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
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
2. Andrew P. Bradley, “The use of area under ROC curve in evaluation of machine learning
4. Sholom M. Weiss, Ioannis Kapouleas, “An Empirical Comparison of Pattern Recognition,
   Neural Nets, and Machine Learning Classification Methods”, Machine Learning.
5. Aik Choon Tan, David Gilbert, “An empirical comparison of supervised machine learning
   Experimental Investigation of the Effect of Discrete Attributes on the Precision of classification
7. Jin Huang, Jingjing Lu, Charles X. Ling, “Comparing Naive Bayes, Decision Trees, and
8. Yong Soo Kim, “Comparison of the decision tree, artificial neural network and linear
   regression methods based on the number and types of independent variables and sample size”,
10. Jae H. Song, Santosh S. Venkatesh, Emily A. Conant, Peter H. Arger, Chandra M.
    Sehgal, “Comparative Analysis of Logistic Regression and Artificial Neural Network for
11. R P Datta, Sanjib Saha, “Applying rule based classification techniques to medical
    databases: An empirical study”, International Journal of Business Intelligence and Systems
    77-No.14, September 2013.
14. WEKA 3.6.9 java based GUI tool popularly used for machine learning and knowledge
    analysis (http://www.cs.waikato.ac.nz/~ml/weka/). Provided by the Machine Learning Group at
    the University of Waikato, Hamilton, New Zealand, 1999-2013.
    Classifiers”, Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, 1995,
    pp. 338-345.
    1993, San Mateo, CA.

Index Terms

Computer Science
Information Sciences

Keywords

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