A Survey on Boosting High dimensional Feature Selection Classification

[1]E.Kokilamani, [2]Dr. R.Gunavathi

[1]M.Phil. Research Scholar, [2]HOD
Sree Saraswathi Thiyagaraja College, Pollachi

Abstract: **Classification problems in high dimensional data with small number of observations are becoming more common particularly in microarray data. Throughout the last two decades, plenty of efficient categorization models and feature selection (FS) algorithms have been planned for high prediction accuracies. The *optimal Linear Programming Boosting (LPBoost)* is a supervise classifier since the boosting family of classifiers. To predict or the feature selection (FS) algorithm applied is not efficient with the accurate data set. The LP Boost maximizes a margin between training samples of dissimilar classes and therefore also belongs to the class of margin-maximizing supervised classification algorithms. Therefore, Booster can also be used as a criterion to estimate the act of an FS algorithm or to estimate the complexity of a data set for classification. LPBoost iteratively optimizes double misclassification costs and vigorously generates pathetic hypotheses to build new LP columns.**

***KEYWORDS*: Optimal Linear Programming Boosting, Prediction, Estimate, Misclassification, Feature Selection**

I. INTRODUCTION

Recent classification techniques achieve well when the number of training examples exceeds the number of features. If, however, the number of features greatly exceeds the number of training examples, then these same techniques can fail. Classification is a supervised learning technique. It arises often from bioinformatics such as disease classifications using high throughput data like micorarrays or SNPs and machine learning such as document classification and image recognition. It tries to find out a function from training data consisting of pairs of input features and categorical output. This function will be used to forecast a class label of any valid input feature. Well known classification methods include (multiple) logistic regression, Fisher discriminant analysis, k-th-nearest-neighbor classifier, support vector machines, and many others.

High-dimensional discriminant analysis plays an important role in multivariate statistics and machine learning. feature selection in the scope of classification problems, clearing up the foundations, real application problems and the challenges of feature selection in the context of high-dimensional data. we focus on the basis of feature selection, providing a review of its history and basic concepts. Then, we address different topics in which feature selection plays a crucial role, such as microarray data, intrusion detection, or medical applications.

There are many interesting domains that have high dimensionality. Some examples include the stream of images produced from a video camera, the output of a sensor network with many nodes, or the time series of functional magnetic resonance images (fMRI) of the brain. Often we want use this high dimensional data as part of a classification task. For instance, we may want our sensor network to classify intruders from authorized personnel, or we may want to analyze a series of fMR images to determine the cognitive state of a human subject. High dimensionality poses significant statistical challenges and renders many traditional classification algorithms impractical to use. In this chapter, we present a comprehensive overview of different classifiers that have been highly successful in handling high dimensional data classification problems. We start with popular methods such as Support Vector Machines and variants of discriminate functions and discuss in detail their applications and modifications to several problems in high dimensional settings. Scalable and efficient classification models with good generalization ability along with model interpretability for high dimensional data problems.

II.LITERATURE SURVEY

Abeel T et al[1] discussed biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high-dimensional data. Surprisingly, the stability with respect to sampling variation or robustness of such selection processes has received attention only recently. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. In addition, a more robust set of markers may strengthen the confidence of an expert in the results of a selection method.

F. Alonso-Atienza et al[2] address the Early detection of ventricular fibrillation (VF) is crucial for the success of the defibrillation therapy in automatic devices. A high number of detectors have been proposed based on temporal, spectral, and time–frequency parameters extracted from the surface electrocardiogram (ECG), showing always a limited performance. The combination ECG parameters on different domain (time, frequency, and time–frequency) using machine learning algorithms has been used to improve detection efficiency. In this study, we propose a novel FS algorithm based on support vector machines

(SVM) classifiers and bootstrap resampling (BR) techniques. We define a backward FS procedure that relies on evaluating changes in SVM performance when removing features from the input space.

David Dernoncourt et al[3] have proposed Abstract Feature selection is an important step when building a classifier on high dimensional data. As the number of observations is small, the feature selection tends to be unstable. It is common that two feature subsets, obtained from different datasets but dealing with the same classification problem, do not overlap significantly. Although it is a crucial problem, few works have been done on the selection stability. The behavior of feature selection is analyzed in various conditions, not exclusively but with a focus on t-score based feature selection approaches and small sample data.

Gordon GJ et al[4] have presented an pathological distinction between malignant pleural mesothelioma (MPM)and adenocarcinoma (ADCA) of the lung can be cumbersome using established methods. We propose that a simple technique, based on the expression levels of a small number of genes, can be useful in the early and accurate diagnosis of MPM and lung cancer. This method is designed to accurately distinguish between genetically disparate tissues using gene expression ratios and rationally chosen thresholds. Here we have tested the fidelity of ratio-based diagnosis in differentiating between MPM and lung cancer in 181 tissue samples (31 MPM and 150 ADCA). We then examined (in the test set) the accuracy of multiple ratios combined to form a simple diagnostic tool. We propose that using gene expression ratios is an accurate and inexpensive technique with direct clinical applicability for distinguishing between MPM and lung cancer.

Guyon et al[5] address the variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data.

A.I. Su et al[6] discussed high-throughput gene expression profiling has become an important tool for investigating transcriptional activity in a variety of biological samples. To date, the vast majority of these experiments have focused on specific biological processes and perturbations. Here, we have generated and analyzed gene expression from a set of samples spanning a broad range of biological conditions. Specifically, we profiled gene expression from 91 human and mouse samples across a diverse array of tissues, organs, and cell lines. We have used this dataset to illustrate methods of mining these data, and to reveal insights into molecular and physiological gene function, mechanisms of transcriptional regulation, disease etiology, and comparative genomics.

D. Dembele et al[7] addresses Microarray technology allows monitoring of gene expression profiling at the genome level. This is useful in order to search for genes involved in a disease. The performances of the methods used to select interesting genes are most often judged after other analyzes (qPCR validation, search in databases...), which are also subject to error. A good evaluation of gene selection methods is possible with data whose characteristics are known, that is to say, synthetic data. We propose a model to simulate microarray data with similar characteristics to the data commonly produced by current platforms. The parameters used in this model are described to allow the user to generate data with varying characteristics. In order to show the flexibility of the proposed model, a commented example is given and illustrated.

Somol P et al [8] discussed Stability (robustness) of feature selection methods is a topic of recent interest, yet often neglected importance, with direct impact on the reliability of machine learning systems. We investigate the problem of evaluating the stability of feature selection processes yielding subsets of varying size. We introduce several novel feature selection stability measures and adjust some existing measures in a unifying framework that offers broad insight into the stability problem. We study in detail the properties of considered measures and demonstrate on various examples what information about the feature selection process can be gained. We also introduce an alternative approach to feature selection evaluation in the form of measures that enable comparing the similarity of two feature selection processes. These measures enable comparing, e.g., the output of two feature selection methods or two runs of one method with different parameters.

Singhal S et al[9] The development of microarray technology has allowed researchers to measure expression levels of thousands of genes simultaneously. To approach this problem we have developed a publicly available simulator of microarray hybridization experiments that can be used to help assess the accuracy of bioinformatic tools in discovering significant genes. After analyzing microarray hybridization experiments from over 50 samples, an estimate of various degrees of technical and biological variability was obtained. We found that the type of normalization approach used was an important aspect of data analysis. Global normalization was the least accurate approach. We provide access to the microarray hybridization simulator as a public resource for biologists to further test new emerging genomic bioinfomatic tools.

Hugo Silva et al[10] In feature selection (FS), different strategies usually lead to different results. Even the same strategy may do so in distinct feature selection contexts. We propose a feature subspace ensemble method, consisting on the parallel combination of decisions from multiple classifiers. Each classifier is designed using variations of the feature representation space, obtained by means of FS. With the proposed approach, relevant discriminative information contained in features neglected in a single run of a FS method, may be recovered by the application of multiple FS runs or algorithms, and contribute to the decision through the classifier combination process. Experimental results on benchmark data show that the proposed feature subspace ensembles method consistently leads to improved classification performance.

III.CONCULSION

The optimal Linear Programming Boosting (LPBoost) is effective algorithm to handle a difficult data set for classification. The Number of feature can easily compared and performance of LP boost will provide a effective results. Linear Boost survey domain is proposed in this literature survey. Finding relevant features directly from a huge number of the features with continuous values using the definition of relevancy is very easy task.

REFERENCES

[1] 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
[2] 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.
[3] 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
[4] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock,S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, and R.Bueno, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer And Mesothelioma," Cancer Research, vol. 62, no. 17, pp.4963-4967, 2002.
[5] I. Guyon, and A. Elisseeff, "An Introduction to Variable and Feature Selection," The Journal of Machine Learning Research, vol.3, pp. 1157-1182, 2003.
[6] A.I. Su, M.P. Cooke, and K.A. Ching, et al., "Large-scale analysis of the human and mouse transcriptomes," Proceedings of the National Academy of Sciences of the United States of America,vol. 99, no. 7, pp. 4465-4470, 2002.
[7] .D. Dembele, "A Flexible Microarray Data Simulataion Model,"Microarrays, vol. 2, no. 2, pp. 115-130, 2013.
[8] P. Somol, and J. Novovicova, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 11, pp. 1921-1939, 2010.
[9] S. Singhal, et al., "Microarray data simulator for improved selection of differentially expressed genes," Cancer Biology and therapy, vol. 2, no. 4, pp. 383-391, 2003.
[10] H. Silva, and A. Fred, "Feature subspace ensembles: a parallel classifier combination scheme using feature selection," Multiple classifier systems, vol. 4472, pp. 261-270, 2007.