Abstract

Several Machine Learning Classification Techniques have been applied in predicting Protein Localization sites of E.coli using a number of techniques. However, research done is limited to no prediction of Localization sites of Proteins on Ecoli0s minimal dataset with the most informative features obtained using different feature selection techniques. This study investigated several Machine learning Classification and Feature Selection Techniques as applied on Ecoli0s minimal dataset. The implementation of classifiers aided in predicting localization sites of E.coli0s minimal subset using its informative features obtained by feature selection techniques. Results were achieved in four parts including; (Data Collection, Cleaning and Preprocessing), Feature selection where the most informative features are selected, Classification where prediction of the localization of proteins is done and then Evaluation of the Classifiers to assess their performance using a number of measures including Accuracy from Cross-validation, and AUROCC to enable in recommending the best Classifier at the end. Among the Classifiers used, Extra Tree Classifier and Gradient Boosting are seen to be the best at performance followed by Random forest as seen from Precision, Recall and F-measure.
scores. AdaBoost is the worst at 83%.

References

7. HT Debas, R Laxminarayan, and title = SE Straus’.
22. L Liqi, Y Sanjiju, and X Weidong et.all. Prediction of bacterial protein subcellular localization by incorporating various features into chou’s pseaac and a backward feature
26. title = Q Yanjun’.

Index Terms

Computer Science Information Sciences

Keywords

Predicting, Ensemble and Non Ensemble Classifiers, and Machine Learning Techniques