An Efficient Distributed Feature Subset Selection Technique on High Dimensional Small Sized Data

Apurva Y. Chaudhari
M.E. Computer (Student)
Department of Computer Engineering
K. K. Wagh Institute of Engineering Education & Research, Nashik
S. P. P. U

Satish. S. Banait
Assistant Professor
Department of Computer Engineering
K. K. Wagh Institute of Engineering Education & Research, Nashik
S. P. P. U

ABSTRACT
Feature subset selection is a crucial phase in modeling accurate classifiers in data mining and machine learning, especially with High Dimensional Small Sized (HDSS) data. LDA can also be used for feature selection as an efficient measure for evaluation of the feature subset. While LDA is applied to feature selection on HDSS data and class imbalance, it meets some difficulties, such as singular scatter matrix, overwhelming, overfitting, and computational complexity. For this purpose, a new LDA based feature selection technique based is proposed which focuses more on minority class with a novel regularization technique. Main objective is to enhance the performance of feature subset selection process using LDA in distributed environment. Sample ratio between both classes has been determined.

Keywords
Feature subset selection, Class emphasis, HDSS, Classification, Regularization

1. INTRODUCTION
Recently, Feature Subset Selection Process is a challenging on High dimensional small sized data [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 Curse of dimensionality [1] [4] [11]. For learning methods to work efficiently, data preprocessing is very crucial for which Feature selection is measured as most promising and important practices for data preprocessing. 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 methods are used to satisfy some common objectives such as eliminating redundant and irrelevant data and improving result directly, increasing the accuracy of classifier, decreasing dimensionality, and to help to avoid slow execution time of learning algorithms. The class imbalance problem is a problematic challenge which many times occur in HDSS data. It means that one class has more number of samples than in case of other class.The class having majority of samples data is referred to as majority 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 [8]. Traditionally, feature selection algorithms always considered majority class for classification purpose, in such case, minority class was neglected. But there is a possibility that the minority class can improve the classification performance as well.So instead of considering only majority class in case of class imbalance, minority class should be given more importance.

Mostly for pattern classification, LDA is used as a classifier. Though, it can also be used as an efficient measure for feature selection [1]. As a substitute for an evaluation purpose of the classification error of LDA classifier, feature selection based on LDA can be used. This will be built on ratio of between-class scatter and within-class scatter. Projection of LDA contains inverse process of the scatter matrix. Also, scatter matrix is singular for HDSS data. For this purpose, regularization technique[1] [13] is the best practice. Regularization practice consists of class and diagonal emphasis. Overfitting [4] is another problem caused. LDA overfits the training data due to small size of samples, and selected feature subsets shows poorer on testing data as compared to training data. This can also be overcome using regularization technique.

Overwhelming [4] problem occurs due to imbalance of class, majority class overwhelms minority class. It inclines to control the process of feature selection, which in advance worsen overfitting problem in minority class. Classification algorithm consists of a set of samples, where every sample is defined by a fixed number of features along with a class label. Linear SVM [12] is used as a classifier due to its remarkable performance on gene microarray data. The performance of classification can deteriorate if it is directly applied on datasets which are of high dimension, imbalanced or of small sample data. Hence we are proposing here more efficient and robust method based on LDA which focuses on minority class which will increase the performance of classification. Furthermore, based on the minority class emphasis the sample ratio of majority and minority is taken into consideration.

2. LITERATURE SURVEY
Recently, many researchers have worked on improving the performance of classifier using feature subset selection algorithm on data with high dimension and having small size. Some of them are briefed as follows:

Jiang Zhu and Zhao Fei [2] proposed a unique approach based on multi-criterion fusion for improving the accuracy and robustness of feature selection algorithm. The unselected features may consist of useful information if not selected reduces the performance of feature selection. So the fusion method is used to exploit the suitable information in the neglected features. For selecting the features, the selection criterions of Fisher Ratio, ReliefF [3] and polynomial support vector machine (PSVM) are considered.

Relief is common feature ranking based method which considers the dependencies between features when features are evaluated [3]. Relief Family consists of Relief, ReliefF and RRelief. Basic Relief algorithm is limited to two-class classification problem. In brief, Relief algorithms provide robustness and are noise tolerant. These algorithms have a larger computational complexity. Also, Relief algorithm voraciously attempts to reduce Bayes Error assessed by the kNN estimator.

A feature subset selection algorithm which is dependent on classifier was proposed called SVMRFE [5]. It has been usually used as a feature selection algorithm because of its worthy classification performance. To overcome the previous problem of selecting genes with correlation techniques, Y.tang [5] proposed a method of selecting genes using SVM methods based on Recursive Feature Elimination. It has been seen that genes which are selected by SVM-RFE gives a better classification performance. SVM-RFE technique automatically eliminated gene redundancy and produces improved and compacted gene subsets.

H. Peng et al. [6] 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 mRMR criterion is proposed. Then, it presented a two stage algorithm for selecting features by relating mRMR [6] with additional classier 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.

Fisher ratio evaluates feature based on its individual characteristics. It is well-known algorithm because of its simplicity and good performance in classification [2] [9].

F. Yang, K. Z. Mao [9] suggested method to increase the robustness of Feature Selection with multiple criteria fusion for feature evaluation. Recursive Feature Elimination algorithm based on Multi criterion fusion is developed called as MCF-RFE. Multiple criteria is 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. The basis criteria used are Fishers ratio, Relief, ADC and AW-SVM. Main goal was to enhance the Feature selection consequences in terms of both Classification stability and performance.

3. PROBLEM DEFINATION

Classification of microarray dataset poses main challenge owing to the large number of features as associated to the number of samples. This is a crucial problem in machine learning which is called as feature selection. Selecting a good subset of features regarding the objective models, an proficient way for reducing dimensionality, eliminating inappropriate data, increasing learning accuracy, and improving result directly can be achieved using feature subset selection. 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, traditional forms of regularization technique to LDA did not focused on minority class with diagonal emphasis. So, based on above criteria, LDA based feature subset selection technique with regularization on HDSS data with class imbalance is proposedwhich focuses on minority class and is experimentally evaluated.

4. IMPLEMENTATION DETAILS

A system is designed to provide analysis of high dimensional dataset having small size and class imbalance. This system provides a solution to identify feature set selection using minority class with the help of new regularization technique: Minority Class Emphasized Linear Discriminant Analysis - MCELDA. This technique overcomes the problems faced by traditional LDA technique such as singular scatter matrix and overfitting and overwhelming. SVM is used as a classifier which produces remarkable results towards gene microarray data.

4.1 Block Diagram

Most important step is data preprocessing. After this, evaluate class, viz., majority class and minority class. Here, minority class is emphasized with regularization technique.

Proposed System is evaluated in distributed environment to reduce the time complexity.

Proposed system is described in Fig. 1.

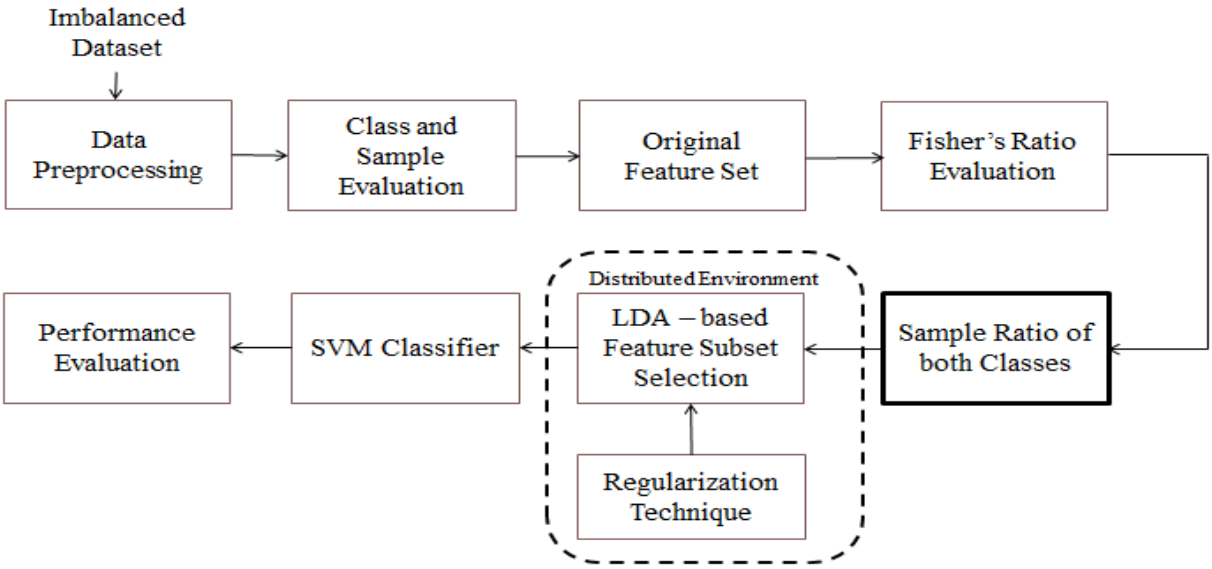

Figure 1: Block Diagram of Proposed System

Taking some support from paper [1] following concepts is described as follows:

4.1.1 Feature Evaluation

For evaluation of features, Fisher's Ratio and LDA based evaluation is used.

a. Fisher Ratio

Fishers Ratio independently evaluates the significance of each feature as given below [1]:

$$FR(X_j) = \frac{(m_{j(1)} - m_{j(2)})^2}{\sigma^2_{j(1)} + \sigma^2_{j(2)}} \quad (1)$$

Where, $m_{j(1)}$ and $\sigma^2_{j(2)}$ are sample mean and variance of feature X_j of class c, c =1, 2.

The feature is more informative if FR value is larger. Whole discriminating power of the feature subset $f^{(t)}$ calculated by Fishers ratio will be:

$$FR(F^{(t)}) = FR(X_1) + FR(X_2) + \ldots + FR(X_t) \quad (2)$$

b. LDA-based Evaluation

LDA considers correlations between features when a group of features are evaluated.

For a given feature subset $F^{(t)}$, all samples are projected by LDA from t-dimensional space to new dimension [1] and then using Fishers ratio the goodness of feature subset will be evaluated using following equation:

$$FR(F^{(t)}) = \frac{wS_B w^T}{wS_W w^T} \quad (3)$$

Where, w is the $1 \times t$ projection vector

S_B *and* S_W are the between-class and within-class scatter matrix defined below:

And w^T is given by,

$$w^T = S_W^{-1}(m_{(1)} - m_{(2)})^T \quad (4)$$

$$S_B = (m_{(1)} - m_{(2)})^T (m_{(1)} - m_{(2)}) \quad (5)$$

When a feature X_j is added into $F^{(t-1)}$ to form $F^{(t)}$, there will be:

$$\Delta m_t = [\Delta m_{t-1}, \Delta m_j] \quad (6)$$

Where, Δm_t is between-class difference of sample mean of feature X_j

$$S_w = S_{W(1)} + S_{W(2)} \quad (7)$$

$$S_{W(C)} = \frac{1}{n_c - 1} \sum_{x \in class\ c} (x - m_{(c)})^T (x - m_{(c)}) \quad (8)$$

Where,

m_c *and* $S_{W(c)}$ are the mean vector and with-class covariance matrix of class c

n_c is the number of samples of class c

Finally, LDA will evaluate a feature subset $F^{(t)}$ as follows:

$$FR(F^{(t)}) = (m_{(1)} - m_{(2)}) S_W^{-1} (m_{(1)} - m_{(2)})^T \quad (9)$$

4.1.2 Regularization Forms to LDA

Regularization techniques [1] [13] are used to deal with a singularity as well as with overfitting problems.

First form of regularization is to add a small positive constant to the diagonal of the scatter matrix. Another form of regularization is the so-called shrinkage technique. It shrinks individual within-class scatter matrices in towards the pooled scatter matrix S_p.

Here, Regularization technique [1] on which we are focusing are on the minority class first and then on the diagonal.

$$S_W(\gamma) = \gamma S_{W(1)} + (1 - \gamma) S_{W(2)} \quad (10)$$

$$S_W(\rho, \gamma) = \rho S_W(\gamma) + (1 - \rho) diag(S_W(\gamma)) \quad (11)$$

Where, $\gamma \in (0, 0.5)$ when n1 > n2 and
$\gamma \in (0.5, 0)$ when n1 < n2

$\rho \in [0, 1]$

4.2 Incremental approach of LDA based Feature Subset Selection

Input: D: Dataset file with pairwise sample

$$< x(1), y(1) >, < x(2), y(2) >, \ldots < x(n), y(n) >$$

Output:$F^{(d)}$: feature subset of d selected attribute from D

Processing:

1. Initialize $F^{(d)} = \{\}$
2. Evaluate Fisher ratio FR of each attribute using eq.(1)
3. Identify attribute X_j with max(FR_j)
4. Add X_j in feature subset i.e. $F^{(t)} = \{X_j\}$
5. Remove X_j from Attribute set X
6. For each attribute i in X
   - calculate between-class difference of sample mean Δm_t using eq.(6)
   - within class scatter matrix S_w as eq. (8)
7. Calculate attribute with maximum relevance
8. Identify attribute with maximum relevance
9. Add Attribute in X_i in feature set
$$F^{(t)} = F^{(t-1)} \cup X_i$$
10. Remove X_j from Attribute set X
11. Repeat steps 6 to 9 until t = d

This technique mainly focuses on minority class by improving overwhelming of majority class hence there is remarkable improvement in classifier's performance on datasets. Compare resultant feature subset with different sample distribution ratio of majority and minority class. Sample ratio distribution is done by considering equal ratio of both classes. To improve efficiency, system is implemented in distributed environment.

5. EXPERIMENTAL SETUP AND RESULT DISCUSSION

5.1 Performance Measure

Performance evaluation metric plays a significant role in assessing both performance of classification and guiding the classifier modeling.

AUC considers class imbalance and hence it is considered as better measure than accuracy.

Table1. Confusion Matrix

	Predictive Class	
True Class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

a. AUC

The Area under a ROC Curve (AUC) is only measure of a classifiers performance for evaluation of which model is superior on average.

$$AUC = \frac{TPrate + TNrate}{2}$$

b. Accuracy

$$Acc = \frac{TP+TN}{TP+FN+FP+TN}$$

5.2 Datasets

Datasets related to microarray data are used for evaluating the effectiveness and performance.

The proposed method is estimated using three publicly available microarray datasets. DLBCL [1], Prostate dataset has been used. Both datasets are of binary classification problems. DLBCL dataset includes 77 samples and 7,129 features. Prostate dataset consists of 12,600 features and 136 samples.

5.3 Results

Following graph shows the AUC comparison between different numbers of feature subsets using LDA-based feature subset selection. Evaluation is carried out on DLBCL dataset and Prostate datasets [1]. Parameters are set to as follows: γ=0 and ρ=0.1

Table 2. Accuracy (in %) of class and diagonal emphasis on DLBCL and Prostate dataset

	No. of Features	DLBCL(Outcome)			Prostate		
		Imbalanced Dataset	Random Oversampling	Random Undersampling	Imbalanced Dataset	Random Oversampling	Random Undersampling
Class Emphasis (γ=0)	50	68.97	70.31	51.92	76.19	88.24	62.75
	100	65.52	70.31	59.62	76.19	88.24	68.75
	150	74.14	73.44	61.54	73.81	92.65	68.75
	200	62.07	65.63	55.77	71.43	92.65	68.75
	250	58.62	67.19	51.92	71.43	92.65	68.75
	300	55.17	64.06	51.92	73.81	92.65	68.75
Diagonal Emphasis (ρ=0.1)	50	82.76	79.69	73.08	80.95	92.65	100
	100	79.31	79.69	73.08	90.48	95.59	93.75
	150	77.59	78.13	82.69	90.48	95.59	93.75
	200	82.76	78.13	82.69	90.48	94.12	93.75
	250	84.48	81.25	84.62	85.71	94.12	87.50
	300	86.21	81.25	82.69	85.71	94.12	87.50

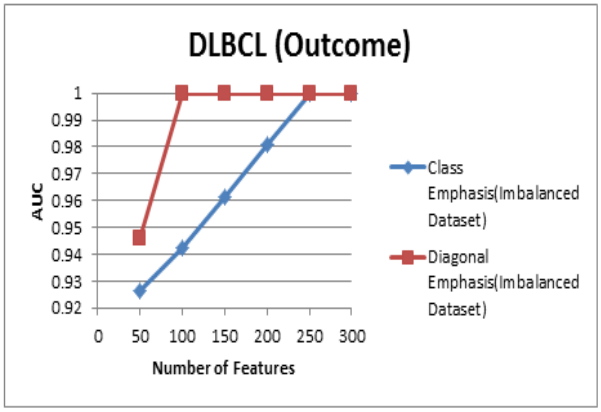

Figure 2 Comparative resultsof Class and diagonal emphasis on DLBCL (Outcome) dataset

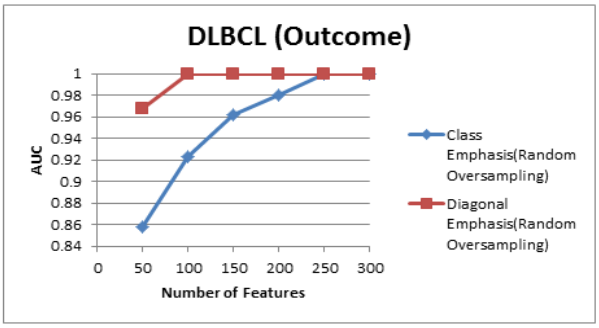

Figure 3 Comparative results of Class and diagonal emphasis using oversampling techniqueon DLBCL (Outcome) dataset

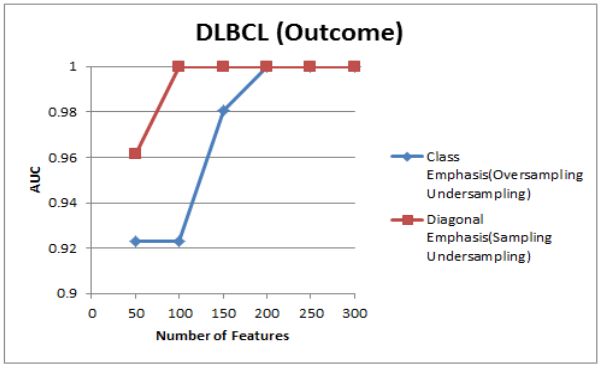

Figure 4 Comparative results of Class and diagonal emphasis using undersampling technique on DLBCL (Outcome) dataset

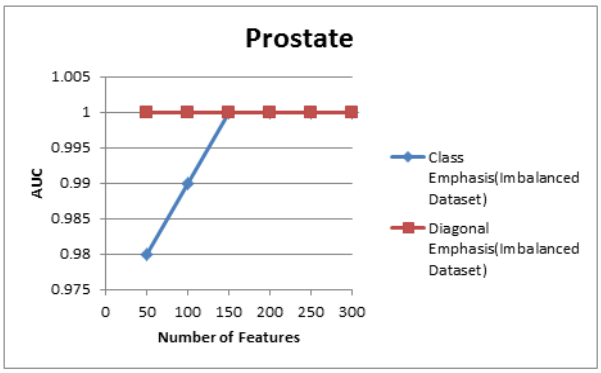

Figure 2 Comparative results of Class and diagonal emphasis onProstate dataset

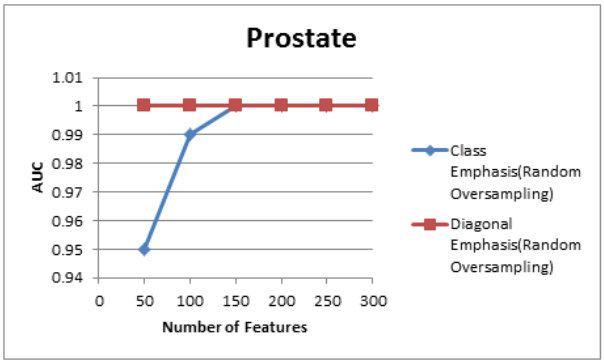

Figure 3 Comparative results of Class and diagonal emphasis using oversampling technique on Prostate dataset

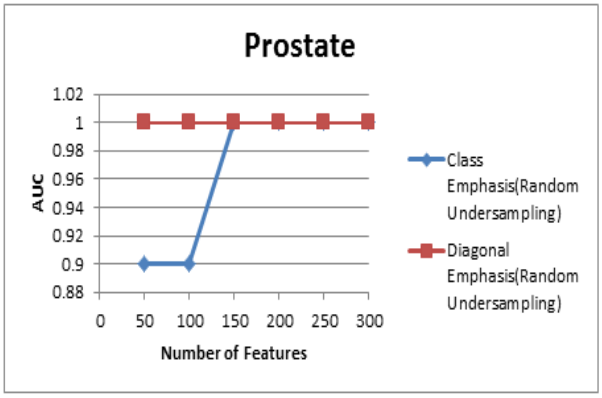

Figure 4 Comparative results of Class and diagonal emphasis using undersampling technique onProstate dataset

With respect to computational time, the proposed distributed system takes 38.465 seconds on DLBCL dataset with 300 feature set.

6. CONCLUSION

LDA-based feature subset selection is performed using regularization technique, which emphasis on class first and then on diagonal of matrix. Sample ratios of both classes are considered by increasing as well as decreasing the samples. Classifier performs better on DLBCL and Prostate dataset when feature subset selection is performed on sample ratio distribution. In distributed environment, time complexity is reduced.

7. ACKNOWLEDGMENT

I would like to express my sentiments of gratitude to all who rendered their valuable help. I am thankful to my guide Prof. S. S. Banait, for his guidance and encouragement in this work. His expert suggestions and scholarly feedback had greatly enhanced the effectiveness of this work. I am also thankful to family for their support.

8. REFERENCES

[1] Feng Yang, K.Z. Mao, Gary KeeKhoon Lee, And Wenyin Tang, Emphasizing Minority Class In LDA For Feature Subset Selection On High-Dimensional Small-Sized Problems, IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 1, January 2015.

[2] JIANG Zhu, Zhao Fei, Feature Selection for High-Dimensional and Small Sized Data Based on Multi Criterion Fusion, Journal of Convergence Information Technology(JCIT) Volume 7, Number 19, Oct 2012.

[3] M. Robnik Sikonja and I. Kononenko, Theoretical and empirical analysis of relieff and rrelieff, Machine Learning, vol. 53, no. 1 2, pp. 23 69, 2003.

[4] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou, On the Class Imbalance Problem.

[5] Y. Tang, Y.Q. Tang, and Z. Huang, Development of two stage SVM RFE gene selection strategy for microarray expression data analysis, IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 4, no. 3, pp. 365381, Jul.Sep. 2007.

[6] H. Peng, F. Long, and C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max relevance, and min redundancy, IEEE Transactions on

Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 12261238, Aug2005.

[7] K. Javed, H. A. Babri, and M. Saeed, Feature selection based on class-dependent densities for high dimensional binary data, IEEE Transactions on Knowledge and Data Engineering, vol. 99, no. PrePrints, 2010.

[8] X. Zhou and K. Z. Mao, The ties problem resulting from counting-based error estimators and its impact on gene section algorithms, Bioinformatics, vol. 22, no. 20, pp. 25072515, 2006.

[9] F. Yang and K. Z. Mao, Robust feature selection for microarray data based on multi-criterion fusion, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 4, pp. 10801092, 2011.

[10] Y. Cheung and H. Zeng, Local kernel regression score for selecting features of high dimensional data, IEEE Transactions on Knowledge and Data Engineering, vol. 21, pp. 17981802, December 2009.

[11] M. Wasikowski and X. wen Chen, Combating the small sample class imbalance problem using feature selection, IEEE Transactions on Knowledge and Data Engineering, vol. 22, pp. 13881400, 2010.

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, " Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, no. 1-3, pp. 389-422, 2002.

[13] Tatyana V. Bandos, Lorenzo Bruzzone, and Gustavo CampsValls, Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis, IEEE Transactions on Geoscience and remote sensing, VOL. 47, NO. 3, MARCH 2009.