Application of R to Data Mining in Hospital Information Systems

Shusaku Tsumoto and Shoji Hirano
Department of Medical Informatics,
Shimane Medical University, School of Medicine,
Enya-cho Izumo City, Shimane 693-8501 Japan

Abstract

Hospital information systems have been introduced since 1980's and have stored a large amount of data of laboratory examinations. Thus, the reuse of such stored data becomes a critical issue in medicine, which may play an important role in medical decision support. On the other hand, data mining from the computer science side emerged in early 1990's, which can be viewed a re-invention of what statisticians, especially those on exploratory data analysis, had proposed in 1970's. The main objective of the present data mining is to extract useful patterns from a large mount of data with statistical and machine learning methods. Especially, it has been reported in medical field that a combination of these two methodologies are very useful: Machine learning method is useful for extracting "hypotheses" which may not be significant from the viewpoint of statistics. After deriving these hypotheses, statistical analysis is used for its validity. Thus, it has been expected that combination of these two methodologies will play an important role in medical decision support, such as intra-hospital infection control, detection of risk factors.

This paper report an application of R to data mining in hospital information systems. As a preliminary tool, we developed a package for data mining in medicine, including the following procedures: (1) Interface for Statistical Analysis, which is based on Rcmdr. (2) Rule Induction, which supports associaton and rough set-based rule induction method. (3) Categorization of Numerical Variables: detection of cut-off point is very important in medical diagnosis. Several methods proposed in data mining and medical decision making are implemented. (4) Clustering, based on R packages. Also, this package supports rough clustering, which gives a indiscernbility-based clustering with iterative refinement of equivalence relations in a data set. (5) Temporal Data Mining (Multiscale Matching and Dynamic Time Warping): these methods have been introduced for classification of long-term time series data. These methods output a distance between two sequences, which can be used for clustering methods. The usage of R gives the following advantages: (a) Rough set methods can be easily achieved by fundamental R-functions, (b) Combination of rough set methods and statistical packages are easily acheived by rich R-packages. In the conference, several aspects of this package and experimental results will be presented.

Keywords: Data Mining, Hospital Information System, Time Series Analysis