International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

A new evaluation measure Q-statistic that incorporates the stability of the selected feature subset

G.Raja & P. Veera Muthu

Post Graduate Student, M.Sc., Computer Science, Besant Theosophical College, Madanapalle.
e-mail: gaddamrajkumar74@gmail.com

Assistant Professor, Besant Theosophical College, Madanapalle.
e-mail ID:; er.veera86@gmail.com

ABSTRACT *Classification problems in high-dimensional data with a small number of observations became more common, especially in micro-study data. Over the past two decades, many basic classification models and feature selection algorithms (FS) have been proposed to increase the accuracy of predictions. However, the result of the FS algorithm based on predictive accuracy will be unstable on differences in the training package, especially in high-dimensional data. This paper proposes a new measurement of the Q-statistic that includes the persistence of the subset of the specific features as well as the accuracy of the prediction. Next, we propose an enhanced FS algorithm that enhances the Q value of the algorithm applied. Experimental studies based on synthetic data and 14 microarray data sets show that Booster not only enhances the Q statistic value but also the predictive accuracy of the applied algorithm unless the data set is intrinsically correct to predict the use of the given algorithm.*

Index Terms—High dimensional data classification, feature selection, stability, Q-statistic, Booster

INTRODUCTION

The presence of high-dimensional data has become more common in many practical applications such as data extraction, machine learning and gene expression analysis of microarray data. The typical microarray data available to the public contains tens of thousands of features with a small sample size, and the size of the features being considered in microarray data analysis is increasing. The statistical classification of data with a large number of features and small sample size (a small problem) is a fundamental challenge [29]. A striking conclusion has been found that

Fisher's simple and popular linear differentiation can be as weak as random guessing as the number of features becomes larger [7], [16]. As mentioned in [14], [59] most of the characteristics of high-dimensional microarray data have nothing to do with the target property, the proportion of relevant features or the percentage of high-regulated or poorly regulated genes compared to normal tissue. Is only 2% 5%. Finding relevant features makes learning easier and predictive. However, the effect should be relatively strong for changes in training data, especially in the biomedical study, where field experts will invest considerable time and effort in this small set of selected features. Hence, the proposed choice must provide them not only with high predictive potential but also with high stability of choice [40].

New Feature Selection Proposal This paper proposes a Q statistic to evaluate the performance of the FS algorithm with the class. This is a hybrid scale of second-class prediction accuracy and stability of specific features. The Booster paper then suggests choosing a subset of a specific FS algorithm. The basic idea of the Booster program is to obtain several data sets from the original data that has been set by reconfiguring the sample area. The FS algorithm is then applied to each set of data sets that have been reformulated for different subsets of features. The combination of these subgroups will be the subset of the features obtained by the Booster of FS algorithm.

Existing System:

One method that is often used is to distinguish the persistent features of the pre-processing step and the use of shared information (MI) to identify relevant features. This is because finding related features based on a convergent MI is relatively simple, while finding features directly related to a large number of features with continuous values using the definition of relevance is a huge task.

Many studies have been conducted on the method of restructuring to create different data sets for the classification problem and some studies are used in the restructuring of the feature area.

The purpose of all these studies is to accurately predict the classification without considering the stability of a subset of the specific characteristics.

Disadvantages of existing system:

Most of the successful FS algorithms have benefited from the high-resolution problems of the forward-looking method but have not been considered a back-to-back method since it is impractical to implement the previous removal with a large number of features.

However, the serious intrinsic problem of choosing forward, however, flipping in the initial feature resolution may result in a completely different subset, so the stability of the set of specific features will be very low, although the selection may achieve very high accuracy. Creating an effective way to obtain a more consistent, high-precision subset is a challenging area of research.

Proposed System:

This paper suggests a Q statistic to evaluate the performance of the FS algorithm with a workbook. This is a hybrid scale of the prediction accuracy of the workbook and the stability of the specific features. The Booster paper then suggests choosing a subset of a specific FS algorithm.

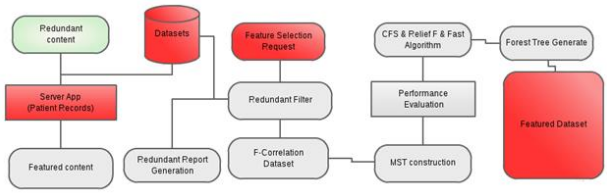
The basic idea of the Booster program is to obtain several data sets from the original data that has been set by reconfiguring the sample area. The FS algorithm is then applied to each set of data sets that have been reformulated for different subsets of features. The combination of these subgroups will be the subset of the features obtained by the Booster of FS algorithm.

Advantages of the proposed system:

Experimental studies show that the algorithm supports not only the value of the Q statistic but also the accuracy of the prediction in the applied workbook.

We observed that the Booster classification methods do not have a significant impact on the accuracy of the prediction and the Q statistic. In particular, the MRMR-Booster performance was distinct in both Q & A improvements.

SYSTEM ARCHITECTURE:



CONCLUSION

This research suggests a statistical measure Q- evaluates the performance of the FS algorithm. Q-statistic calculations for stability of a subset of specific properties and accuracy of prediction. Booster suggested to enhance the performance of the existing FS algorithm. Experimentation with synthetic data and 14 microarray data sets showed that the proposed booster improves predictive accuracy and Q statistic for known FS algorithms: FAST, FCBF, and mRMR. We also noted that the classification methods applied to Booster did not have a significant impact on the accuracy of the prediction and the Q statistic. In particular, MRMR-Booster demonstrated that it was prominent both in improving the accuracy of the prediction and the Q- statistic. It was noted that if the FS algorithm is effective but it is not possible to obtain high accuracy performance or Q statistic for some specific data, Booster of the FS algorithm will enhance performance. However, if the FS algorithm itself is not effective, Booster may not be able to achieve high performance. Booster performance depends on the performance of the applied FS algorithm.

Author Details

Student Name : G.Raja

Post Graduate Student,

M.Sc., Computer Science,

Besant Theosophical College,

Madanapalle.

Mobile:9441008192

e-mail: gaddamrajkumar74@gmail.com

Guide Details

P. VEERA MUTHU

Assistant Professor,

Besant Theosophical College,

Madanapalle.

Mobile:9700740744

e-mail ID:; er.veera86@gmail.com