Supporting of Category in High Measurement Data

G.Sravanthi[1], Thimma Reddy[2]

[1]*M.Tech Student, Dept of SE, Joginpally B.R. Engineering College. TS, India.*

[2] *Assoc. Prof, Dept of SE, Joginpally B.R. Engineering College. TS, India.*

Abstract- The problems of classification in high-dimensional data with a small number of observations became more common, especially in microarray data. In the last two decades, many effective classification models and feature selection algorithms (FS) have been proposed for high precision accuracy. However, the result of the FS algorithm based on the accuracy of the unstable prediction will be in the differences in the training package, especially in high-dimensional data. This document proposes a new evaluation of the Q statistic that includes the stability of the set of specific sub properties, as well as the accuracy of the prediction. Then, we suggest the support for the FS algorithm that improves the value of the Q statistic for the applied algorithm. Experimental studies based on synthetic data and 14 sets of microarray data show that Booster not only improves the statistical value Q, but also the prediction accuracy of the predicted algorithm unless it is difficult to determine the data set intrinsically using the specified algorithm.

Keywords- High dimensional data classification, feature selection, stability, Q-statistic, Booster

I. INTRODUCTION

High-dimensional data have become more common in many practical applications, such as data extraction, automated learning and analysis of microarray gene expression data. The typical microarray data available to the public contains tens of thousands of small-sized features for the sample and the volume of features that are considered in the microarray data analysis is increasing. The statistical classification of the data with a large number of characteristics and a small sample size (under the problem of the sample) is a fundamental challenge [1]. We found a surprising finding that the simple and common analysis of linear linear discrimination can be as weak as random guessing where the number of features increases [2], [4]. As mentioned in [3], most of the characteristics of the high-dimensional data are not related to the objective characteristic and the proportion of related characteristics or the percentage of genes that are organized or organized in relation to the appropriate natural tissues is only 2% to 5% The search for related characteristics simplifies the learning process and increases the accuracy of the prediction. However, the result should be relatively strong for differences in training data, especially in the biomedical study, since field experts will invest considerable time and effort in this small set of selected characteristics. Therefore, the proposed selection should provide not only a high forecasting potential, but also a high stability of choice. Hot topics in automated learning [5],. One method that is often used is to degrade the reputation of persistent features in preprocessing and the use of mutual information (MI) to identify related characteristics. Because finding related characteristics in an unreliable MI base is relatively simple, while finding features directly related to a large number of features with persistent values by defining links is a formidable task. You can use the methods used in the problems of statistical selection of the variable, such as the front and rear selection and their combination in fixed service problems. Most FS algorithms successful in high-dimensional problems used the forward selection method, but they were not considered a delay method because it is not practical to perform the elimination of actions with a large number of functions. There is a serious problem with the forward selection. However, the decrease in the resolution of the main characteristic may result in a completely different subset of the characteristic, so the stability of the selected feature set will be very low, although the selection may result in a very high resolution [6] . This is known as the stability problem in FS. Research in this area is a relatively new area, where the design of an effective way to obtain a more stable subset of high precision is a difficult area of research. This document suggests a Q statistic to evaluate the performance of the FS algorithm with a workbook. This is a hybrid scale for the accuracy of the workbook prediction and the stability of the selected functions. The Booster article then suggests choosing a subset of a particular FS algorithm. The basic idea of Booster is to obtain multiple data sets from the original data set by resampling the sample area. The FS algorithm is applied to each of the data sets that were sampled for different subsets of characteristics. The combination of these specific subsets will be the subset of the Booster of FS algorithm. Scientific studies indicate that the compatibility with an algorithm not only improves the statistical value Q, but also the accuracy of the prediction of the workbook applied. Several studies based on the sampling technique have been conducted to create different datasets for the classification problem and some studies are using sampling sampling in the characteristics area. The purpose of all these studies is to predict the accuracy of the classification without considering the stability of the set of specific sub-characteristics. This article is organized as follows. Section 2 describes the preprocessing steps to find related weak characteristics based on the t-test and to eliminate inappropriate MI-based functions. It introduces a new evaluation standard for Q statistics and investigates its characteristics. Section 4 gives the reinforcement algorithm and provides some background

II. RELATED WORK

One method that is often used is to degrade the reputation of persistent features in pre processing and the use of mutual information (MI) to identify related characteristics. Because finding related characteristics in an unreliable MI base is relatively simple, while finding features directly related to a large number of features with persistent values by defining links is a formidable task. Several studies based on the sampling technique have been conducted to create different datasets for the classification problem and some studies are using sampling sampling in the characteristics area. The purpose of all these studies is to predict the accuracy of the classification without considering the stability of the set of specific sub-characteristics.

III. RESEARCH METHODOLOGY

This document suggests a Q statistic to evaluate the performance of the FS algorithm with a workbook. This is a hybrid scale for the accuracy of the workbook prediction and the stability of the selected functions. The Booster article then suggests choosing a subset of a particular FS algorithm. The basic idea of Booster is to obtain multiple data sets from the original data set by re sampling the sample area. The FS algorithm is applied to each of the data sets that were sampled for different subsets of characteristics. The combination of these specific subsets will be the subset of the Booster of FS algorithm. Experimental studies indicate that the support of the algorithm not only improves the statistical value Q but also the accuracy of the prediction of the workbook applied. We observed that the Booster classification methods did not have a significant impact on the accuracy of the prediction and the Q statistic. In particular, the performance of the RMR-Booster was different in the improvement of the accuracy of the prediction and the Q statistic.

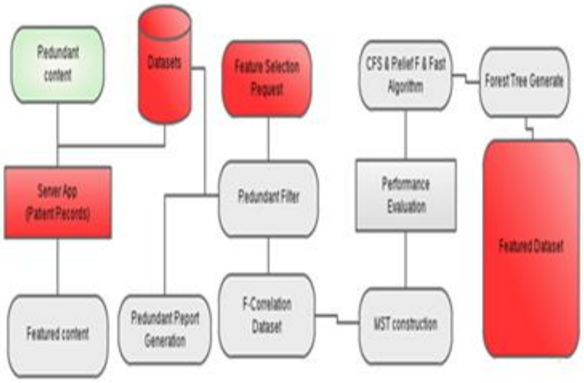

Fig.1: The system Architecture

Algorithm Evaluation process of FS

Input: FS algorithm s,
number of folds k, original data set D and k-folded data subsets D_i, i = 1; . . . ; k.

```
for i =1 to k do
    D_i =D- Di # apply D_i to s-Booster5
    V  i f-Booster$_5$(D_i)
    ai ← Classifier (Di)
end for
Q compute Q using k-pairs of (V * i, ai )
```

IV. CONCLUSION

It has been observed that if the FS algorithm is effective but can not obtain a high accuracy or Q statistic for some specific data, FS Booster will reinforce the performance algorithm. However, if the FS algorithm itself is not effective, Booster may not perform well. The performance of the reinforcement depends on the performance of the applied FS algorithm. If Booster does not provide high performance, it means two possibilities: the data set is basically difficult to predict or the applied FS algorithm is not effective with the established data set. Therefore, Booster can also be used as a criterion to evaluate the performance of the FS algorithm or to evaluate the difficulty of a data set for classification.

V. REFERENCES

[1]. T. Hastie, The Elements of Statistical Learning, New York, NY, USA: Springer, 2009.

[2]. P. J. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations," Bernoulli, vol. 10, no. 6, pp. 989–1010, 2004.

[3]. D. Dembele, "A flexible microarray data simulataion model," Microarrays, vol. 2, no. 2, pp. 115–130, 2013.

[4]. J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," Ann. Statist., vol. 36, no. 6, pp. 2605–2637, 2008.

[5]. T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with

ensemble feature selection methods," Bioinformatics, vol. 26, no. 3, pp. 392–398, 2010.

[6]. C. Kamath, Scientific data mining: a practical perspective, Siam, 2009.