Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

Receiver Operating Characteristic Analysis by Data Mining

Seong-Won Rhee[1] Jea-Young Lee[2]

Abstract

Data Mining is used to discover patterns and relationships in huge amounts of data. Researchers in many different fields have shown great interest in data mining analysis. Using the classification technique of data mining analysis, the available model for Receiver Operating Characteristic(ROC) method is presented. We present that this may help analyze result of data mining techniques.

1. Introduction

Recently, our ability that create and collect data by great development of computer technology is increasing fast. Thereby corporation, school and public institution which construct database system through computerization could possess bulk data of complex structure. One of technology that extract useful information from such stored bulk data is data mining. So, data mining has become a research area with increasing importance. Researchers in many different fields have shown great interest in data mining analysis. Data mining, which is also referred to as knowledge discovery in databases, means a process of non-trivial extraction of implicit, previously unknown and potentially useful information such as knowledge rules, constraints, and regularities from data in databases (Piatetsky-Shapiro and Frawley, 1991). Much technology are studied in data mining. Among them, spread-studied technology are classification and clustering. The classification technique of data mining analysis is involved very with the logistic regression model and ROC analysis in statistics.

There are many methods to obtain the critical region of tests for hypotheses in statistics. In clinical experiments, some methods to determine the critical region for diagnostic decision were also proposed, such as decision trees, decision matrices, Bayesian analysis and ROC analysis. Receiver operating characteristic analysis has been used increasingly in medical research. Metz (1978) introduced the basic principles of ROC analysis and Hanley and McNeil (1982, 1983) introduced the meaning and use of the area under a ROC curve and proposed the method of comparing areas under ROC curves. Altman (1992), Zweig and Campbell (1993), Schoonjans *et al.* (1995), and Lee and Rhee (1998) studied ROC analysis.

[1] CDO, ResearchNet Co., Sincheon-Dong, Dong-Ku, Taegu, 701-020, Korea.

[2] Associate Professor, Department of Statistics, Yeungnam University, Kyongsan, 712-749 Korea

In this paper, we study an ROC analysis by the logistic model using the results of the classification technique, the neural network, in data mining analysis for the normality test.

2. ROC Analysis as a tool for evaluating new diagnostic tests

Receiver operating characteristic (ROC) curves are used to evaluate diagnostic tests when test results are not binary. Originally, this method was developed in the field of signal detection theory for the purpose of aiding radar operators during World War II to distinguish a signal from the chaotic sounds of war. Today it is receiving a considerable amount of attention from medical researchers as a tool for evaluating new diagnostic tests. They describe the inherent capacity of the test for distinguishing between truly diseased and nondiseased subjects. In biomedical applications, the two states are often referred to as diseased and nondiseased, or D+ and D- for short. Central to this analysis is the ROC curve, which displays diagnostic accuracy as a series of pairs of performance measures. Each pair consists of a true positive fraction(TPF) and the corresponding false positive fraction(FPF) for the given definition of test (t) positivity, t+. These fractions are calculated from the D+ and D- respectively, TPF as the proportion of (t+, D+) among those D+, and FPF as the proportion of (t+, D-) among those D-. In the medical literature, the term TPF is called sensitivity and the complement of the FPF is called specificity. The ROC curve is to plot in a series of the sensitivity versus 1-specificity(FPF) pairs. So a more accurate test will be located on an ROC curve closer to the top left corner than a less accurate one.

3. Classification Technique-Neural networks

Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification rules, decision trees, mathematical formulae, or neural networks. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units (Han and Kamber, 2001). Neural networks exist in many different varieties. Three of the most commonly used models are the self organizing map which often called the Kohonen net, the back-propagation net, and the Boltzmann Machine.

REFERENCES

[1] Altman, D. G. (1992). *Practical statistics for medical research,* London : Chapman and Hall.

[2] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a

receiver operating characteristic (ROC) curve, *Radiology*, Vol. 143, 29-36.

[3] Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases, *Radiology*, Vol. 148, 839-843.

[4] Lee, J.-Y. and Rhee, S.-W. (1998). Q-Q, P-P 플롯의 변동 통계량에 대한 ROC 분석, *The Korean Communications in Statistics*, Vol. 5, No. 1, 205-215.

[5] Han, J. and Kamber, M. (2001). *Data Mining : Concepts and Techniques*, Academic Press.

[6] Metz, C. E. (1978). Basic principles of ROC analysis, *Seminars in Nuclear Medicine*, Vol. 8, 283-298.

[7] Piatetsky-Shapiro, G. and Frawley, W. J. (1991). *Knowledge Discovery in Databases*, AAAI/MIT Press.

[8] Schoonjans, F., Zalata, A., Depuydt, C. E., and Comhaire, F. H. (1995). MedCalc: a new computer program for medical statistics, *Computer Methods and Programs in Biomedicine*, Vol. 48, 257-262.

[9] Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, Vol. 39, 561-577.