Paper Title: Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction

1. Summary

Despite of abundant data available within the healthcare systems, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper uses data mining techniques that are in use in today’s medical research particularly in Heart Disease Prediction.

Three different supervised machine learning algorithms i.e. Naive Bayes, K-NN, Decision List algorithm have been used for analysing the dataset. Tanagra tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared.

After number of experiments, outcome reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute enough for heart disease prediction.

1. Main strengths and contributions:

* Decision support systems aids in achieving clinical tests at a reduced cost.
* Genetic algorithm used reduces the actual data size to get the optimal subset of attributes enough for heart disease prediction.
* Tanagra tool used is a powerful system that contains clustering, supervised learning, meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms.

1. Main weaknesses:

* Attribute values need to be normalized before clustering to avoid domination between high and low value attributes.
* Real data is not used to validate the algorithms.

1. Other interesting thoughts raised by the paper: like,

* Incorporation of genetic algorithm in decision tree outperforms the other data mining techniques and gives the best accuracy.