Review of Data Preprocessing in Data Mining and Data Summarization and Visualization

Shrikant Patro

School of computer science and engineering, VIT Chennai , Tamil Nadu, India 600 127

Email: Shrikantjagannath.2018@vitstudent.ac.in

Abstract:

Data mining is a process of extracting useful information pattern, trend from the dataset. The model defined over the given dataset is playing important role in decision making task. The quality of data is responsible for quality data mining task. The data is usually susceptible to missing values, noisy data, and inconsistent data outlier data. Data preprocessing is the essential step before machine learning that prepare data suitable for applying different algorithm and improve the quality of data. Preprocessing of data include several technique like cleaning integration, transformation and reduction. Data Summarization is numerical summarization of data that include the measure of central tendency, dispersion and measure of position and outlier detection. Visualization of data through different graphical method.

Introduction

The primary reason behind the Data preprocessing is Knowledge discovery (KDD) process. The efficient implementation of KDD process is possible through preprocessed data. The Raw data are highly vulnerable to missing values, outliers and inconsistent because of their huge size, multiple source and gathering methods. The output of data mining task will be affected due to poor quality data. The appropriate dataset must be selected and preprocessed to enhance reliability. This preprocessing will transform the data into suitable form for data mining process. The mining procedure will be applied include clustering, classification regression, etc. This process will show the trend in the data. This representation of trends in the data can be represented through visualization. This visualization include Bar graph, Pie chart, and scatter plot to represent the relationship between the variables or attributes of the datasets.

Methodology

The process data preprocessing begins with data cleaning include missing values, smooth noise data, recognize outliers and correct inconsistency All the data cleaning routine may not be applicable for the all dataset but sufficient cleaning must be done for the correct analysis of data. If there are record with the missing values either can be ignored during processing or they can be filled. There different technique that can be used to fill the missing values they are described as follows are:

1. **Ignore the tuple:** when very few instances with at least one missing value exist. If there are very few instances with such phenomenon comparatively then it's better to ignore such instances and process data on the basis of the complete records.

2. **Fill the missing value using imputation:** To fill the missing value

we must observe the pattern in the data. The pattern is observed with respect to mean. For Continuous missing variable, if most of the value around mean then mean imputation is used. If the data is left skewed or right skewed then median imputation is used. For categorical variable we must use mode base imputation the most frequently occurring value for the given variable must be used for imputation.

3. **Use the global constant to fill the missing value:** The missing value is replaced with the particular constant which is similar for all record. The Label like NA , nan can be used for the missing value.

4. **Use the most probable value for the missing value:** This approach is used with technique like inference based regression using a decision tree induction or Bayesian formalism.

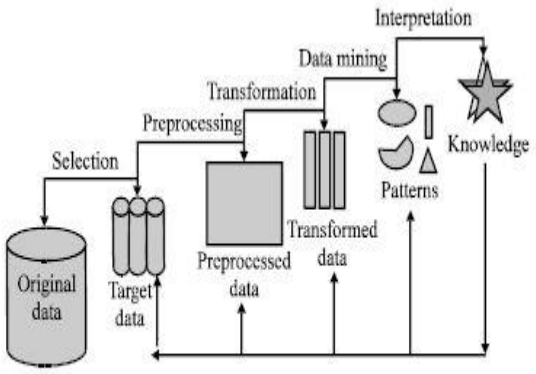

Fig.1: Knowledge discovery steps

The presence of noisy data will affect the mining process a lot. Noise is a random error or deviation measured in specific record which do not follow trends Noisy data means that there is as error in data or outliers which deviates from normal. The pattern followed by majority of the data points is considered as normal trend while the very few instance that deviates from the normal instances is considered as outlier.

In order to understand the nature of the data we can summarize data numerically. This numerical summarization of data include measure of central tendency, measure of dispersion and measure of position and outlier. Determining the measure of central tendency means determining mean, median for the raw data. Also, determining the mode of the variable for the row data. The measure of dispersion include determining range of variable for raw data, standard deviation, variance for the raw data, The measure of the position of the outlier include determining Z-score , interpret percentile , Determine and interpret quartiles , Determine and interpret the interquartile range. Finally, check and set of data for outliers. The organizing the qualitative data and representation through bar graph and pie charts. The primary moves to understand the distribution of the data include following steps: (

(a)Organize the discrete data in table.

(b)Construct histogram for discrete data.

(c)Organizing continuous data in table.

(d) Construct the histogram of continuous data.

(e)Draw Stem and leaf plot.

(f)Draw dot plot

(g)Identify the shape of the distribution.

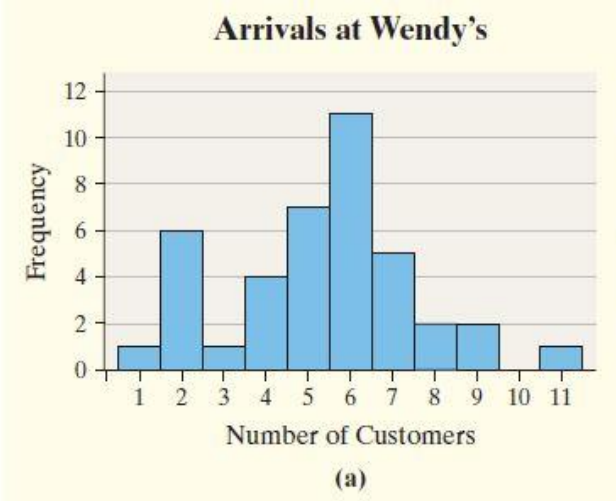

Figure 2

Additional display of quantitative data is achieved in following step

(a)Construct frequency polygon.

(b)Create cumulative frequency and relative frequency table.

(c)Construct frequency and relative frequency ogives.

(d)Draw time series graph.

Database- Woman's Fertility dataset

There are different CSV file for different phenomenon associated with woman fertility dataset. This dataset is collected by WHO with the help of 100 volunteers a semen sample analyzed according to WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits. Dataset characteristics: Multivariate, Attribute characteristics: Real, Associated task: Classification, Regression, Number of instances:100 , Number of attribute : 10, Missing values : NA , Area: Life , Date Donated: 2013-01-17 and number of web hits: 142847.

The further description is specifically regarding the "Blood Transfusion Service Center". This include field like Recency (Months), Frequency (Times), Monetary (c.c. blood), Time (months) and donated. Donated is binary attributed stating in terms of 1 and 0 to represent whether donated or not.

Conclusion:

Final conclusion about the above include, the dataset available for data analysis, Knowledge discovery (KDD) is not always appropriate for analysis. Preprocessing step are required followed by numerical and graphical summarization to understand the pattern and nature of the data. This is building block for data analysis.

References:

[1] "Review of Data Preprocessing Technique in Data Mining " , Suad A. Alasadi and Wasim S.Bhaya, College of information technology , University of Babylon , Iraq

[2]Fertility dataset- UCI Machine learning Repository

[3]Statistics informed decision using data – Michael Sullivan.