ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Integration of Feature Generation Scheme and Boosters for Micro Array Data Classification

[1]P. Deepa Final Year-IT,[2]P. Soundarya Final Year -IT,[3]S. SanthiPriya, AP/IT
Department Of Information Technology,
Mahendra Engineering College for Women. Kumaramangalam, Tiruchengode, Tamilnadu, India
Mail ID: [1]**deepa66@gmail.com**, [2]**soundaryaprakash01@gmail.com**

Abstract

The large scale data analysis operations are carried out with the support of the data mining or machine learning methods. Dimensionality is the key issue in the data mining and machine learning applications. The high dimensional data analysis requires huge computational resources and processing time. Dimensionality reduction methods are applied for better visualization, data compression, noise removal and understandability and generalization factors. The data size is controlled with the dimensionality reduction tasks.

The feature selection models are applied to reduce the dimensionality in the high dimensional data environment. The subset selection with relevancy factor is considered in the future selection process. Statistical methods are applied in the feature selection process. The weakly relevant features are discovered using the T-test model. The irrelevant features are removed with F-test model. The Q-statistics measures are applied to evaluate the features. The booster algorithm is applied for the feature improvement process. Naïve Bayes algorithm is used for the classification process.

The static features are discovered using the feature selection methods. The feature extraction methods are applied to fetch dynamic features in the micro array data values. The compound feature generation mechanism is applied to combine feature selection and extraction tasks. The feature integration operations are carried out with multiple ratio based feature relationships. Boosting method is integrated with the compound feature generation scheme. The classification process is carried out using the Naïve bayes algorithm with generated feature values.

Index Terms: High dimensional data classification, Feature selection, Feature extraction, feature generation and Naïve Bayesian classifier

1. Introduction

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches.

The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categorie. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the

generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to overfit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper.

With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. Pereira et al., Baker et al. and Dhillon et al. employed the distributional clustering of words to reduce the dimensionality of text data.

2. Related Work

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because: (i) irrelevant features do not contribute to the predictive accuracy and (ii) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed FAST algorithm falls into the second group.

Traditionally, feature subset selection research has focused on searching for relevant features. A well known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function [4]. Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi-class problems, but still cannot identify redundant features.

Along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated as well. CFS, FCBF and CMIM are examples that take into consideration the redundant features. CFS is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. FCBF is a fast filter method can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. CMIM iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from these algorithms, our proposed FAST algorithm employs clustering based method to choose features.

Recently, hierarchical clustering has been adopted in word selection in the context of text classification. Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word by Baker and McCallum. As distributional clustering of words are agglomerative in nature, and result in sub-optimal word clusters and high computational cost, Dhillon et al. proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. Butterworth et al. proposed to cluster features using a special metric of Barthelemy-Montjardet distance, and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on Barthelemy-Montjardet distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

Hierarchical clustering also has been used to select features on spectral data. Van Dijk and Van Hullefor [11] proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. [8] presented a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information. Their feature clustering method is similar to that of Van Dijk and Van Hullefor except that the former forces every cluster to contain consecutive features only. Both methods employed agglomerative hierarchical clustering to remove redundant features.

Quite different from these hierarchical clustering based algorithms, our proposed FAST algorithm uses minimum spanning tree based method to cluster features. Meanwhile, it does not assume that data points are grouped around centers or separated by a regular geometric curve. Moreover, our proposed FAST does not limit to some specific types of data.

3. Feature Selection and Booster for Classification

The presence of high dimensional data is becoming more common in many practical applications such as data mining, machine learning and microarray gene expression data analysis. Typical publicly available microarray data has tens of thousands of features with small sample size and the size of the features considered in microarray data analysis is growing. The statistical classification of the data with huge number of features and small sample size presents an intrinsic challenge. A striking result has been found that the simple and popular Fisher linear discriminate analysis can be as poor as random guessing as the number of features gets larger.

As was reported in [14], most of the features of high dimensional microarray data are irrelevant to the target feature and the proportion of relevant features or the percentage of up-regulated or down-regulated genes compared with appropriate normal tissues is only 2% □5%. Finding relevant features simplifies learning process and increases prediction accuracy. The finding, however, should be relatively robust to the variations in training data, especially in biomedical study, since domain experts will invest considerable time and efforts on this small set of selected features. Hence, the proposed selection should provide them not only with the high predictive potential but also with the high stability in the selection.

There have been lots of researches on the FS during the last two decades, and the research continues to be still one of the hot topics in machine learning area [1], [9], [12]. 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.

Methods used in the problems of statistical variable selection such as forward selection, backward elimination and their combination can be used for FS problems. 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 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. This is known as the stability problem in FS. The research in this area is relatively a new field [3], [5], [2], [10], [7], and devising an efficient method to obtain a more stable feature subset with high accuracy is a challenging area of research.

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.

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. Several studies based on resampling technique have been done to generate different data sets for classification problem [13] and some of the studies utilize resampling on the feature space [6]. The purposes of all these studies are on the prediction accuracy of classification without consideration on the stability of the selected feature subset.

4. Problem Statement

The weakly relevant features are discovered using the T-test model. The irrelevant features are removed with F-test model. The Q-statistics measures are applied to evaluate the features. The booster algorithm is applied for the feature improvement process. Naïve Bayes algorithm is used for the classification process. The following problems are identified from the existing system. Feature extraction operations are not supported in the system. Heterogeneous ratio based feature integration operations are not supported. Low dimensionality reduction levels and Limited classification accuracy levels.

5. Feature Generation Scheme and Boosters for High Dimensional Data Classification

The high dimensional data classification is performed on features. The features are analyzed with statistical measures. The features are improved with boosting process. The system is divided into six major modules. They are Data Preprocess, Feature Selection, Compound Feature Generation, Feature Analysis, Boosting Process and Classification Process.

The data preprocess module is used to perform data cleaning operations. Dimensionality reduction operations are carried out under the feature selection process. The feature selection and feature extraction operations are combined in compound feature generation module. Feature quality is evaluated in the feature analysis module. The boosting process is build to improve the features. The gene data are categorized under data classification module.

5.1. Data Preprocess

The gene micro array data values are collected as textual data files. The data populate process is used to transfer the textual data into Oracle database. The noisy and redundant data values are corrected under the data cleaning process. The missing values are updated using the aggregation data substitution method.

5.2. Feature Selection

The feature selection operations are carried out with data relevancy identification process. The T-test model is applied to discover weakly relevant features. The irrelevant features are identified using the F-test model. The relevancy and redundancy factors are analyzed in the feature selection process.

5.3. Compound Feature Generation

The original features are identified using the feature selection process. The transformed features are discovered using the feature extraction process. The feature selection and feature extraction tasks are integrated in the compound feature generation process. The integrated features are discovered using the Compound Feature Set Generation (CFG) algorithm.

5.4. Feature Analysis

The optimal features are identified using the feature analysis model. The q-statistics is used for the feature analysis process. The q-statistics measure is calculated for the discovered features. The selected features are passed to the boosting process. The feature selection process is carried out with three algorithms. The Minimal-redundancy-maximal-relevance (mRMR) algorithm considers the redundancy and relevancy factors in the feature selection process. The Fast Correlation-Based Filter (FCBF) algorithm uses the correlation factors. The data partitioning is applied in the Fast clustering bAsed feature Selection algoriThm (FAST).

5.5. Boosting Process

The boosting process is applied to improve the selected feature. The q-statistics measure is improved in the boosting process. The booster algorithm is applied in the feature improvement process. The resampling process is applied on the sample space.

5.6. Classification Process

The cancer severity level is identified in the classification process. The Naive Bayes (NB) classifier is used in the gene analysis process. The learning process is applied to identify the class patterns. The testing process is used to assign the class values for the gene data.

6. Conclusion

The feature selection methods are applied for the dimensionality reduction process. The classification operations are carried out on the selected features. The Q-statistics and Naïve bayes algorithms are used for the classification and feature analysis process. The compound feature generation scheme is applied to improve the feature selection process. The cancer gene classification process is performed on features. The feature selection and extraction methods are integrated in the system. Static and dynamic features are integrated in the feature generation process. The system achieves high accuracy levels with minimum computational overhead.

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] Y. Han and L. Yu, "A variance reduction framework for stable feature selection," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 428–445, 2012.

[3] S. Alelyan, "On feature selection stability: A data perspective," PhD dissertation, Arizona State Univ., Tempe, AZ, USA, 2013.

[4] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transactions On Knowledge And Data Engineering Vol:25 No:1 Year 2013.

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

[6] F. Alonso-Atienza, J. L. Rojo-Alvare, A. Rosado-Mu~noz, J. J. Vinagre, A. Garcia-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," Expert Syst. Appl., vol. 39, no. 2, pp. 1956–1967, 2012.

[7] N.Meinshausen and P. Buhlmann, "Stability selection," J. Roy. Statist. Soc.: Series B (Statist.Methodol.), vol. 72, no. 4, pp. 417–473, 2010.

[8] Krier C., Francois D., Rossi F. and Verleysen M., Feature clustering and mutual information for the selection of variables in spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162, 2007.

[9] A. J. Ferreira and M. A. T. Figueiredo, "Efficient feature selection filters for high dimensional data," Pattern Recog. Lett., vol. 33, no. 13, pp. 1794–1804, 2012.

[10] Z. He and W. Yu, "Stable feature selection for biomarker discovery," Comput. Biol. Chem., vol. 34, no. 4, pp. 215–225, 2010.

[11] Van Dijk G. and Van Hulle M.M., Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, International Conference on Artificial Neural Networks, 2006.

[12] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 1–14, Jan. 2013.

[13] K. M. Ting, J. R. Wells, S. C. Tan, S. W. Teng, and G. I. Webb, "Feature-subspace aggregating: Ensembles for stable and unstable learners," Mach. Learn., vol. 82, no. 3, pp. 375–397, 2011.

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