Abstract

The medical data are multidimensional and hundreds of independent features in these high dimensional databases need to be considered and analyzed, for valuable decision-making information in medical prediction. Most data mining methods depend on a set of features that define the behavior of the learning algorithm and directly or indirectly influence the complexity of
the resulting models. Hence, to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed. Feature selection is a preprocessing step which aims to reduce the dimensionality of the data by selecting the most informative features that influence the diagnosis of the disease. We propose a feature selection embedded Hybrid Prediction model that combines two different functionalities of data mining; the clustering and the classification. The F-score feature selection method and k-means clustering selects the optimal feature subsets of the medical datasets that enhances the performance of the Support Vector Machine classifier. The performance of the SVM classifier is empirically evaluated on the reduced feature subset of Diabetes, Breast Cancer and Heart disease data sets. The proposed model is validated using four parameters namely the Accuracy of the classifier, Area Under ROC Curve, Sensitivity and Specificity. The results prove that the proposed feature selection embedded hybrid prediction model indeed improve the predictive power of the classifier and reduce false positive and false negative rates. The proposed method achieves a predictive accuracy of 98.9427% for diabetes dataset, 99% for cancer dataset and 100% for heart disease dataset, the highest predictive accuracy for these datasets, compared to other models reported in the literature.

**Reference**

- Paraskewas Orfanidis and David J. Russomanno, 2008. Preprocessing enhancements to improve data mining algorithms. International Journal of Business Intelligence and Data Mining 3(2), 196-211.
An Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets (pp. 856–863).

An Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets

Index Terms

Computer Science  
Pattern Recognition

Key words

Medical Data Mining  
F-score  
Support Vector Machine Classifier  
Accuracy  
Sensitivity  
Specificity  
Area Under ROC Curve