Indian Journal of Science and Technology, *Vol 8(11), DOI: 10.17485/ijst/2015/v8i11/71766, June 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Accurate and Stable Feature Selection Powered by Iterative Backward Selection and Cumulative Ranking Score of Features

G. Victo Sudha George[1*] and V. Cyril Raj[2]

[1]Department of Computer Science and Engineering
[2]Engineering and Technology, Dr. M.G.R Educational and Research Institute University, Chennai-600095, Tamil Nadu, India; sudhajose72@gmail.com, cyrilraj@drmgrdu.ac.in

Abstract

This paper focuses on a stable feature selection framework using Cross Validation technique and SVM-RFE. Though SVM-RFE has outperformed many of its counterparts in feature subset selection for accurate cancer classification, its greediness in selecting optimal feature subset affect the stability of selection process in successive runs that brings down the confidence on the selected features. In this paper, we propose an iterative backward feature selection method using SVM-RFE motivated by cross-validation technique. Cumulative Ranking Score (CRS) is a parameter formulated to determine the class discrimination ability of each feature. The proposed method is applied on the publically available breast cancer dataset and found top 10 highly discriminative genes. Later the SVM classifier is trained using the top 10 genes identified by the proposed method and the original SVM-RFE separately and tested. It is proved that the proposed method has improved the classification accuracy significantly compared to the original SVM-RFE.

Keywords: Cross Validation, Cumulative Ranking Score, Stable Feature Selection, SVM-RFE

1. Introduction

Feature selection has been an active research area in data mining communities as it allows improving the comprehensibility of the resulting classifier models substantially[1]. It selects a fraction of input features from a dataset with very large number of features by removing features with little or no suggestive information. Various feature selection algorithms have been developed[2,3] with an intention of enhancing the classification precision while bringing down the dimensionality.

Besides high precision, another important concern is the stability (the insensitivity of the result of a feature selection algorithm to variations to the training set) of feature selection. This subject is particularly essential for applications where feature selection is used as a knowledge-discovery tool to identify characteristic features and clarify the observed occurrence. For example, in Microarray analysis, biologists are concerned in discovering a small number of features that explain the mechanism driving different behaviors of microarray data[4]. A feature selection algorithm often decides on largely different subsets of features under variations to the training data, though most of these subsets are on par with each other in terms of classification performance[5–7]. Such unsteadiness reduces the poise of domain experts in validating the selected features experimentally. Therefore it is essential to establish a robust methodology to select the significant variables not vulnerable of selection bias[8] and to use appropriate statistical indicators to enumerate and assess the significance of the results.

The stability of feature selection is an intricate affair. Recent studies on this concern[5,6] have shown that the stability of feature selection results is influenced by various

*Author for correspondence

factors namely data distribution, mechanism of feature selection, and sample size. Moreover, the stability of feature selection results should be investigated together with the predictive performance of the selected features.

Generally, the feature selection algorithms find out features that differentiate between classes but not the extent of contribution of each feature in discriminating between classes. To tackle this problem, we put forward a parameter CRS to calculate the degree of participation of each feature in discriminating between classes. This parameter aggregates the ranking of the features obtained from different subsets of samples being generated using Cross Validation approach from a dataset. Due to its successful use in selecting informative genes for cancer classification, SVM-RFE gained immense popularity and is well known as one of the most competent feature selection method as given by [9–12]. Hence in this work we have used it as a baseline. Though SVM-RFE has its own merits it has limitations like greediness in selecting optimal feature subset and not reusing the feature once it is removed in iteration. And it is proved that the reuse of features formerly removed during the SVM–RFE process can progress the performance of SVM classifier[13].

In the proposed method sample variation is introduced from the same dataset using Cross Validation technique to improve the reliability of the SVM-RFE which indirectly addresses the problem, curse of dimensionality (small sample size and more number of features). Here the Microarray gene expression data is used to validate the proposed method as it is a good example for the above addressed issues.

2. Materials and Methods

2.1 Datasets

Here the Microarray, breast cancer data set GSE15852[14] from GEO is used to identify the highly discriminative genes. It consists of the gene expression profiles of 22285 genes for 86 tissue samples among which 43 are breast tumors and 43 are normal tissues. To reduce the complexity of computation, an initial filtering was carried out using BRB Array Tool to eliminate the irrelevant noisy genes by setting $p<0.001$ and fold change as 3. The resultant 613 genes expression profiles are used by the proposed method to identify the highly discriminative genes.

In this study, we consider discrimination between two classes of samples, the breast cancer tissues and the normal tissues. The data sets are represented in the form of data matrix. The total size of the matrix is N x n where N is the total number of samples and n is the total number of genes in each sample.

2.2 Format of the Input Data Matrix

Each row represents a gene and each column a sample. Each row contains the expression value of a gene of all the samples. The data format used is given in Figure 1.

2.3 SVM-RFE

Although simpler feature selection methods are existing[15], SVM-RFE is used as a baseline as it has acknowledged good classification performance and is widely used in Microarray data analysis. Fundamentally, SVM-RFE is a multivariate iterative backward feature selection method in the sense that it considers feature interaction while evaluating the relevance of features. At each iteration the algorithm trains a linear SVM classifier based on the remaining set of features, ranks them according to the squared values of feature weights in the optimal hyper plane, and eliminates a feature with the least weight, from full set of features. This Recursive Feature Elimination process continues until all features have been removed or a desired condition is met. We used SVM-RFE a soft-margin based SVM using linear kernel.

2.4 K-fold Cross-validation

In k-fold cross-validation, the original sets of samples are divided randomly into k equal sized subgroups. Of the k subgroups, a single subgroup is treated as the validation data for testing the classification model being built, and the remaining k – 1 sub groups are used for training the classifier. The process is repeated k times (the folds), with each of the k subgroups used only once for validating the classification model being build. Thus, if we are selecting d out of D features in this way, k different feature sets of dimension d may be chosen. Now, the features with high degree of occurrence[16,17] might be used in the classification system in future. In this work, 9-fold cross validation is used to enhance the stability of feature selection against sample variation. Moreover during each iteration the cumulative ranking of all the features are computed. Finally the features with high cumulative ranking are chosen.

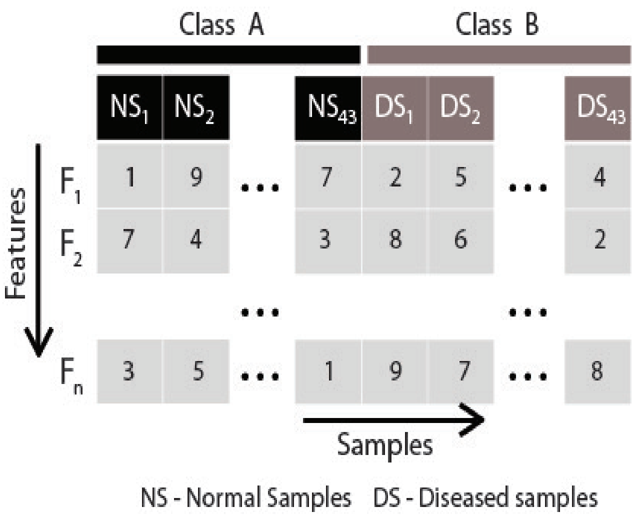

Figure 1. Input data matrix format.

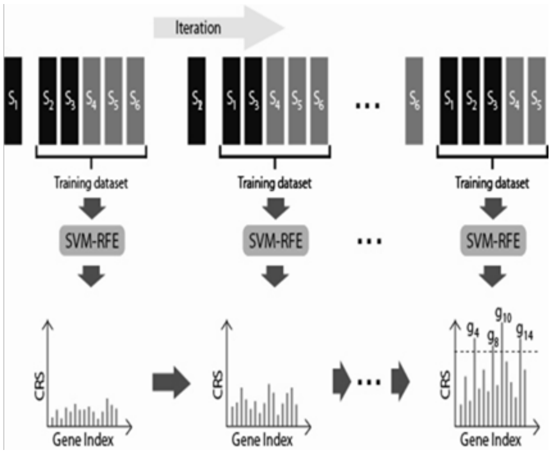

Figure 2. Frame work of the proposed method.

2.5 Proposed Iterative Backward Feature Selection Method

The proposed method is an iterative feature selection approach using SVM-RFE. Though SVM-RFE is one of the dominant methods for feature selection its performance can go down due to small sample size, sample variation and noisy data. Here 9-CV is used to generate various training sets from the same set of samples. The 86 samples available in the dataset is divided into 9 groups of subsamples out of which 8 groups of size 10 and the last group having 6 samples. Each time 10 samples are eliminated from the input and the residual samples are used as training set. In each iteration a gene ranking set is generated and Gene Ranking Matrix (GRMAT) is constructed from which CRS is computed.

Finally genes having high CRS are selected and validated against classification accuracy. The frame work of the proposed method is as given in Figure 2. The steps to be followed are given below.

1. Divide the data set into k subgroups.
2. Select a subgroup as testing dataset and keep all the remaining samples as training set.
3. Train SVM-RFE using the training set and generates the ranking of all the genes.
4. The rank of the genes GR^1_k ,………., GR^n_k is stored in $GRMAT_{k,n}$.
5. Repeat steps 2 to 4 and k times.
6. Finally the $GRMAT_{k,n}$ contain the gene index number of all the genes of all the iterations.
7. Compute CRS of all the genes using $GRMAT_{k,n}$ and select the highly robust and discriminative genes having high CRS values.
8. Validate the classification model using the selected high CRS scored genes.

2.6 Cumulative Ranking Score (CRS)

CRS is the parameter ($0 \le CRS \le 1$) that defines the class discrimination ability of all the genes, which is influenced by the ranking of the genes. It is calculated for all the genes in each iteration. This parameter is formulated to select a robust and precise set of genes which are claimed to be the true biomarkers for a disease. CRS ≤(i) ≤ (1) holds the CRS of the gene with i as gene index. Finally the genes having CRS greater than the threshold value 0.20 are selected. The genes with high CRS are the highly discriminative genes. The pseudo code for computing CRS is given below.

```
Pseudo Code for computing CRS of all the genes
For i= 1 to n // for all the genes
    CRS(i) =0
End For
For R=1 to n// for all the  genes
For S= 1 to N// for each sample
CRS(GRMAT(S,R)=CRS(GRMAT(S,R)+(n+1)–R
End for
End For
For i= 1 to n
CRS(i) = CRS(i)/(N* n) // Normalized CRS value
End For
```

3. Results and Discussion

3.1 Evaluation

The classification accuracy is computed against the top CRS scored genes using Linear SVM and LOOCV. A classification model is built and classification test was conducted for each test sample by adding a gene in decreasing order of CRS. The process as stated above is repeated for all test samples. For each cumulative sample size the classification accuracy, the percentage of correctly classified test samples are calculated.

Figure 3 is the CRS graph for all genes in which the horizontal axis indicates the gene index and the vertical axis indicates the CRS value. This CRS graph is drawn by applying the proposed Iterative SVM–RFE to the entire samples. We see that CRS of the top 7 genes are significantly higher than those of the other genes. Thus, these genes are expected to be crucial for the discrimination between the two classes.

Figure 4 shows the classification accuracies with a cumulative increase of the number of genes (up to the 30th) in decreasing order of CRS. As expected, the classification accuracy exhibits a rapid increase up to the 7th gene and accomplishes 100% result. An unstable and a fall in classification accuracy is observed from the 8th gene onwards. This proves the fact that the discriminative power of the gene weakens with the increase of ranks. Nevertheless, it is noteworthy that it restricts the accuracies from being greater than 89 %.

Figure 5 shows the classification accuracies obtained by SVM-RFE, the original method that trains only a single dataset. Note that the horizontal axis is the cumulative number of genes ranked by one time execution of SVM-RFE. We see that the classification accuracy of SVM-RFE is considerably inferior not more than 84% compared to that of proposed iterative method that learns different training datasets by Cross-Validation, and preferentially extracts features showing a stronger discriminant power.

3.2 Literature Proof of the Top 10 Genes

Here the biological relevance of the top 10 genes from literature is discussed and also the details of the top 10 genes identified by our method are given in Table 1. The examination of the functional analysis of KRT19 in human breast cancer was carried out in [18] and result shows KRT19 is a potential tumor suppressor. To evaluate CD24 protein expression in breast cancer an experiment was carried out in [19]. The result proved that CD24 expression in primary breast cancer might be a new marker for more aggressive breast cancer biology. The differential expression of both basal-like cytokeratins (KRT5, KRT6A, KRT6B, KRT14, KRT16, KRT17, KRT23, and KRT81) and luminal cytokeratins (KRT7, KRT8, KRT18, and KRT19) across the subtypes TNBC tumor subtypes is witnessed by [20]. The result of the experiment carried out in [21] shows that GATA3 plays an integral role in breast luminal cell differentiation and is implicated in breast cancer progression. The Tumor-Associated Calcium Signal Transducer 2 (TACSTD2) gene has been reported to be highly expressed in many types of human epithelial cancers, and is associated with tumor metastasis and poor prognosis[22]. The association of AGR2 in estrogen receptor (ER)-positive tumors is identified by [23].

The conclusion says, AGR2 is a promising drug target in breast cancer and may serve as a useful prognostic indicator as well as a marker of breast cancer metastasis.

The result of [24] shows that cell type dependent modification of Wnt signaling components after EpCAM over expression in breast cancer cell lines, which results in marginal functional changes.

It is demonstrated in [25] Comprehensive genomic profiling of relapsed CDH1-mutated ILC revealed actionable genomic alterations in 86% of cases. It identified in[26] estrogen is activated expression of MUC1/SEC in human breast cancer epithelial cells. The literature proof clearly shows that all the top 10genes identified by the proposed method play a crucial role in breast cancer which is a good indication that the proposed method is effective in identifying highly robust and class discriminative genes.

4. Conclusion

A stable feature selection method inspired by CV technique and SVM-RFE has been proposed. The iterative method generates different training set from the same data set for each iteration. Cumulative Ranking Score is calculated from the successive iterations for all the genes. Finally the genes are ranked based on their CRS.

The proposed method is validated using a breast cancer data set. We have proved that our method has succeeded in identifying highly robust and class discriminative genes with good biological relevance. Also it is shown that the classification accuracy has improved significantly while using the highly ranked genes selected by the proposed method compared to the original SVM-RFE algorithm.

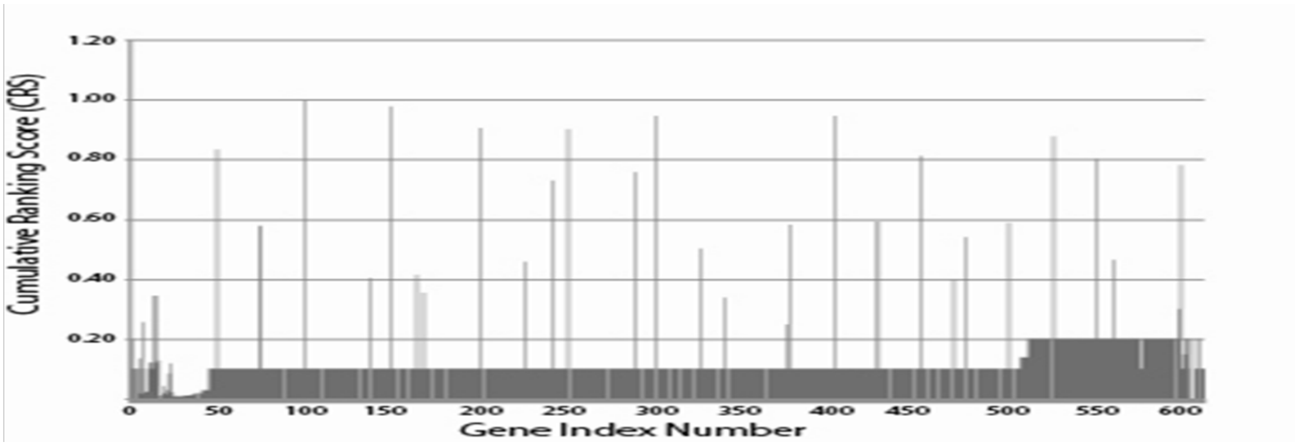

Figure 3. CRS graph drawn for all the genes.

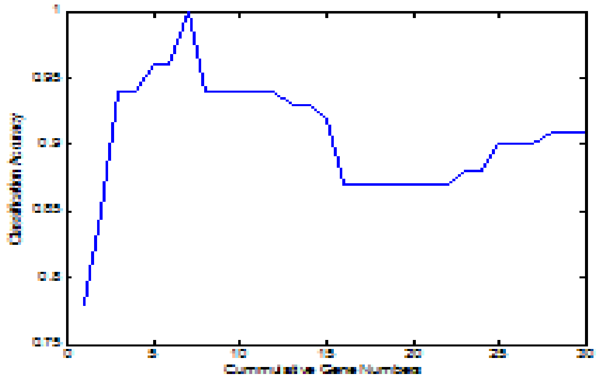

Figure 4. Classification accuracy of the proposed method.

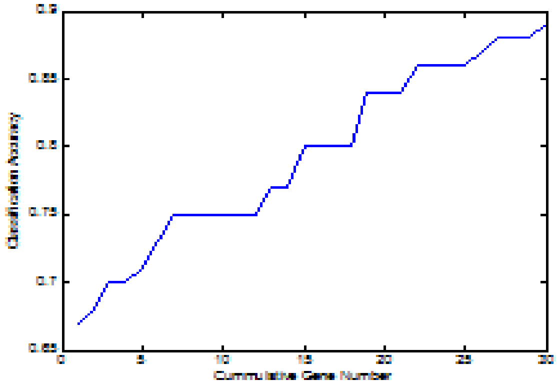

Figure 5. Classification accuracy of original SVM-RFE.

Table 1. Details of the Top 10 Genes

Cumulative Ranking Score (CRS)	Probe Set	Symbol	Name	Accession Number
1	201650_at	KRT19	keratin 19	NM_002276
0.98	208650_s_at	CD24	CD24 molecule	BG327863
0.95	209016_s_at	KRT7	keratin 7	BC002700
0.95	209602_s_at	GATA3	GATA binding protein 3	AI796169
0.91	202286_s_at	TACSTD2	tumor-associated calcium signal transducer 2	J04152
0.90	201596_x_at	KRT18	keratin 18	NM_000224
0.88	209173_at	AGR2	anterior gradient 2 homolog (Xenopus laevis)	AF088867
0.84	201839_s_at	EPCAM	epithelial cell adhesion molecule	NM_002354
0.81	201131_s_at	CDH1	cadherin 1, type 1, E-cadherin (epithelial)	NM_004360
0.80	213693_s_at	MUC1	mucin 1, cell surface associated	AI610869

5. References

1. Liu RSH, Motoda H, Zhao Z. Feature selection: An ever evolving frontier in data mining. JMLR: Workshop and Conference Proceedings 10: The Fourth Workshop on Feature Selection in Data Mining; 2010. p. 4–13.
2. Zeng Z, Zhang H, Zhang R, Zhang Y. A hybrid feature selection method based on rough conditional mutual information and naive Bayesian Classifier. ISRN Applied Mathematics. 2014; 2014:382738.
3. Das K, Ray J, Mishra D. Gene selection using information theory and statistical approach. Indian Journal of Science and Technology, 2015; 8(8):695–701.
4. Kumar AP, Valsala P. Feature selection for high dimensional DNA microarray data using hybrid approaches. Bioinformation. 2013; 9(16):824–8.
5. Student S, Fujarewicz K. Stable feature selection and classification algorithms for multiclass microarray data. Biology Direct. 2012; 7(33).
6. Li Y, Jiangsu. FREL: a stable feature selection algorithm. IEEE Transactions on Neural Networks and Learning Systems. 2014.
7. Loscalzo S, Yu L, Ding C. Consensus group stable feature selection. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09); 2009. p. 567–76.
8. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA. 2002. p. 6562–6.
9. Yang Z-M, He J-Y, Shao Y-H. Feature selection based on linear twin support vector machines. Procedia Computer Science. 2013; 17:1039–46.
10. Huang M-L, Hung Y-H, Lee WM, Li RK, Jiang B-R. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. The Scientific World Journal. 2014; 795624.
11. Mundra PA, Rajapakse JC. SVM-RFE with MRMR filters for gene selection. IEEE Transactions on NanoBioscience. 2010; 9(1):31–7.
12. Samb ML, Camara F, Ndiaye S, Slimani Y, Esseghir MA. A novel RFE-SVM-based feature selection approach for classification. International Journal of Advanced Science and Technology. 2012; 43.
13. Pau Ni IB, Zakaria Z, Muhammad R, Abdullah N, et al. Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context. Pathol Res Pract.2010; 206(4):223–8.
14. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A, Benítez JM, Herrera F. A review of microarray datasets and applied feature selection methods. Inform Sci. 2014; 282:111–35.
15. Tian Rui, Basu MK, Capriotti1 E. Contrast Rank: a new method for ranking putative cancer driver genes and classification of tumor samples. Bioinformatics. 2014; 30(17):i572–8.
16. Chin Lynda, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. Genes & Dev. 2015; 25:534–55.
17. Schuleruda H, Albregtsen F. Many are called, but few are chosen. Feature selection and error estimation in high dimensional spaces. Comput Meth Programs Biomed. 2004; 73:91–9.
18. Ju JH, Yang W, Lee KM, Oh S, Nam K, Shim S, Shin SY, Gye MC, Chu IS, Shin I. Regulation of cell proliferation and migration by keratin19-induced nuclear import of early growth response-1 in breast cancer cells. Clin Cancer Res. 2013; 19(16):4335–46.
19. Hosonaga M, Arima Y, Sugihara E, Kohno N, Saya H. Expression of CD24 is associated with HER2 expression and supports HER2-Akt signaling in HER2-positive breast cancer cells. Canc Sci. 2014; 105(7):779–87.
20. Lehmann BD, Bauer JA, Chen Xi, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Investig. 2011; 121(7):2750–67.
21. Cimino-Mathews A, Subhawong AP, Illei PB, Sharma R, Halushka MK, Vang R, Fetting JH, Park BH, Argani P. GATA3 expression in breast carcinoma: utility in triple-negative, sarcomatoid, and metastatic carcinomas. Hum Pathol. 2013; 44(7):1341–9.
22. Lin H, Huang JF, Qiu JR, Zhang HL, Tang XJ, Li H, Wang CJ, Wang ZC, ZQ F, Zhu J. Significantly upregulated TACSTD2 and Cyclin D1 correlate with poor prognosis of invasive ductal breast cancer. Exp Mol Pathol. 2013; 94(1):73–8.
23. Salmans ML, Zhao F, Andersen B. The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker. Breast Cancer Res. 2013; 15(2):204.
24. Gostner JM, Fong D, Wrulich OA, Lehne F, Zitt M, Hermann M, Krobitsch S, Martowicz A, Gastl G, Spizzo G. Effects of EpCAM over expression on human breast cancer cell lines. BMC Cancer. 2011; 11:45.
25. Ross JS, Wang K, Sheehan CE, et al. Relapsed classic E-Cadherin (CDH1)–mutated invasive lobular breast cancer shows a high frequency of HER2 (ERBB2) gene mutations. Clin Cancer Res. 2013; 19:2668–76.
26. Lacunza E, Baudis M, Colussi AG, Segal-Eiras A, Croce MV, Abba MC. MUC1 oncogene amplification correlates with protein over expression in invasive breast carcinoma cells. Canc Genet Cytogenet. 2010; 201(2):102–10.