive data based on semantics of the application the user has.

B. Unsupervised Learning

The model is not provided with the correct results during the training. It can be used to cluster the input data in classes on the basis of their statistical properties only.

The blocking-based approach aims to put a relatively small number of uncertainties and lessen the confidence of sensitive guideline, but, the problems were: a competitor can regularly infer the hidden values if he applies a smart inference performance. This can be overcome by including many uncertainties, but the process becomes difficulty like, both 0's and 1's must be hidden, because if only 1's was hidden the attacker would simply replace all the uncertainties with 1's and would recover easily the original database.

The steps involved are:

- Identifying generally developing item sets discovering all the transactions that support the item sets.
- Recapture all the available Association rules.
- Finally hide these association rules by declining their support/courage.

The rules are mined according to the user-entered value for confidence. Any rule that has a confidence greater than the entered value is treated to be valuable. The

Association Rules are a lawful input to the next module. From the set of Association Rules, some sensitive Rules are established and then the development of hiding the rules is carried out.

Major Tasks in Data Pre-processing

1) Data cleaning
2) Data integration
3) Data transformation
4) Data reduction

1) Data cleaning

Data cleaning is a starting step of data pre-processing method. Cleaning on the data is one maintain problem in constructing of data warehousing and mining because real world data are very dirty in form of noisy, missing values in tuple and inconsistence. Data cleaning is use to work to clean the data in form by smoothing noisy data, filling in missing values. 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

1) Ignore the tuple
2) Fill in the missing value manually
3) Use a global constant to fill in the missing value
4) Use the attribute mean to fill in the missing value
5) Use the attribute mean for all samples acceptance to the same class.
6) Use the most probable value to fill in the missing value.

– Ignore the Tuple:

This is mostly done when the class label is lost. This method is not very helpful, unless the tuple contains several character with missing values.

b) Fill in the lost value manually:

In general, this approach is inefficient and may not be feasible given a large data set with many missing values. This problem is corrected by following procedures or techniques

1) Binning
2) Regression
3) Clustering

c) Use a global constant to fill in the missing value:

displace all missing attribute values by the same number, such as a label like "Undefinite". If loss values are disturbed by, say, "unknown,"

d) Use the attribute mean to fill in the missing value:

Use the mean value to displace the loss value for a tuple.

e) Use the most feasible value to fill in the lost value:

This may be set with lapse, inference based tools using a Bayesian formalism, or accord tree induction.

2) Data Integration:

Data integration involves integrating data from multiple databases, data cubes, or files. Data mining often requires data integration-the merging of data from multiple data stores. 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. The following points listed different type of data integration technique and their example

a) Data integration

Combines data from multiple basic into a consistent store

b) Schema integration

participate metadata from different sources

c) Detecting and resolving data value conflicts

For the same real world material, attribute values from different sources are different.

3) Data Transformation:

Data transformation operations, such as control and aggregation, are adding data pre-processing procedures that would supply toward the success of the mining process. 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. Then be needed to transform the data one arrange to another arrange use of Smoothing, aggregation, generalization, control technique. Data transformation can involve the following

a) Stable:

which works to remove noise from the data? Such techniques include binning, regression, and clustering.

b) Aggregation:

where summary or aggregation operations are applied to the data.

c) Simplification of the data:

where low-level data are replaced by higher-level concepts through the use of concept hierarchies.

d) Normalization:

It technique useful of ANN classification algorithm. min-max normalization, z-score normalization and decimal scaling are use in normalization technique.

e) Attribute Construction:

Where new character is constructed and added from the given set of sign to help the mining process.

4) Data Reduction:

Data reduction method can be applied to process a cheap representation of the data set that is smaller in volume, yet hard maintains the purity of the real data. Obtain a reduced representation of the data set that is much smaller. The reduced data sets produce the more or less same analytical results as that of real volume. Mining data and easy data are store in a database, data warehouse. We obtain the decrease data volume help of Data cube aggregation, Dimensionality drop, Binning.

a) Data cube aggregate:

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

b) Data compression:

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

c) Dimensionality minimization:

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

d) Numerosity minimization:

Where the data are displaced or estimated by a changed, small data representation such as parametric method or nonparametric methods.

e) Discretization and model hierarchy step:

where rare data values for attributes are displaced by ranges or high conceptual step.

II. PRIVACY

Privacy will be assigned to as an individual's craving and capacity to keep certain information about them hidden from others. Defining privacy in a lawful context has historically been a troublesome process which still hampers new privacy case.

- Secure distribution of data between organizations– Being able to share data for collective profit without compromising competitiveness
- Confidentialisation of openly available data – Establishing that individuals are not detectable from aggregated data and that inferences regarding individuals are forbid
- Anonymization of private data – Individuals and management modifying or randomizing information to conserve privacy.

A. Ethics

Ethics will be invoked to as a set of moral rules or a system of values which models the behaviour of individuals and organizations. It is the perfect way of doing things which as assessed by society and often imposed through law (such as anti-discrimination legislation). To act ethically involves acting for the benefit of the association. It is perfectly available to act unethically yet lawfully.

Two ways can be taken to mitigate the effects of ethical arrangement. Firstly, privacy safeguarding mechanisms can be put in place that maximum access to data, shorten the scope of queries or perturb, hide or eliminate data so that undesired responses do not occur. Unfortunately, this can also change the quantity of a mining system to make beneficial results. The second path is thus to allow unrestricted mining but to apply an alerting process to notify users to the possibly sensitive of rules,

To manage rather than defeat the risk. A large problem that then needs to be overcome with this approach is that sensitivity is situation dependent and thus global measures of sensitivity cannot be accepted. This is the problem accepted by this work.

B. Sensitivity Values and Sensitivity Hierarchies

We store the set of privacy and ethical sensitivity values for all attribute or attribute expense in which we have a major interest. Allowing values to at attribute value level has the advantage of giving a more refined way in which to assign ratings. In our system we arbitrarily used a range 0 . . . 10 with 0 indicating no special sensitivity.

C. Sensitivity Combination Function

A Sensitivity Combination Function (SCF) is used to consider a rule's rating based on each item's privacy and ethical values, their location in the antecedent or subsequent, the number of items in the item set, and so on. It can easily be seen that the manner in which the SCF duty is central to the item-based ratings being accurately translated into ratings for the resulting rules.

III. EXPERIMENTAL RESULTS

There are many techniques of data mining. The most common techniques used in the field of data mining are followings.

A. Artificial Neural Networks

Non-linear predictive models that study through teaching and resemble organic neural networks in formation. This predictive model uses neural networks and asset the patterns from sizable databases.

B. Decision Trees

Set of accord are represented by Tree-shaped design. These decisions make rules for the classification of a dataset under the immense databases. Specific decision tree form includes Classification and Regression Trees (CART) and Chi Square Automated Interaction Disclosure (CHAID).

C. Genetic Algorithms

Development techniques that use progress such as genetic sequence, mutation, and natural collection in a design based on the concepts of progression.

D. Nearest Neighbor Method

A facility that classifies each evidence in a dataset based on a combo of the classes of the k record(s) most related to it in a classical dataset (where k ³ 1). This is frequently called the k-nearest acquaintance technique.

E. Rule Induction

The eradication of useful if-then rules from data based on analytical understanding between different reports of database.

Many of these technologies have been in use for more than a decade in specialized analysis means that work with nearly small volumes of data. These facilities are now expending to integrate precisely with industry-standard data warehouse and OLAP terrace [8]. The appendix to this white paper arranges a glossary of data.

F. Sensitivity

A database of data warehouse keeps perfect information about the activity or company. Some data items of warehouse are sensitive and some are general. The sensitive or confidential information become be removed by other information of database. This separation can be kept by the help of stamp or tag. The access right for sensitive information from database is not for all. There should be a method regarding connection of company sensitive information by each means of data mining.

IV. CONCLUSION

Data pre-processing is a main give out for both data warehousing and data mining, as real-world data tend to be imperfect, blaring, and inconsistent. Data preparation includes data cleaning, integration, data transformation and data reduction. Data cleaning method are clean the noisy of data, perfect on imperfect data and remove unwanted data. Data integration combines data from multiples sources to form a coherent data store. Meta data correlation analysis, data conflict detection, Data alternation method change form of data and data reduction reduces the volume of database by plot integration.

The performance of the new algorithms when compared to the existing approaches is simple and does not require much time for implementation. The side effects in

terms of new rules and lost rules are also minimized. There are currently no systems, that the authors are aware of, that is available to data miners who are concerned about the potential sensitivity of the information that they are extracting from a database. Then Generalization is applied to the masked data. In order to increase the data utility, suppression is avoided for these attributes.

REFERENCES

[1] S.S.Baskar, Dr. L. Arockiam, S.Charles ‖ A Systematic Approach on Data Pre-processing In Data Mining‖ COMPUSOFT, An international journal of advanced computer technology, 2 (11), November-2013 (Volume-II, Issue-XI). ISSN:2320-0790
[2] Jasdeep Singh Malik, Prachi Goyal, Mr.Akhilesh K Sharma ― A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules‖ address-Assistant Professor, IES-IPS Academy, Rajendra Nagar Indore – 452012 , India
[3] Kamlesh kumar pandey, Narendra Pradhan "An Analytical and Comparative Study of Various Data Preprocessing Method in Data Mining"International Journal of Emerging Technology and Advanced Engineering,ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 10, October 2014.
[4] Data Pre-processing & Mining Algorithm, Knowledge & Data Mining & Pre-processing, 3rdedition, Han & Kamber.
[5] Salleb, Ansaf and Christel Vrain, "An Application of Assosiation Knowledge Discovery and Data Mining (PKDD) 2000, LNAI 1910, pp. 613-618, Springer Verlag (2000).
[6] Praveena and Jaiganesh "A Literature Review on Supervised Machine Learning Algorithms and Boosting Process" International Journal of Computer Applications (0975 – 8887) Volume 169 – No.8, July 2017.

