Abstract

In this paper, twenty well known data mining classification methods are applied on ten UCI machine learning medical datasets and the performance of various classification methods are empirically compared while varying the number of categorical and numeric attributes, the types of attributes and the number of instances in datasets. In the performance study, Classification Accuracy (CA), Root Mean Square Error (RMSE) and Area Under Curve (AUC) of Receiver’s Operational Characteristics (ROC) is used as the metric and come up with some findings: (i) performance of classification methods depends upon the type of dataset variables or attributes such as categorical, numeric and both (mixed), (ii) performance of classification methods on categorical attributes is superior than on numeric attributes of a dataset, (iii) classification accuracy, RMSE and AUC of a classification method depends on the number of instances in datasets, (iv) classification performance decreases in case of instances decreases for both categorical as well as numeric datasets, (v) top three classification methods are established after comparing the performance of twenty different classification methods for the categorical, numeric and both (mixed) attribute datasets, (vi) out of these twenty different classification
Data Classification based on Decision Tree, Rule Generation, Bayes and Statistical Methods: An Empirical Comparison

methods Bayes Net, Naïve Bayes, Classification Via Regression, Logistic Regression and Random Forest method performs best on these medical datasets.

References

1. Jiawei Han, Micheline Kember, Jian Pei, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.

Index Terms

Computer Science

Information Sciences

Keywords

Data Mining; Classification; Classification Accuracy; RMSE; ROC; Confusion Matrix