Feature Selection Methodology in Quality Data Mining

Nam Ho Soo, Yulius Halim
Department Industrial Engineering, Dongseo University
San 69-1 Jurye2-Dong Sasang-Gu, Busan, Korea

Abstract

In many literatures, data mining has been used as a utilization of data warehouse and data collection. The biggest utilizations of data mining are for marketing and researches. This is solely because of the data available for this field is usually in large amount. The usability of the data mining is expandable also to the production process. While the object of research of the data mining in marketing is the customers and products, data mining in the production field is object to the so called 4M1E, man, machine, materials, method (recipe) and environment. All of the elements are important to the production process which determines the quality of the product. Because the final aim of the data mining in production field is the quality of the production, this data mining is commonly recognized as quality data mining. As the variables researched in quality data mining can be hundreds or more, it could take a long time to reveal the information from the data warehouse. Feature selection methodology is proposed to help the research take the best performance in a relatively short time. The usage of available simple statistical tools in this method can help the speed of the mining.

Keywords: Quality Data Mining, Feature Selection, Data filter, ANOVA, Two-way contingency table

1. Introduction

Data mining, a method for extracting a hidden pattern from a large amount of data, is a powerful tool which is used by industries recently. The advance technology that is flowed into the knowledge of industries gave a great support to the tools especially the advance of information technology. The power of data mining can be seen in the application of predicting the future and approaching the knowledge that is unknown by the means of known knowledge. Data mining make the approach by accessing the data available. It will dig and extracting the important and valuable information and pattern from the abundant raw data. Mostly, the power of data mining is located in the ability in finding a pattern that is unexpected by experts. The fast and steadiness of data mining have replaced the old methods and tools which is time consuming when it faces a large amount of data. Most of the methods in data mining could be applied quickly and easily in the computer so that it will do the rest while the design is done by the experts.

Most big companies have paid a big attention to data collection and storage. They provide a big storage for data so that they can get general information. While the data tell about common information, it can also tell about more valuable information which is buried deep in the data storage. In order to get the data buried, data mining can be used. The question rose such as, "who will respond the A type promotion? Why?" could be answered by means of data mining. It will approach the answer by analyzing the data provided.

Further, the usage of the data mining can be applied also to the production process stage. The data mining in this case can be called as the Quality Data Mining (QDM) since the data mining is done to reveal the information about the quality. Quality improvement program such as TQM, Six sigma and many others get big advantages when quality data mining approach is implemented to the production field. Quality data mining implement the structured step in revealing the behavior of quality in production process and improving the quality. The knowledge of the production process behavior in

a production field should be gathered first to map the problem and find the right solution.

2. The Need of Feature Selection

The system of data mining is mostly about the processing of a lot of data to output the knowledge contained. Many systems, to give the best output, should to make a focus to point of the result. Any parts of the inputs that do not support the output mean that it damages the output. The usage of the computer and its bustling development has shown the need of high speed systems. To build a high speed system, the computation of useless part of the input should be reduced so that the system could work faster. The system of data mining also should be designed to accommodate the performance and right algorithm for a better system. In the computation of the data mining system, the variables or factors should be considered also. There are some factors that give no effect to the output but otherwise, add the errors of the computation in the system.

The variables that are handled by the system could be called as the features of the problem. The features that belong to the problem can be hundreds which will slow down the system if they all are used to give the output of the system. Another problem is that some of the features have no effect to the output as mentioned before. To get the features which give effects to the output and boost the speed of the system, one should select the feature by filtering it. The selection of the features can be done in many ways, such as Delphi decision making which involving the experts to filter the right features, brain storming which involving the experts or any person that have direct contact with the features available and statistical testing.

Delphi decision making and brain storming can be categorized as the people subjective selection. The selection based on experts is good when the experts have the knowledge on all of the features available. Once the features are not mastered by the experts wholly, the selection done by the experts would be questionable. The selection of the features by the experts also could be doubted while the situation is dynamically changed. While the decision making by human or experts is still needed for the last decision should be taken, the usage of statistical testing is very useful to select the key features in a faster way by an analysis based on the data available. The next section discusses more about the methodology proposed by using the statistical approach.

3. Feature Selection Methodology

The system designed here proposes the using of chi-square independence test and analysis of variance (ANOVA) as the statistical test for the selection of features. The usage of the homogenous test and ANOVA is not done serially but it is used alternatively. The difference between both of the tests is located at the data type where it can be divided to two kinds of data, the categorical data and continuous data. When the data type is categorical for the response variables, then the selection of the features is done by the homogenous test. The ANOVA test is used when the data type is continuous. This method can be separated to four kinds of selection based on the data type of the exploratory variables or features and the response variables. The methodology is clearly described in a quadrant as shown in Table 1.

Table 1
Feature Selection Methodology Quadrant

Response variable	Exploratory variables	
	Categorical	Continuous
Categorical	χ^2 test	χ^2 test with transformation
Continuous	ANOVA	ANOVA with transformation

A. Method for categorical exploratory variables and categorical response variable

Chi-square homogenous test can be used when the response variable is the categorical data type. To test the independency between the response variable and the exploratory variable we can apply the two-way contingency table that use test statistics as the significance test. The test statistics then can be compared to the value of the real χ^2 distribution that has α significance level and df (degree of freedom). If the value of the test statistics has over the value of the real chi-square distribution, the two variables is said to violate the independency requirement, thus the two variables has high dependency and important to be understood the dependency between the variables and it is selected as the variables or features that need to be analyzed further.

B. Method for continuous exploratory variables and categorical response variable

As mentioned above, the algorithm used for the categorical response variable is the chi-square homogenous test. This kind of test requires the exploratory variables to be categorical while the variables type here is the continuous data type. To adopt the homogenous test method, one should transform the data into the categorical data type. The transformation could take many ways. One of the most popular methods is the Sturgis' rule (David Lane (2003)). This rule is used mostly at the histogram development. The aim of the transformation is to group the continuous values to a finite number of classes defined. Each class has a certain range of values which will not overlap each other. Generally, the width of the class range for all classes is equal.

Sturgis' rule use the logarithm to determine the number of classes needed to be built. The classes' width range is equal so that the range can be easily computed. The Sturgis' rule is to set the number of the classes as close as possible to $1+\log_2 N$ where N is the number of cases.

The other rule that is possible to be used and similar to the Sturgis' rule is the Rice rule (David Lane (2003)). The Rice rule uses the cube root as the tools to determine the number of classes needed to build. The Rice rule is to set the number of the classes to $2\sqrt[3]{N}$ where N is the number of cases.

The next step after the continuous exploratory variable has transformed to the categorical exploratory variable is to determine whether the exploratory variable or the feature have effect to the response variable or not. If it is found to be significant, then the variable is selected and analyzed in further analysis. The algorithm used to test the significance of the effect of the exploratory variable to the response variable is the chi-square homogenous test which is the same with the first method.

One thing can be added to this method is the column merging. The rule at the first method that mention about the minimum value for the expected value is five can be extended here to the merging for the classes that have expected value lower than five. This merging cannot be applied to the exploratory variables that are the categorical type because the merging will give a wrong meaning to the analysis.

C. Method for categorical exploratory variables and continuous response variable

Analysis of variance should be used when the response variable is the continuous data type. Contrary to the response variable, the exploratory variables are required to be categorical. This method will analyze the exploratory variables one by one and test them independently. This method also did not test the interaction between two or more exploratory variables. Since the needs for the ANOVA are only one by one variable and no interaction, the ANOVA used here is the one way ANOVA. The number of categories determines the number of population in the ANOVA computation. ANOVA have the ability in detecting the significance of the effect of an exploratory variable to the response variable. Thus, this ability is directly the form of feature selection. The one that have significant effect is considered as the feature selected and otherwise.

D. Method for continuous exploratory variables and continuous response variable

This method is similar to the third method above but the requirement of the ANOVA to have categorical exploratory data type is violated here. To overcome the problem, one may transform the continuous response variable to categorical response variable. The method of the transformation is the same to the second method above. The steps in this method can be summarized as follows,

i. transforming the continuous exploratory variable to the categorical exploratory variable
ii. compute the significance of the difference among categorical exploratory variables by ANOVA

The key point in feature selection is the speed of analysis. It is required to have a relatively high speed analysis and give the result that will help the speed of the following analysis of the quality data mining but still keep the most valuable information in the data mining to be processed further. By the requirements, one of the details in the feature selection should be dropped. Some of the continuous data cannot be treated easily by performing the ANOVA method directly. The requirement of the ANOVA method is that the data used is normally distributed. Some of the data may violate the requirement thus cannot use this method. Most of these kinds of data are the defect number that is distributed in the form of Poisson distribution. But in consideration to the speed of the analysis, this method is still considered to be used as an approach.

The whole work of this method is summarized in the Figure 1.

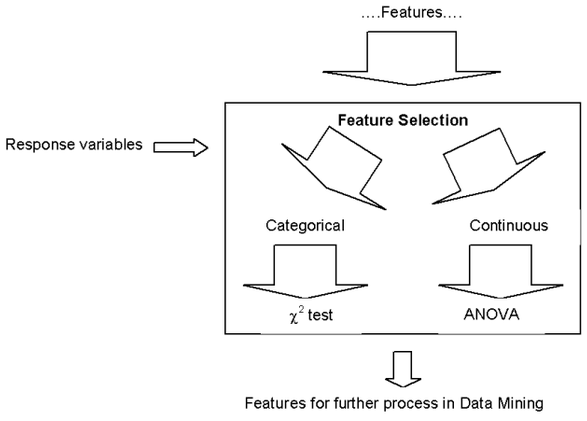

Figure 1 Feature selection process diagram

4. Concluding remarks

Speed or performance has become an important point in quality data mining since the data handled is abundant whether it is in cases or variables. The fast filtering of variables would be very advantageous to cut the time needed for data processing in data mining. Feature selection method in this paper is very useful to achieve fast filtering in data mining. The usage of ANOVA and χ^2 contingency table can help the speed of filtering or screening variables since the computation is relatively simple and the computing time is not too much.

References

David L. (2003). Histograms. Connections Project, June 2003.

Jacob Cohen, Patricia Cohen (1984). Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, March 1984.

Jay L. Devore (1995).Probability and Statistics for Engineering and the Sciences. Duxbury Press, 1995.

H.S., Nam (2003) Quality Data Mining for BOEHydis, Teaching Materials.

Paul S. and Hilbrand K. (2002). Data Cleansing in Oracle 9i, A Practical Guide. IT Consultancies Group.