Using feature selection as accuracy benchmarking in clinical data mining.

ABSTRACT

Automated prediction of new patients' disease diagnosis based on data mining analysis on historical data is proven to be an extremely useful tool in the medical innovation. There are several studies focusing on this particular aspect. The objective of this study is two-fold. First, we look into three different classifiers, which are the Naïve Bayes, Multilayer Perceptron (MLP) and Decision Tree J48 to predict the diagnosis results. Next, we investigate the effects of feature selection in such experiments. We also compare the experimental results with the study of Comparative Disease Profile (CDP) using the same dataset. Results have shown that the Naive Bayes provides the best result in terms of accuracy in our experiments and in comparison with CDP. However, we suggest using Multilayer Perceptron since the variables used in our experiments are inter-dependent among each other. In addition, MLP has shown better accuracy than CDP.

Keyword: Data mining; Healthcare; Heart disease; Multilayer perceptron; Naive bayes; J48.