International Journal of u- and e- Service, Science and Technology
Vol.9, No. 11 (2016), pp.421-428
http://dx.doi.org/10.14257/ijunesst.2016.9.11.37

A Theoretical Approach towards Data Preprocessing Techniques in Data Mining – A Survey

Lilly Florence M and Swamydoss D

Adhiyamaan College of Engineering, Hosur, India
lillyflorence.mca@adhiyamaan.in

Abstract

Nowadays, a huge amount of data is stored in system from web, mobile, social media, and health centre or from any production company. Unfortunately these data cannot be used as it is. For knowledge discovery, we must use the quality input. To convert this data into quality input, we have many algorithms and methods in data mining. The preprocessing techniques are used for convert the noise data into quality data. So preprocessing steps are the most important steps in data mining. This paper presents all data preprocessing methods. Here the authors describes the four main process namely data cleaning, data integration, data reduction and data transformation.

Keywords*: data preprocessing, PCA, Wavelet transform, normalization*

1. Introduction

Data analysis is the essential process to take decision in our real life world. It is very basic need in discovery of knowledge from science to engineering and from small business to a large one. Data about a particular concept are stored in the heterogeneous format. Processing these data gives knowledge about the concept. When we analysis and discover knowledge, the data analysis involves discovery of new knowledge. Data processing is the important task in real world concern. In that data preprocessing is the major task in data mining. Data preprocessing involves several tasks like data cleaning, data integration, data transformation, data reduction and data discretization. Data cleaning is the process of fill the missing values, smooth noisy data, identify outliers and remove inconsistencies. Data integration integrate multiple sources of data together. Data transformation involves normalization and aggregation. Data reduction obtains the reduced form of data in volume but produces the same result. Data discretization is part of data reduction but with the numerical data. By applying these preprocessing techniques on the data, we can get quality output. Once we mine the quality data, we can get the quality result. And also we can improve the quality of mining process.

"Where can we apply preprocessing techniques?" all the preprocessing techniques can be applied on real world data. Since real world data tend to be incomplete, noisy and inconsistent. To make these data informative or ready for knowledge discovery, we need to apply the preprocessing techniques. Let us take on scenario that to analyze the online sales data to introduce new discount for a particular product. So we need the sales data which may not be a quality data. To get the quality sales data we have to apply preprocessing techniques on sales data. So in data mining preprocessing is necessary to get a quality of result.

There are many software tools are available for data mining techniques including for preprocessing. But analyzing the data without the help of tools will lead to understand the nature of data. In this paper the authors discussed various preprocessing techniques and analyze with real world example.

ISSN: 2005-4246 IJUNESST

2. Data Preprocessing Techniques

Data have quality if it produce the required result. "*What are the factors tell about quality of data?*" There are many factors comprises of quality including accuracy, completeness, consistency, timeliness, believability and interpretability. Let us try to analyze the sales detail in the sales data warehouse. For our analysis we would like to include certain attributes. For this analysis several tuples have non value, moreover users of the sales database system have reported errors, unusual values and inconsistencies in some transaction. This leads incompleteness of data. So inaccuracy, incompleteness and inconsistent is the reality in many real world database and data warehouse. There are many reason for data inaccurate, incomplete and inconsistency, there is error at data entry, technology limitation and attribute of interest may not be available. Major task in preprocessing are;

- Data cleaning
- Data integration
- Data reduction and
- Data transformation

2.1 Data Cleaning

Data quality is a main issue in quality knowledge discovery. Data quality problem occur anywhere in information system. These problems are solved by using data cleaning process. Generally data cleaning reduces errors and increases the quality of data. It is a time consuming task, but it is an important task for quality result. So we cannot ignore it. Data quality mining is the recent mining approach apply on large data to recovery quality problem in large database. Data cleaning process involves;

a. Fill in missing values
b. Identify outliers and smooth noisy data
c. Correct inconsistent data

a. Fill in missing values

If we deal with missing data, there are many techniques that can be used. Selecting the right technique is depends on the problem domain, data domain and interest of the user. The usual way of handling the missing data is ignore the tuple. If the percentage of such row is high, the performance will be low. The second most commonly used method is use a global constant to fill in for missing values. We can use a constant lie "unknown", "NA" or "Hyphen". But we should choose the constant depends on type of data value, it will be meaningful. The third method for handling missing data is by using attribute mean. Replace the missing values of an attribute with mean value for the attribute in warehouse. Lastly we can use the most probable value to fill in the missing value. Here we can construct a decision tree. The methods discussed here bias the data. The filled in value may or may not be correct. It is important that some time the missing value may not imply the result. Some attributes are intensely left blank if the same are to be provided in the later. Hence we can try out the best to clean the data and to improve the quality of result.

b. Identify outliers and smoothing the noisy data

Noise is a random error or difference in measured variable. The usual technique handle for noise data is binning method, clustering, combined computer human inspection and regression. In binning method, sort the attribute values and partition them into bins. Then smooth it by mean, median or bin boundaries. In the clustering method the attribute values are grouped, then detect the irrelevant values and smooth the data. Outlier may be identified by human or computer. Regression involves fit the values in a straight line using two variables. So one value can be used to predict the other value. Using regression we can smooth out the noise.

c. Correct inconsistent data

In some transaction of database there may be some inconsistent data. Some of inconsistent data may correct manually with the help of other sources. The data must be analyze based on unique rules, consecutive rules and null rules. A unique rule says that the value of target attribute is unique. The consecutive rule say that there is unique value between the lowest and highest value for the attribute. A null rule specify the blank or question mark and how such situation to be handled.

2.2 Data Integration

Data integration combines the data from multiple sources. These sources may include multiple databases, data cubes or flat files. There are some issues while integrating data from multiple sources; they are schema integration, redundancy. Schema integration can be handle by Meta data of databases. Some redundancy can be detected by correlation analysis.

Correlation coefficient,

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \overline{\mathrm{A}})\,(b_i - \overline{\mathrm{B}}\,)}{n\sigma A\sigma B}$$

Where n is the number of tuples $\overline{\mathrm{A}}$ and B are the mean value and σA and σB are standard deviations. If $r_{A,B} > 0$, A and B are positively correlated. If $r_{A,B} = 0$, both A and B are independent. Correlation analysis for discrete data, we can use chi-square test,

$$\mathrm{X}^2 = \sum \left(\frac{Observed\ Value - Expected\ Value}{Expected\ Value} \right)$$

Where Observed value $= \frac{Count\ (A) - Count(B)}{n}$.

The third issue in data integration is detection and resolve of data value conflicts. This is due to different format of data integrated together. When integrating attributes from one database to another, special attention is required in the structure of data. Careful integration of data from multiple source may avoid redundancy and improve the quality of the data.

2.3 Data Reduction

As the size of data set increases, exploration, manipulation and analysis become more complicated and resource consuming. Data reduction process reduces the volume of data set. Strategies for data reduction includes the following;

a. Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube. The cube created at the lower level of abstraction is known as base cuboid. The cube at the highest level of abstraction is referred as apex cuboid.

b. Attribute subset selection, it reducing the data set size by removing the irrelevant attributes. The main aim of attribute subset selection is to find a minimum set of attributes such that the result is very closed by obtained using all attributes. Basic heuristic methods of attribute subset selection includes;

    i. Stepwise forward selection, this procedure starts with an empty set, next the best attribute added with the set, so that each subsequent step, the best attribute is added.

ii. Stepwise backward elimination, this procedure starts with the full set of attributes and remove irrelevant attribute unit the minimized set is reached.

iii. Decision tree induction, a tree is constructed by relevant attributes. The set of attributes appear in the tree form the reduced set of attributes.

c. Dimensionality reduction, in this data encoding are applied to get a reduced form of original data. The data compression process will be in two form; lossy compression and lossless compression. There are two popular and effective lossy method of dimensionality reduction namely wavelet transforms and principal component analysis.

i. Wavelet transforms: it is a linear signal processing techniques when it applied to a data vector X, transform it to a numerically different vector X' of wavelet coefficients. When this technique applied on data it retain a small fraction of the strongest wavelet coefficients. This technique also works to remove noise without smoothing out the features of the data. The general procedure for applying a discrete wavelet transform uses a hierarchical pyramid algorithm that split the data into two halves at each iteration. The method follows the following steps;

a. The length of the data must be power of 2.

b. Two operations namely smoothing and weighted difference applied on each transactions. The results in two set of data.

c. The two functions are recursively applied in the set of data in the previous step until the resulting set of data obtained.

d. The values obtained from the above iteration is the wavelet coefficient of the transformed data.

Wavelet transform have many real world application including the compression of fingerprint images, computer vision, analysis of time series data and data cleaning.

ii. Principal Components Analysis (PCA), PCA searches for k-n dimensional orthogonal vectors that can be used to represent the data where k<n. The actual data are projected on a smaller space, PCA often reveals relationships that were not previously suspected. The steps involved in PCA as follows;

a. The input data must be normalized for a particular range first.

b. PCA compute k orthogonal vector, each point in a direction perpendicular to the others. These vectors are called principal component.

c. The principal component are stored in decreasing order also it serve as a new axes for the data.

d. Since the principal component are stored in decreasing order, it eliminate the weaker component. Using the strongest component we can reconstruct the original data.

d. Neumerosity reduction

We can reduce the data volume by choosing alternative form of data representation using Neumerosity reduction technique. This technique have two forms parametric and no-parametric. For parametric method a model is used to estimate the data. So in this method a data parameter is needed to store instead of storing actual data. But in non-parametric method, it stores reduced representation of the data. This method includes histogram, clustering and sampling.

a. Histogram, it partition the data into disjoint subsets or buckets. If the bucket represent only a single attribute, the bucket is said to be singleton bucket. In this method the bucket are determined and data are partition using partition rule, some of the partition rule including equal-width, equal-frequency, v-optimal and maxdiff partition rule.

b. Clustering, it treat data tuples as objects. They partition the objects into group or clusters. The quality of the cluster depend on the diameter of the cluster. Centroid distance is an alternative measure of cluster quality, it is defined as the average distance of each object from the cluster centroid. The cluster representation replaces the action data.

c. Sampling, it consider a small random sample of data instead of large data set. Random sampling apply on data either with replacement or without replacement of sample. We can also use cluster sample and stratified sample method for sampling. Sampling is most commonly used to estimate the answer to an aggregate query.

e. Data discretization and concept Hierarchy Generation

Data discretization technique can be used to reduce the size of values by dividing the range of values into intervals. Then the intervals are replaces the actual values. This technique may be supervised or unsupervised. A concept hierarchy can be used to reduce the data by replacing low level or high level concept. For example the age attribute value is replaced with middle-age, youth, senior. Concept hierarchy can be used for numerical data and categorical data.

i. Concept hierarchy generation for numerical data

It can be constructed based on discretization. This can be achieved by binning method, histogram analysis, entropy-based discretization, and χ2 merging and cluster analysis. We have already discussed binning and histogram analysis in the previous section, here we have explain entropy-based discretization and χ^2 merging.

a. Entropy-based discretization

It is a supervised, top-down splitting technique. This method select a minimum entropy as a split point, and recursively partitions the resulting interval to get the hierarchal discretization. The basic steps involved in entropy based discretization of an attribute A within the set is as follows;

1. Each point of A can be considered as split point to partition the range of A
2. Entropy based discretization takes place by

$$\text{InfoA(D)} = \frac{|D1|}{|D2|}\, Entropy(D1) + \frac{|D2|}{|D1|}\, Entropy(D2)$$

Where D_1 and D_2 correspond to the tuples in D satisfying the condition split point>=a<split point, D is the data tuples. The entropy (D_1) is calculated by

$$\text{entropy(D1)} = \sum_{i=1}^{m} p_i log_2(p_i)$$

Where p_i is the probability of classes in D_1. Therefore when selecting a split point for attribute A, we select the attribute value that gives minimum expected information requirement.

3. The process of finding the split point is recursively applied to each partition applied until the minimum information requirement is less than the threshold.

b. χ^2 merging Analysis

This is a supervised and bottom-up approach method. The basic concept is the relative class frequency should be consistent within an interval. Initially each distinct value of a numerical attribute is considered as on interval. χ^2

tests are performed for every pair of adjacent intervals then the adjacent interval with least χ^2 values are merged together. This merging process repeat until it reach its threshold value.

ii. Concept hierarchy generation for categorical data.
Categorical data are discrete data, categorical attribute have finite number of distinct value. The concept hierarchy can be generated for categorical data by partial ordering of attributes at schema level by users, specification of a portion of a hierarchy by data grouping and specification of set of attributes.

2.4 Data Transformation

Data transformation involves in transforming the data suitable for mining process. This process does not correct the existing attribute value or add new attribute instead it prepare for mining. Data transformation may involve the following;

- Smoothing
- Aggregation
- Generalization
- Normalization
- Attribute construction

Smoothing is the technique used for cleaning. Aggregation is used in constructing the data cube for analysis of the data. In generalization, the data value will be replaced as general terms for example the age attribute value can be replaced as young, middle-aged and senior. Normalization is useful for classification, there are many methods for normalization. In this paper we discussed three mostly used methods namely min-max normalization, z-score normalization and decimal scaling normalization.

Min-max normalization, in this normalization, the values are normalized within the given range. The formula is given by;

$$\frac{v - min_A}{max_A - min_A}\,(new - max_A - new - \ min_A) + (new - min_A)$$

Where v is the new value in the required range.

Z-Score Normalization, here data is normalized based on the mean and standard deviation. The formula is given by;

$$d' = \frac{d - mean(p)}{std(p)}$$

Where mean(p) is sum of all attributes values of p and std(p) is standard deviation of all values of p.

Decimal scale normalization, this normalization depends on the movement of decimal point of attribute value. The formula is;

$$d' = \frac{d}{10^m}$$, where m is the smallest integer.

3. 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. In this paper we have studies all the methods and techniques used for data preprocessing. We discussed four steps of preprocessing namely, data cleaning, data integration, data reduction and data transformation. Once we applied this technique we can get the quality of data for quality mining process. Now-a-days there are several tools are available to carry out the preprocessing steps. Although several methods are used by analysis, data preprocessing remains an active and important area of research.

References

[1] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," J. Knowledge and Information Systems, vol. 1, no. 1, 1999, pp. 5–32

[2] Cooley, B. Mobasher and J. Srivastava. "Data Preparation for Mining World Wide Web Browsing Patterns," In Journal of Knowledge and Information Systems, vol. 1, no. 1, 1999. pp. 5-32.

[3] Etzioni O. 1996. The World Wide We b: quagmire or gold mine. Communications of theACM36: number 11(November), 65 –68

[4] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Second Edition, Elsevier, 2006.

[5] Katharina Morik.The Representation Race –Preprocessing for Handling Time Phenomena. In ECML , Lecture Notes in Computer Science, pages 4–19. Springer, 2000. Invited talk.

[6] R. Kosala, H. Blockeel. "Web Mining Research: A Survey," In SIGKDD Explorations, ACM press, 2(1): 2000, pp.1-15

[7] Petr Aubrecht and Zden k Kouba. Metadata Driven Data Transformation . In SCI 2001, volume I, pages 332–336. International Institute of Informatics and Systemics and IEEE Computer Society, 2001.

[8] R. Srikant, R. Agrawal. "Mining sequential patterns: Generalizations and performance improvements," In 5th International Conference Extending Database Technology, Avignon, France, March1996, pp. 13-17.

[9] W.W.W Consortium the CommonLogFileformat http://www.w3.org/Daemon/User/Config/ Logging.html#common- Log file-format, (1995)

[10] W3C Extended Log File Format, Available at http://www.w3.org/TR/WD-logfile.html (1996).

[11] Google Website. http://www.google.com

Authors

Lilly Florence M completed her UG and PG at Monanmaniam Sundarnar University, Tirunelveli, India. She done her Ph.D at Mother Teresa Womens University, KOdaiknnal, Tamilnadu. She is currently working as Professor in Adhiyamaan College of Engineering, Hosur, Tamilnadu. Her area of interest is Data Mining, Cloud Computing and Software Reliability.

Swamydoss D Completed his UG and PG at Bharathidasen University, he did his Ph.D. at Anna University, Chennai, India. Currently he is working as a professor in Adhiyamaan College of Engineering, Hosur, Tamilnadu. His area of interest is Networking, Web Management.