**Booster in High Dimensional Data Classification**

**ABSTRACT:**

Classification problems in high dimensional data with a small number of observations are becoming more common especially in microarray data. During the last two decades, lots of efficient classification models and feature selection (FS) algorithms have been proposed for higher prediction accuracies. However, the result of an FS algorithm based on the prediction accuracy will be unstable over the variations in the training set, especially in high dimensional data. This paper proposes a new evaluation measure Q-statistic that incorporates the stability of the selected feature subset in addition to the prediction accuracy. Then, we propose the Booster of an FS algorithm that boosts the value of the Q-statistic of the algorithm applied. Empirical studies based on synthetic data and 14 microarray data sets show that Booster boosts not only the value of the Q-statistic but also the prediction accuracy of the algorithm applied unless the data set is intrinsically difficult to predict with the given algorithm.

**EXISTING SYSTEM:**

* One often used approach is to first discretize the continuous features in the preprocessing step and use mutual information (MI) to select relevant features. This is because finding relevant features based on the discretized MI is relatively simple while finding relevant features directly from a huge number of the features with continuous values using the definition of relevancy is quite a formidable task.
* Several studies based on resampling technique have been done to generate different data sets for classification problem and some of the studies utilize resampling on the feature space.
* The purposes of all these studies are on the prediction accuracy of classification without consideration on the stability of the selected feature subset.

**DISADVANTAGES OF EXISTING SYSTEM:**

* Most of the successful FS algorithms in high dimensional problems have utilized forward selection method but not considered backward elimination method since it is impractical to implement backward elimination process with huge number of features.
* A serious intrinsic problem with forward selection is, however, a flip in the decision of the initial feature may lead to a completely different feature subset and hence the stability of the selected feature set will be very low although the selection may yield very high accuracy.
* Devising an efficient method to obtain a more stable feature subset with high accuracy is a challenging area of research.

**PROPOSED SYSTEM:**

* This paper proposes Q-statistic to evaluate the performance of an FS algorithm with a classifier. This is a hybrid measure of the prediction accuracy of the classifier and the stability of the selected features. Then the paper proposes Booster on the selection of feature subset from a given FS algorithm.
* The basic idea of Booster is to obtain several data sets from original data set by resampling on sample space. Then FS algorithm is applied to each of these resampled data sets to obtain different feature subsets. The union of these selected subsets will be the feature subset obtained by the Booster of FS algorithm.

**ADVANTAGES OF PROPOSED SYSTEM:**

* Empirical studies show that the Booster of an algorithm boosts not only the value of Q-statistic but also the prediction accuracy of the classifier applied.
* We have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and Q-statistic. Especially, the performance of mRMR-Booster was shown to be outstanding both in the improvements of prediction accuracy and Q-statistic.

**SYSTEM REQUIREMENTS:**

**HARDWARE REQUIREMENTS:**

* System                            :         Pentium Dual Core.
* Hard Disk                       :         120 GB.
* Monitor                          :         15’’ LED
* Input Devices                 :         Keyboard, Mouse
* Ram                                :         1GB.

**SOFTWARE REQUIREMENTS:**

* Operating system           :         Windows 7.
* Coding Language           :         JAVA/J2EE
* Tool                               :         Eclipse
* Database                        :         MYSQL