International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 7 673 – 677

Booster in High Dimensional Data Classification

Ambresh Bhadrashetty
Assistant Professor,
Dept. of Studies in Computer Applications (MCA),
Visvesvaraya Technological University,
Centre for PG studies, Kalaburagi
ambresh.bhadrashetty@gmail.com

Vishalaxi
Student, MCA VI Semester,
Dept. of Studies in Computer Applications (MCA),
Visvesvaraya Technological University,
Centre for PG studies, Kalaburagi
vishalaxi.kaba@gmail.com

Abstract—Classification problems specified in high dimensional data with smallnumber of observation are generally becoming common in specific microarray data. In the time of last two periods of years, manyefficient classification standard models and also Feature Selection (FS) algorithm which isalso referred as FS technique have basically been proposed for higher prediction accuracies. Although, the outcome of FS algorithm related to predicting accuracy is going to be unstable over the variations in considered trainingset, in high dimensional data. In this paperwe present a latest evaluation measure Q-statistic that includes the stability of the selected feature subset in inclusion to prediction accuracy. Then we are going to propose the standard Booster of a FS algorithm that boosts the basic value of the preferred Q-statistic of the algorithm applied. Therefore study on synthetic data and 14 microarray data sets shows that Booster boosts not only the value of Q-statistics but also the prediction accuracy of the algorithm applied.

Keywords-*High dimensional data, Feature selection, Q-statistic, Booster*

*****

I. INTRODUCTION

High dimensional data is being a common factor in many practical applications like data mining, microarray gene expression data analysis and machine learning. The available microarray data is having plenty number of features with small sample size and size of the feature which is to be included in microarray data analysis is growing. It is a tough challenge to consider the statistical classification of data with plenty number of feature and a small sample size.

Many of features in high dimensional microarray data are being irrelevant to the considered target feature. So, to increase the prediction accuracy, finding a relevant features is necessary. The feature should be selected in such a manner that it should not only provide high predictive potential but also a high stability in the selected feature.

Feature Selection

Feature extraction and feature selection are utilized as two primary systems for Dimensionality Reduction. Getting another feature, from the current elements of datasets, is named as feature extraction. Feature Selection is the way toward choosing a subset of features from the whole gathering of accessible features of the dataset. In this manner for feature selection, no preprocessing is required as if there should be an occurrence of feature extraction. Typically the goal of feature selection is to choose a subset of elements for data mining or machine learning applications. Feature selection can be accomplished by utilizing administered and unsupervised strategies. The procedure of Feature selection is constructed mostly with respect to three approaches i.e. filter, wrapper and embedded [6] (Fig. 1).

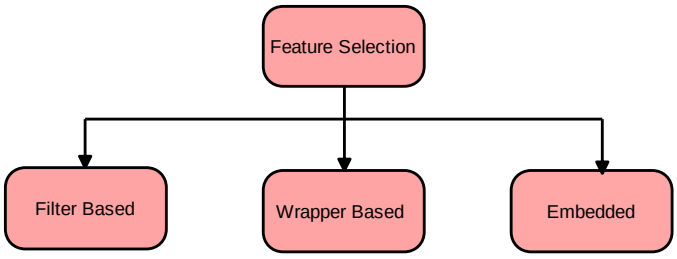

Figure 1: Three approaches of Feature Selection

Filter based feature selection Algorithm: Taking off the features on a few measures (criteria) go under the filter approach of feature selection. In the filter based element determination approach, the decency of a featureis assessed utilizing statistical or inherent properties of the dataset. Considering these properties, a feature is decreed as the most reasonable feature and is selected for machine learning or data mining applications. A portion of the normal methodologies of feature selection are Fast Correlation Based Filter, (FCBF) Correlation based Feature Selection (CFS)

Wrapper based feature selection Algorithms: In the wrapper approach of feature selection, subset of feature is created and decency of subset is assessed utilizing some classifier.

Embedded based feature selection Algorithms*:*In this approach, some classifier is utilized to rank features in the dataset. Based on this rank, a feature is selected for the required application. SVM-RFE is one of the implanted feature selection approaches

A new proposal for feature selection

In this paper we propose Q-statistics to evaluate the performance of an FS algorithm with a classifier. Q-statistic is a hybrid measure to check the prediction accuracy of the

features being selected. Then we also propose a Booster on the selected feature subset from a given FS algorithm.

Booster is introduced to obtain several data sets from the original data set by resampling on sample space. To obtain different feature subsets, FS algorithm is applied on resampled data sets which is given by Booster. Here the Booster boosts not only the value of Q-statistics but also the prediction accuracy of the classifier applied.

II. LITERATURE SURVEY

In 2004, L. Yu, and H. Liu have found that the Feature selection is applied to decrease the quantity of features in number of application programs where data has specific hundreds or generally thousands of features. The feature selection methods put attention on discovering related features. Therefore they have depicted that feature selection alone is not going to be sufficient for effective feature selection of high-dimensional data. So they have describedfeature redundancy and present to perform, examined redundancy analysis and feature selection. Therefore a latest working frame is established that dissociate relevance analysis and redundancy analysis. And then they are going to construct a correlation-basedmethod for relevance and redundancy analysis, and managing a verifiable study basically of its effectiveness comparable with representative methods[1].

In 2005, J. Stefanowski made a study to Ensemble approaches to learning algorithms that develop an arrangement of classifiers and afterward group new instances by consolidating their expectations. These approaches can beat single classifiers on extensive variety of classification problems. It was proposed an expansion of the bagging classifier coordinating it with feature subset selection. More-over, we analyzed the utilization of different methods for incorporating answers of these sub-classifiers, specifically a dynamic voting rather than straightforward voting combination rule. The extended bagging classifier was assessed in an experimental comparative study with standard approaches[2].

In 2005 H. Peng, F. Long, and C. Ding says that Feature selectionis an essential issue for pattern classification systems. We think about how to choose great feature as per the maximal measurable statistical dependency in view of mutual information. Because of the trouble in specifically implementing the maximal dependency condition, we initially determine an equivalent form, called minimal-redundancy-maximal-relevance model (mRMR), for first-order incremental feature selection. At that point, they introduce a two-stage feature selection algorithm by consolidating mRMR and other more refined feature selectors (e.g., wrappers). This enables us to choose a smaller arrangement of predominant features with ease. We perform extensive test comparison of our algorithm and different strategies utilizing three distinct classifiers and four unique data sets. The outcomes affirm that mRMR prompts promising improvement feature selection and classification accuracy[3].

In 2012, A.J. Ferreira, and M.A.T. Figueiredo have faced that Feature selection is a focal issue in machine learning and pattern recognition. On large datasets (as far as dimension as well as number of cases), utilizing seek based or wrapper techniques can be cornputationally restrictive. Additionally, many filter methods in light of relevance/redundancy evaluation likewise take a restrictively long time on high-dimensional datasets. So the author has proposed proficient unsupervised and supervised feature selection/ranking filters for high-dimensional datasets. These techniques utilize low-complexityrelevance a redundancy criteria, material to supervised, semi-supervised, and unsupervised getting the hang of, having the capacity to go about as pre-processors for computationally serious methods to concentrate their consideration on smaller subsets of promising features. The experiment comes about, with up to 10(5) features, demonstrate the time proficiency of our strategies that bring down speculation mistake than best in class methods, while being significantly more straightforward and faster. [4]

In 2014, D. Dernoncourt, B. Hanczar, and J. D. Zucker have worked and found that the Feature selection is a significant step during the construction of a classifier on high dimensional data. Feature selection leads to be unstable because of the small number of observations. The two feature subsets considered from different datasets but having the same classification problem will not overlap extensively. Few works have been done on the selection stability, to find the solution for the stable data. The working of feature selection is analyzed in many different conditions with small sample data. The analysis is done in three steps: The first one is theoretical using simple mathematical model; the second one is empirical and based on artificial data; and the last one is on real data. All three analysis gives the same results and are understanding over the feature selection high dimensional data. [5]

III. PROBLEM DEFINITION

The research on feature selection is still in process since past two decades. Mutual information (MI) is oftenly used to select a relevant features. Forward selection and backward elimination are the two methods used in the statistical variable selection problem. Forward selection method is utilized in many of the successful FS algorithms in high dimensional data. Backward elimination method is not used in practical application because of huge number of features.

A problem with the forward selection method is, change in a decision of the initial feature, which creates a different features subset and varies in the stability. It is known as stability problem in FS.

IV. SYSTEM ARCHITECTURE

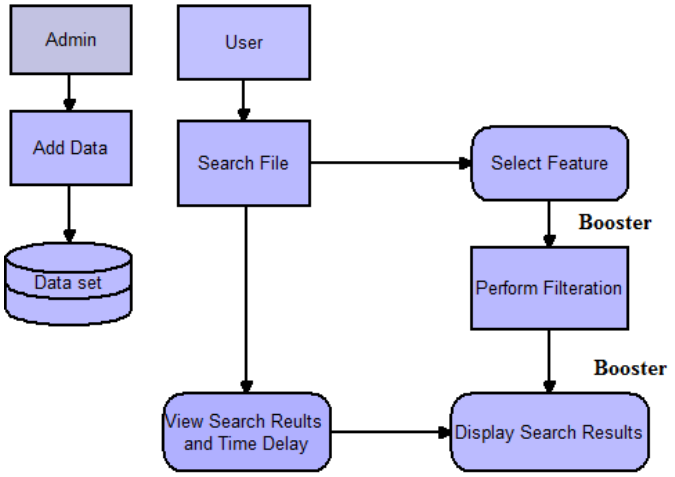

Figure 2: System Architecture

The application is all about a search engine. Here the user can search for the file which is required. Before searching it is necessary that the file should be present in the database. So for that reason admin will add the file into the database which further creates the data set. Once the file is stored in database a user can search for the required file. When the user search for a file multiple operations are performed. It first selects the feature, perform the filtration on the selected feature and then display the search result. Here Booster helps to obtain the results as soon as possible, it works from the time of search till the results are displayed. And a Q-statistic measure shows that how many number of files are present related with the search keyword which we have inserted. Booster also shows the time delay to extract the file. After obtaining the searched file, a user can download the document or images, he can also view the time delay and file count.

V. IMPLEMENTATION

We are implementinga latest feature selection algorithm that specifically mark number of concerns with earlier specified work

$$Let\ D = \{(x_n, y_n)\}\ N=1 \subset RJ \times \{\pm 1\}$$

be as a basic training data set, wherever x_n is mentioned as the n_{th} sample of information constituting of J features, y_n is the equivalent class labeling unit, and also J ≫N. For clearness, we going to examine only binary kind of issues. The specified algorithm is made in general way to label multiple-class issues. Here we first going to interpret the marginal value. Provided a distance function, we going to calculate two nearest neighboring units of every sample unit x_n, usually one from the similar class unit, and also the other one specifically from the distinct class commonly referred as nearest miss or specific NM value. The marginal value of mentioned x_n is next generally calculated as

$$\rho_n = d(x_n, NM(x_n)) - d(x_n, NH(x_n)),$$

where d(·) can be basically a specific distance function. We going to make use of the standard Manhattan useful detachment to interpret a specified model marginal value and considering nearest neighboring units, at the time other standard definition values are going to be utilized. This marginal specification is utilized in implicit way in the familiar RELIEF algorithm, and first specified in mathematical way in basically for feature selection process of the characteristics. An instinctive exposition of this marginal value is a calculation part as to how the quantity of features of x_n is going to be manipulated specifically by noise generally prior being uncategorized. Hence next considering the b marginal theory study, a classifying unit that reduces a marginal-dependent error based operation usually make a general assumption better on not seen basic test information. Next one naturally plan then is going to be scaling every feature, and therefore acquire a weighted feature space related value, characterized by a vectoring unit w which is considered to be a nonnegative value, therefore a marginal-associated error operation in the instigated attribute space related is reduced. Next considering the margin based value of x_n, calculated with regard toward w, basically provided by:-

$$\rho_n(w) = d(x_n, NM(x_n)|w) - d(x_n, NH(x_n)|w) = wTz_n,$$

herez$_n$ is given as $|x_n - NM(x_n)| - |x_n - NH(x_n)|$, and also considering is component-wise absolutely considering operative unit. Point to be noted here that basic $\rho_n(w)$ is specified as a linear operation of component w, and generally contains similar appearance as the model marginal value of standard Support Vector Machines, provided by $\rho\ SVM(x_n) = wT\varphi(x_n)$, by making use of a mapping related operation $\varphi(\cdot)$. A consequential dissimilarity, although, is that by constructing the basic magnitude based value of every element of w in the specific above marginal description returns the relevanceof the comparable feature in a learning process. Hence this is not the situation in standard Support Vector Machines excluding while a particularly linear main most important part unit is utilized, although is going to be catch only basic linear discriminance informative unit. It is to be noted that basically the marginal therefore described needs only informative about the basic neighborhood process of x_n, at the same time no presumption is performed about usually the undisclosed information distributive process. The meaning is that by locally considering educating we are going to transfigure an inconsistent not sequential issue into a group of locality sequential ones. The locally uniformness of a non-sequential issue making it operational is to calculate the featuring based weight values by making use of a sequential based model generally that has been detailed manner made study generally in the specific literature part. Hence it also makes possible the mathematically considering examining of the technique. The major issue with the above specified marginal description, although, usually is that the closest neighboring units of a provided sampling unit are not known before educating. Next in the occurrence of number of thousands of not related characteristics, the closest neighboring units described basically in the real space is going to be entirely distinct from those particularly in the generated space part. To consider for the unsureness in describing locally considered informative data, we going to construct a probable based working model, generally where the nearest neighboring units of a provided sampling units are served as hidden based operating units. Sub-sequent the concepts of the expected-maximizing technique, we going to calculate the marginal value by calculating the presumption of $\rho_n(w)$ through averaging specifically out the not visible variable values.

Input: Information D = {(xn, yn)}N=1⊂RJ× {±1},
part particular width σ, regularizing parameter λ, stop model θ

 ISSN: 2321-8169
673 – 677

Yield: Include weights w
Instatement: Set w(0)= 1, t = 1 ; 1
rehash 2
Figure d(xn, xi|w(t−1)), ∀xn, xi ∈ D; 3
Figure P (xi =NM(xn)|w(t−1)) and 4
P (xj =NH(xn)|w(t−1)) as in (4) and (5);
Give Answer for v by the method for inclination based plummet an incentive by making utilization of a 5 the refreshing guideline by and large indicated in (10);
w(t)j= v2j, 1 ≤j ≤J ;
6t = t + 1; 7
until kw(t)− w(t−1)k < θ ;
8w = w(t).

VI. RESULTS

Figure 3: Upload Document

In the above figure admin will be uploading the file into the database. The file should have a name and a small description related with file which is going to be uploaded. Admin can also upload an image.

Figure 4: View Document

In figure 4 admin can view all the files which uploaded into the database. He can also delete the file which are no more existing.

Figure 5: Search for the file

In this figure a user can search for the required file by entering the search keyword. The search result provides the accurate file and gives the time taken to search a file. It also shows a number of files related with a search keyword as file count. During the search booster reduces the time to extract the file and displays it.

VII. CONCLUSION

We anticipated a measure Q-statistic to assess the performance of a Feature Selection algorithm. Q-statistic the accounts both for the stability of selected feature subset and the prediction accuracy. The paper proposed here is for the Booster to boost the execution of a current FS algorithm. Experimentation with synthetic data and 14 microarray data sets has demonstrated that the recommended Booster enhances the prediction accuracy and the Q-statistic of the three surely understood Feature Selection algorithms: FAST, FCBF, and mRMR. Likewise we take in notice that the classification methods applied to Booster do not have much effect on prediction accuracy and Q-statistics.

The Performance of mRMR-Booster be appeared near be extraordinary together in the prediction accuracy along with Q-statistic This was examined with the intention of if a FS algorithm is proficient however couldn't get superior in the high performance in accuracy or the Q-statistics for some particular data, Booster of the Feature Selection algorithm will boost the performance. If a FS algorithm itself is not productive, Booster will most likely be unable to acquire high performance. The execution of Booster relies upon the performanceof FS algorithm applied

REFERENCES

[1] L. Yu, and H. Liu,"Efficient feature selection via analysis of relevance and redundancy," The Journal of Machine Learning Research, vol. 5, no.2, pp. 1205-1224, 2004.

[2] J. Stefanowski,"An experimental study of methods combining multiple classifiers-diversified both by feature selection and bootstrap sampling," Issues in the Representation and Processing of Uncertain and Imprecise Information, pp. 337-354, 2005.

[3] H. Peng, F. Long, and C. Ding, "Feature selection based onmutual information criteria of max dependency, max-

 ISSN: 2321-8169
673 – 677

relevance,and min-redundancy,"IEEE transactions on pattern analysis andmachine intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.

[4] A.J. Ferreira, and M.A.T. Figueiredo,"Efficient feature selection filters for high dimensional data," Pattern recognition letters, vol. 33, no. 13, pp. 1794-1804, 2012.

[5] D. Dernoncourt, B. Hanczar, and J.D. Zucker,"Analysis of feature selection stability on high dimension and small sample data," Computational Statistics and Data Analysis, vol. 71, pp. 681-693, 2014.

[6] Amit Kumar Saxena and Vimal Kumar Dubey"A Survey on Feature Selection Algorithms" International Journal on Recent and Innovation Trends in Computing and Communication, vol 3

[7] S. Alelyan,"On Feature Selection Stability: A Data Perspective," PhD dissertation, Arizona State University, 2013.

[8] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, and W. Pedrycz, "Measuring relevance between discrete and continuous features based on neighborhood mutual information," Expert Systems with Applications, vol. 38, no. 9, pp. 10737-10750, 2011.

[9] B. Wu et al.,"Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," Bioinformatics, vol. 19, no. 13, pp. 1636-1643, 2003.

[10] N. Meinshausen, and P. Buhlmann,"Stability selection," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 72, no. 4, pp. 417-473, 2010.

[11] L.I. Kuncheva,"A stability index for feature selection," Artificial Intelligence and Applications, pp. 421-427, 2007.

[12] R.V. Jorge, and A.E. Pablo,"A review of feature selection methods based on mutual information," Neural Computing and Applications, vol. 24, no. 1, pp. 175-186, 2014.

[13] G. Gulgezen, Z. Cataltepe, and L. Yu,"Stable and accurate feature selection," In: Machine Learning and Knowledge Discovery in Databases, pp. 455-468, 2009.

[14] F. Alonso-Atienza, and J.L. Rojo-Alvare, et al., "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," Expert Systems with Applications, vol. 39, no.2, pp. 1956-1967, 2012.