A Survey on Efficient Selection of Subset of Feature Technique on HDSS Data

Apurva Chaudhari\textsuperscript{1} Prof. Satish S. Banait\textsuperscript{2}
\textsuperscript{1,2}Department of Computer Engineering
\textsuperscript{1}PG Student \textsuperscript{2}Assistant Professor

KKWIEER, Nashik, (Savitribai Phule Pune University, Pune) India

Abstract — Feature subset selection has the main attention of the research in the areas for which datasets possess high dimensional variables. During Classification, the high dimensional feature vectors of microarray data impose a high dimensional cost and the risk of over fitting. Hence there is a necessity to reduce the dimension with the help of feature selection. This survey paper considers Feature subset selection on classification for biomedical datasets with a less samples and large features or variables. Commonly, the performance of a classifier is degraded due to irrelevant features of high dimensional data. A conventional form of regularization gives major class an equivalent or more emphasis, but here main focus is on Minority class so that overall Classifier’s performance can be improved.

Key words: Feature Subset Selection, Linear Discriminant Analysis, High Dimensional Small Sized data

I. INTRODUCTION

Recently, Feature subset selection process is a challenging scope on High dimensional small sized data \cite{1} having scope from research and investigation point of view. The data with high dimension and having small size problems consists of considerably more Features than instances, so before constructing a classification model, Feature selection generally helps as a crucial phase owed due to the Curse of dimensionality \cite{1}. For the purpose of usage of machine learning methods efficiently, data pre-processing is very crucial for which Feature selection is measured as most promising and important practices for data pre-processing. Feature selection or subset selection involves selecting the Feature subset that exploits the accuracy of classification. Good Feature subset comprises of the minimum number of features that exploits the classification accuracy. Selection of features possesses two main approaches, namely, Individual Evaluation and Subset Evaluation. Here, we are only dealing with Subset Evaluation/Selection. Feature selection proves to be the utmost way for reducing the dimensionality of High dimensional data. In Subset Evaluation, candidate feature subsets are built using search strategy. In machine learning such kind of subset evaluation / selection has some problems like class imbalance \cite{1,2,3,4,5,6}. Class imbalance problem means one class has more number of samples than in case of other class. Due to class imbalance problem, the classifier is mostly affected for a specific data set, which could lead overall accuracy but very low performance on the Minority class \cite{6}. Knowing what the class imbalance problem is, many approaches are present to handle with this problem. These approaches are classified into 2 major categories: Sampling based approaches & Feature selection approaches. Sampling based approaches are further classified, namely, under sampling, oversampling and hybrid sampling approaches. While selecting the features from reference dataset there are few problems like a large number

of samples with few feature sets, and a large feature set with very small samples \cite{7,8,10,11}.

II. FEATURE SUBSET SELECTION

Dimensionality reduction of the feature vectors which define each object offers numerous benefits. In Feature selection method, High dimensional data comprises many redundant and irrelevant features \cite{1,5,17}. Redundant or irrelevant features may disturb the accuracy of Classification algorithms in negative way. Moreover, reducing the amount of features may help reduction of the cost of obtaining data and might make it informal to understand the classification models. There are several methods for Dimensionality reduction studied in Literature. Specific mutual methods follow transformations of the original features to lower dimensional spaces. Consider Linear Discriminant Analysis that reduces the dimensions of the data \cite{1}. Additionally, the Linear combination of features makes it tough in interpreting the consequence of the original features on Class discrimination. For all above reasons; main focus should be on methods that select original feature subsets.

III. LITERATURE SURVEY

Recently, many researchers have worked on improving the robustness and accuracy of classifier using Feature subset selection algorithm on data with high dimension and having small size. Some methods considered majority class while others considered minority class. Each has its own merits and demerits. Some of them are briefed as follows:

Feng Yang et al. \cite{1} proposed a method which focuses on minority class and is Feature selection method which is based on LDA termed as MCE-LDA with regularization technique. This method tried to overcome the difficulties occurred by LDA, when feature selection is applied on data with high dimension and having small size with class imbalance. A conventional form of regularization focuses on mainly on majority class, but this paper focuses on minority class with new regularization technique. It has showed that MCE-LDA \cite{1} produces feature subsets which have admirable performance in both robustness and classification.

Xinjian Guo et al. \cite{3} investigated various remedies for class imbalance problem at four different levels. Problem called Class imbalance is not actually caused by Imbalance occurred in class, but also due to imbalanced class may produce small samples which will cause in degradation. One direct way to handle with this Class imbalance problem is in the modification class distributions. Distributions can be balanced and can be achieved by removing some instances from the majority class and adding some instances in the minority class \cite{5}. SMOTE is one of the over-sampling methods. SMOTE produces synthetic minority samples by oversampling the minority class. Another iterative boosting
algorithm places the diverse weights on the training distributions. This focuses more on the imperfectly categorized samples in the subsequent iteration. Boosting efficiently modifies the distribution of the training data, which can be measured as an advanced sampling technique [3].

K. Javed et al. [5] proposed a Feature ranking algorithm called as Class Dependent density based Feature Elimination for binary data sets [5]. Authors focused on the supervised feature problem and theoretical analysis had showed that CDFE computes the weights more powerfully in comparisons of mutual information measure which is used for feature filter-wrapper approach for selecting the final feature subset. With High dimensional datasets, Feature selection with Feature ranking algorithms is modest and has effective computations but redundant data cannot be easily removed. Consequently, Feature subset selection algorithm evaluates data redundancies but it becomes unrealistic for computations on High dimensional datasets. By combining Feature ranking as well as Feature subset selection methods in the form of a two stage algorithm can be addressed for above problems.

JIANG Zhu, Zhao Fei [10] proposed a unique approach based on multi-criterion fusion for improving the robustness and accuracy of Feature selection algorithm. It is seen that the unselected features may consists of useful information if not selected. This may reduce the performance of feature selection. So the fusion technique is used to exploit the suitable information in the neglected features. After selecting the features from samples, the results of feature selection are used to train the classifier called Polynomial-SVM. For selecting the features, the selection criterions of Fisher Ratio, ReliefF and polynomial support vector machine (PSVM) are considered. Comparing MCF-PSVM with other three methods, it is found that the accuracy of feature identification for MCF-PSVM method was better than ReliefF and PSVM but less than Fisher Ratio.

F. Yang, K. Z. Mao [13] suggested method to increase the robustness of Feature Selection with multiple criteria fusion for feature evaluation. Multiple criteria are used for the Feature evaluation and it inclines to be less sensitive to the incorrect valuation, and hereafter, the toughness of the Feature selection algorithm is enhanced. Main goal to enhance the Feature selection consequences in terms of both Classification stability and performance.

Cawley, Talbot [14] proposed the substantial improvement of sparse logistic regression (SLogReg) approach. Authors showed that a modest Bayesian method can be occupied to remove the Regularization constraint completely by incorporating it systematically by using “an uninformative Jeffreys’s prior”. This enhanced Algorithm is called as Bayesian logistic regularization (BLogReg) is usually of two or three orders of degree closer than the sparse logistic regression algorithm. BLogReg algorithm unrestricted from Selection bias in case of performance estimation.

H. Peng et al. [15] considered the selection of worthy Features for Maximal statistical dependent criteria created on Mutual information. Previously, there was trouble in directly employing the Maximal dependency condition, consequently for first order incremental feature selection, corresponding method called as minimal redundancy maximal relevance criterion is proposed. Then, it presents a two stage algorithm for selecting features by relating mRMR [15] with additional classifier Feature selectors. At very low cost, this allowed in selecting a compressed set of more Features. It has been observed that mRMR makes improvement in feature selection as well as in the classification accuracy.

M. Robnik Sikonja and I. Kononenko [16] studied Features, parameters required and the uses of RELIEF family of algorithms. RELIEF, RELIEF-F and RRELIEF together form a RELIEF family. RELIEF algorithms are common and popular estimators of attribute. Conditional dependencies are good identified by these algorithms. They offer an integrated view on attribute estimation in classification and regression. Basic Relief algorithm is limited to two class classification problem. ReliefF overcomes the disadvantage of basic Relief algorithm and can deal with multiclass problem. RRelief is the basic Relief with regression. In brief, Relief algorithms provide robustness and are noise tolerant. These algorithms have a larger computational complexity.

I. Guyon et al. [17] addressed the small subset of selection of genes problem from wide forms of Gene expression data. Previously, a classifier is built using present training samples from cancer and normal patients, which was appropriate for genetic diagnosis and discovery of drugs. For overcoming the genes selection problem with correlation methods, a new method is proposed for selecting the genes using Support vector machine approaches which are based on Recursive Feature Elimination. It was noticed that genes which were selected by proposed technique produce a better classification performance. SVM-RFE process reduces gene redundancy repeatedly and produces better and compacted gene subsets.

IV. SUPPORT VECTOR MACHINE

Support Vector Machine is a search for optimum best hyperplane as a decision function in a High dimensional space. Usually, linear Support Vector Machine has somewhat poor performance in contrast, but when it is used only with slightly skewed datasets, linear SVM accomplishes best performance [12]. Very small research work is carried out on learning from small samples, but there exist a number of problems would benefit greatly from research by doing this task. Biological data analysis [6], [14], [16], [17] problems often have very small sample sizes but large feature sets. Support Vector Machine is a technique in attaining the optimum boundary that is most distant from the vectors nearest to the boundary in both of the sets.

V. CONCLUSION

Survey carried out and objective of proposed system suggest that the approach for feature selection and classification for high dimensional small sized dataset should be smart enough so that class imbalance, over-fitting and overwhelming problems should be avoided. In conventional form, Majority class had equal or more emphasis, but there may be possibility that minority class may improve overall performance of Feature selection algorithm on the data possessing high dimensional and with small size [1]. Also in future, we can consider instance ratio of minority and majority classes for improving the robustness and accuracy.
REFERENCES


