A Kernel-Based Multivariate Feature Selection Method For Cancer Data Classification

Megha Purohit[1]
[1]Student, Computer Engineering, SAL Institute of Technology & Engineering Research Ahmedabad, India.

ABSTRACT

High dimensionality and small sample sizes, and their inborn danger of over fitting, pose extraordinary difficulties for building proficient classifiers in malignancy data grouping. Consequently a feature selection procedure ought to be directed preceding data classification to improve prediction performance. Overall, filter techniques can be considered as essential or assistant selection system on account of their effortlessness, adaptability, and low computational many-sided quality. Nonetheless, a progression of inconsequential cases demonstrate that filter techniques result in less precise execution since they disregard the conditions of features.

Albeit few publications have committed their regard for uncover the relationship of features by multivariate-based techniques, these strategies depict connections among elements just by linear techniques. While straightforward linear combination relationship limits the transformation in execution.

In this paper, we utilized kernel method for svm-RFE with MRMR way to deal with find inalienable nonlinear connections among features and also amongst feature and target. So as to uncover the viability of our technique we played out a few analyses and thought about the outcomes between our technique and other aggressive multivariate-based features selectors. In our examination, we utilized three classifiers (support vector machine, neural system and average perceptron) on two gathering datasets, to be specific two-class and multi-class datasets (principally focused on svm).

Exploratory results show that the execution of our technique is superior to anything others, particularly on three hard group datasets, to be specific Wang's Breast Cancer, Gordon's Lung Adenocarcinoma and Pomeroy's Medulloblastoma.

Keyword: - *Machine Learning, Multi class classification, Feature Selection*

1. INTRODUCTION

The field of machine learning is flourishing by the feature selection which depends on the data mining strategies. As of late, numerous high measurement/small example issues of territories, for example, natural language processing, biological data, monetary and budgetary, system, telecom and restorative data examination required to convey feature selection before upgrading a supervised learning or unsupervised learning. There are a few managed data mining strategies that it is hard to determine which one coagulates better with the bio-informatics data.

Along these lines, appraisal of data mining techniques is generally completed to choose an effective technique to renounce the bio-informatics issues. Correspondingly, there are numerous adjustments and variants of feature selection recommended by literature however everything relies on upon the data like money, natural, galactic and so on. In this manner, assessment of every methodology is important to know which FS technique can be utilized for specific classification.

Various articles gave correlation either among classification techniques or feature selection strategies which can't affirm best blend of FS technique and classifier. Besides, classification headways like binary and multi class classifiers ought to be assessed with feature selection technique are henceforth, an analysis required that can better assess every classifier with every feature selection technique.

2. FEATURE SELECTION AND CLASSIFICATION ADVANCEMENTS

Filter, wrapper and embedded methods are habitually used to carry out a comparison study to evaluate the better method suitable for biological dataset.

2.1 Filters

Filter techniques choose variables without taking care of its type. Filter method gives superiority to the least captivating variables. The added variables will be an allotment of the model classification to allocate or statistics prediction. These techniques are accurately able in ciphering time and able-bodied to over fitting [1]. Although, filter techniques have an inclination to pick out outmoded variables due to the fact that they do not keep in mind the relationships between variables. Consequently, they are especially used as a pre-process method.

2.2 Wrappers

Excessive dimensionality is a top notch trouble for bio informatics dataset. The crucial reason of wrapper feature determination is building a model that utilizing a planned element subset and the use of the presence of this model as a score for the advantage of that subset. While developing a model, various options must be made in the best approach to assemble and look at the model. While this model might be built utilizing the whole preparing set and after that has its general execution assessed contrary to that equivalent preparing set, this would conceivably bring about over fitting [2]. Wrapper techniques assess subsets of variables which grant, dissimilar to filter approaches to deal with find the conceivable associations between variables [3].

2.3 Embedded

As of late, embedded strategies have been proposed to decrease the order of machine learning. They are attempting to blend the advantages of each first procedure. The machine learning algorithms take advantages of their own variable determination algorithms. Thus, it needs to understand that what an great choice is which confines their misuse [4]. Partially on account of the higher computational intricacy of wrapper and a lesser degree embedded approaches, these procedures have not got great arrangements as long as the filter proposition [5].

2.4 Classification

Thus, Final best featured set is connected on either classification or clustering. Proposed analysis is centered around to a great degree appreciated and progressive supervised learning classification which depends on a model which can predict classes of cases from the data set. On the off chance that we discuss medical data, supervised learning like decision trees, simulated neural systems, SVM (Support vector machine), regression tree, KNN (K Nearest Neighborhood) has demonstrated fine results [6, 7, 2]. An assortment of classification methods have been displayed subsequent to recent years for medical applications. Classification strategies were comprehensively classification into one class or binary arrangement, multi class classification and hierarchy multi class classification.

Here classification property is bringing about to just two discrete qualities. They depend on 1. Indirect methodology and they are one against one, one against one, all against all and directed acyclic graph SVM 2. Direct approach endeavor to discover separate limits for all classes in one stage [8, 9, 10]. Numerous articles turned out based on these essential systems for multi class grouping [11, 12]. Despite the fact that they are being utilized generally have a few drawbacks that they are capable to form only one measure at a time henceforth it devours more computational power and even costly. Likewise is troublesome and protracted numerical execution [13]. There is presumably no multiclass method that beats the entire set. The selection of the procedure must be made depending on the requirements like the wanted level of exactness, the time accessibility for advancement and preparing. It additionally relies on which sorts of issues are emerging. However, selecting the pleasing one is an exceptionally tough assignment.

3. METHODS

Filter methods might be isolated into two classes, univariate-based methods and multivariate-based methods. Univariate method procedures have pulled in much enthusiasm because of their low many-sided quality and quick general execution for over the top dimensionality of microarray data analysis [14]. Nonetheless, a couple of valuable features disposed of through univariate techniques may likewise have striking commitment for arrangement.

Along these lines, the vital cause in their less exact general execution is that they ignore the results of capacity co operations [15]. The utilizations of multivariate filter methods are simple bivariate-

essentially based techniques which are about in view of entropy (or restrictive entropy) and common insights, comprising of MRMR, CFS and a few variations of the Markov blanket filter approach. However, they also abandon probably redundant variables which can bring about a performance loss [16].

Partial least squares (signified as PLS), which shares the qualities of various regression and feature transformation strategies (which incorporates accepted connection analysis and fundamental part assessment), has set up to be valuable in conditions when the quantity of found variables are impressively more than the scope of perceptions [17].

In various expressions, PLS is a well known technique to determine issues when there might be intemperate multi collinearity amongst functions. SlimPLS, PLSRFE, and TotalPLS are multivariate-fundamentally based feature selection techniques that have been proposed by the method for Gut kin et al. furthermore You et al., individually.

3.1 Kernel PLS RFE with MRMR

K-PLS RFE is one of the prevalent uses of a class of multivariate statistical analysis technique presented by [18], and a famous regression system in Chemo metrics [19]. It varies from different strategies in developing the principal relations between two matrices (X and Y) by method for latent variables called segments, prompting a closefisted model which imparted qualities to other regression and feature transformation systems [20]. The objective of K-PLS RFE with MRMR is to figure vectors of its X-weight (v), Y-weight (c), X-score (t) and Y-score (u) by an iterative technique for the improvement issue:

$$\arg\max \|v\| = 1, \|c\| = 1 \operatorname{cov}(t,u) = \operatorname{cov}(X_v, Y_c)$$

Where $t = X_v$ and $u = Y_c$, are called segments of X and Y, respectively.

At the point when the initial two segments t_1 and u_1 are acquired, the second pair t_2 and u_2 is separated from their residuals $E_x = X - t_1 p^T$ and $E_Y = Y - t_1 q^T$, separately.

Here p and q are called the loadings of t concerning X and Y, respectively.

This procedure can be rehashed until the required stop condition is satisfied. The detail description of the algorithm can be found in [21]. The kernel version of PLS uses a nonlinear transformation Ø (.) to map gene expression data into a higher-dimensional (even unending dimensional) kernel space K; i.e. mapping Ø: $X_i \in IR^D \rightarrow Ø(X_i) \in K$. However, we don't have to know the particular numerical articulation of nonlinear mapping, we just need to express the whole algorithm as far as dot products between sets of inputs and substitute kernel function K(,,.) for it. This is supposed to call the "Kernel trick".

In classification to state dot product operation in the algorithm, we can restrict v to have a place with the linear spans of the points. They can therefore be communicated as:

$$v = (\Phi(x_1), \cdots, \Phi(x_N))\beta^{\Phi}$$

$$t^{\Phi} = \begin{pmatrix} \Phi(x_1) \\ \vdots \\ \Phi(x_N) \end{pmatrix} v = \begin{pmatrix} \Phi(x_1) \\ \vdots \\ \Phi(x_N) \end{pmatrix} (\Phi(x_1), \cdots, \Phi(x_N))\beta^{\Phi} = \mathbf{K}_X \beta^{\Phi}$$

Let K_x (X_i, X_j) be a feature of the Gram matrix K_x in feature space and h is the coveted number of features. Collapsing Y will, be that as it may, be required for kernel partial least squares.

The primary part for kernel PLS can be resolved as Eigen vector of the following square kernel version matrix for $\beta^{Ø}$: $\beta^{Ø}\lambda = K_Y K_X \beta^{Ø}$, where l is an Eigen value. The measure of the kernel matrix K_Y K_X is N×N. Subsequently, regardless of what number of variables are in the first matrices X and Y, the measure of these kernel matrices won't be get influenced by it.

Therefore, the combination of PLS with kernel creates an intense algorithm that will solve this issue quickly and adequately with MRMR approach.

3.2 The importance of each feature

In original space, let T is a set of features, T= {t_1, t_2, t_3... t_n} the addition of variant clarification of T to Y is given by

$$w_i = \sqrt{D \frac{\sum_{l=1}^{h} \Psi(Y,t_l) v_{il}^2}{\sum_{l=1}^{h} \Psi(Y,t_l)}}, \; i \in \{1,2,3,\dots,D\}. \quad (1)$$

Where h is the quantity of features and V_{il} is the weight of the i^{th} feature for the l^{th} segment.

$$\Psi(Y, t_1) = \Sigma_{j=1}^{c} \Psi(y_j, t_l)$$

It is the connection amongst t_l and Y, where Y (i, j) is correlation function. The bigger estimation of w_i is the more explanatory force of the i^{th} feature to Y. It is important that the above condition can likewise be utilized as a part of kernel space. The reason is holding of condition $\emptyset(y_j) = y_j$ because here y_j is a class label. So the expression $\psi(\emptyset(y_j), t_l^{\emptyset})$ can be expressed as $\psi(y_j, t_l^{\emptyset})$, here $t_l^{\emptyset} \epsilon T^{\emptyset}$ and $T^{\emptyset} = \{t_1^{\emptyset}, t_2^{\emptyset}, \dots, t_h^{\emptyset}\}$.

Table 1: Algorithm kPLS RFE with MRMR

1. Input: Dataset
2. Output: give highest ranked and highest weighted features
3. Begin
4. Set β
5. Given set of features, S⊂G
6. Ranked set of features, R={ }
7. Repeat
8. Train linear SVM with feature set S
9. Calculate the weight of each feature w_i
10. Calculating the component
11. Calculating the contribution of the l^{th} component γ_l,

$$\gamma_l = \frac{\sum_{i=1}^{M} N_i m_{il}^{\Phi}}{\sum_{i=1}^{M} N_i};$$

12. L=l+1
13. End while
    H=l-1
14. Calculating the weight of each feature w via equation-1)
15. Return w
16. For each feature i ϵ S do
17. Compute $R_{s,i}$ and $Q_{s,i}$
18. Compute γ_i
19. End for
20. Select the feature with smallest ranking score, $i^* = \arg\min\{\gamma_i\}$
21. Update $R = R \cup \{i^*\}$; $S = S \setminus \{i^*\}$;
22. Until all features are ranked
23. End: output R

4. ANALYSIS AND RESULTS

4.1 Data Set Details

In this experiment, we have data sources which are mentioned below:

Table 2 the cancer classification datasets used in the paper

Class	Dataset	Sample	Feature	Class
Two-class	AMLALL	72	7129	2
	Breast	209	22283	2
	Lung	86	7129	2
	Prostate	102	12600	2
	DLBCL	77	7129	2
	Medulloblastoma	60	7129	2
Multi-class	Stjude	215	12558	7
	Lymphoma	62	4026	3
	SRBCT	83	2308	4
	MLL	72	8685	3
	Lung	203	3312	5

AML ALL (A) [22]
There are two sections containing the preliminary (train), 38 bone marrow tests from two classes: 27 instances of intense lymphoblastic leukemia (ALL) and 11 instances of intense myeloid leukemia (AML); free (test), 34 tests from two classes: 20 instances of ALL and 14 instances of AML. Every case is portrayed by expression levels of 7129 tests from 6817 human genes.
Source: http://wwwgenome. wi.mit.edu/cgi-bin/cancer/datasets.cgi;

Breast Cancer (B) [23]
The dataset utilized the raw force Affym-etrix CEL records and standardized the data by RMA systems. A last expression matrix containing 22283 elements and 209 examples, 71 of which are from patients, the rest 138 specimens are ordinary examples.
Source: http://math.bu.edu/people/sray/software/prediction;

Lung Cancer (L) [23]
This dataset contains 86 tests: 24 are tumor tests and 62 are typical controls, 7129 genes with most elevated intensity over the samples are considered. Source: http://math.bu.edu/people/sray/software/prediction/;

Prostate Cancer (P) [24]
This dataset contains 52 prostate tumor tests and 50 ordinary specimens with 12600 genes. An autonomous arrangement of testing tests is produced from the training data, 25 tumor and 9 ordinary examples are separated by Singh's production.
Source (training): http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi;

DLBCL (D) [25]
The objective of this dataset is to recognize diffuse huge B-cell lymphoma (DLBCL) from follicular lymphoma (FL) morphology. This dataset contains 58 DLBCL tests and 19 FL tests. The expression profile contains 7129 genes. Source: http://wwwgenome.wi.mit.edu/mpr/prostate;

Medulloblastoma (M) [15]
Patients result forecast for focal sensory system embryonic tumour. Survivors are patients who are alive after treatment whiles the disappointments are those who succumbed to their infection. The dataset contains 60 patient examples, 21 are
Survivors and 39 are failures. There are 7129 genes in the dataset. Source:
http://www-genome.wi.mit.edu/mpr/CNS;

Stjude (S) [14]
The dataset has been divided into six diagnostic groups, BCR-ABL (9 samples), E2APBX1 (18 samples), Hyper diploid > 50 (42 samples), MLL (14 samples), T-ALL (28samples) and TEL-AML1 (52 samples)), and one that includes diagnostic samples (52samples) that did not now in shape into any one of the above groups. There are 12558genes. Source: http://www.stjuderesearch.org/data/ALL1;

Lymphoma (Ly) [16]
The dataset consists of measurements of 4026 genes from 62 patients. The sufferers are classified into 3 classes: lymphoma and leukemia (DLCL, forty two samples), follicular lymphoma (FL, 9 samples) and chronic lymphocytic leukemia (CLL, eleven samples). Source: http://llmpp.nih.gov/lymphoma;

SRBCT (SR) [17]
The dataset carries 83 samples and 2,308 gene expression values. It may be divided into four divisions, the Ewing family of tumours (EWS), Burkett lymphoma (BL), neuro blastoma (NB) and rhabdomyosarcoma (RMS). Some of the 83 samples, 29, 11, 18, and 25 samples belong to training EWS, BL, NB and RMS, respectively.
Source:http://www.biomedcentral.com/content/supplementary/1471-2105-7-228-S4.tgz.

MLL (ML) [17]

The dataset includes 72 samples in 3 training classes, acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), and mixed-lineage leukemia gene (MLL), which have 24, 28, 20 samples, respectively. In our test, we obtained a dataset with 72samples and 8685 genes.
Source:http://www.biomedcentral.com/content/supplementary/1471-2105-7-228-S4.tgz.

Lung (Lu) [17]
The total of this dataset carries 203 samples with 12600 genes in 5 data classes, adenocarcinoma (139), squamous cell lung carcinomas (21), pulmonary carcinoids(20), small-cell lung carcinomas(6) and regular lung (17). We received a dataset with203 samples and 3312 genes.
Source:http://www.biomedcentral.com/content/supplementary/1471-2105-7-228-S4.tgz.

5. COMPARISON OF GENES

Table 3 Description of genes reported by existing published papers and ranked by our method

Accession number	Gene description	Rank
X95735_at	Zyxin	4
M23197_at	CD33	8
U22376_cds2_s_at	C-myb	74
M27891_at	Cystatin C	21
M16038_at	LYN	11
M84526_at	DF(adipsin)	9
M27783_s_at	ELA2 Elastatse 2	80
U50136_ma1_at	LTC4 synthase	3
Y12670_at	Leptin receptor	2
U46499_at	Glutathione	96
L09209_s_at	Amyloid beta	48
U46751_at	p62	19
M55150_at	Fumarylacetoacetate	7
M83652_s_at	Properdin	22
M80254_at	CyP3	17
X17042_at	Proteoglycan 1	10
U82759_at	HoxA9	8

In our first experiment, we used two datasets, namely the Leukemia data (two-class) of [26] and the Lymphoma data (three class) of [27], to compare our method with previous works with respect to the selected genes.
For the Leukemia data, we collected several most important genes (in table 3) that were published in several papers. It can readily be seen that three probes, X95735_at, M27891_at and M23197_at were reported by five published papers, and their ranking by our method are 4th, 17st and 8st, respectively. We notice that there are many overlapping of genes among the list of papers.
For Leukemia data, the top-ranked 10 features obtained by our procedure are shown in table 4 in which genes are in columns from 1 to 10. There is a worthwhile result achieved by our method, that is, it obtained the genes with the highest weight.
Many of these genes are known as differentially expressed genes by many foregoing studies. 10 out of 40 genes are listed in this table that was also selected by [26], which shows the effectiveness of our method.
The top 10 genes ranked by our procedure are listed in table 5. From the table, we can see that important genes can be captured easily by our method. There are many genes that are also chosen by [28].
Table 3 illustrates the differentially expressed genes for two datasets, namely the Leukemia data and the Lymphoma data. No single gene is uniformly expressed across the class; all these genes as a group appear correlated with class which is illustrating the effectiveness of the Kernel PLS method.
In Table 4 the top panel is consist of three genes GENE1622X, GENE2402X and GENE1648X.Bottom panel compose of three genes, namely GENE1602X, GENE681X and GENE1618X.
In Table 4 the top panel shows three probes highly express in AML and the bottom panel shows three probes more highly expression in AML.

6. COMPARISON OF SEVERAL MULTIVARIATE-BASED FEATURE SELECTORS

Table 4 Top ranked 10 features for Leukemia data

1.	M23197_at	6.	X04085_ma1_at
2.	Y12670_at	7.	M55150_at
3.	U50136_ma1_at	8.	U82759_at
4.	X95735_at	9.	M84526_at
5.	D49950_at	10.	X17042_at

In our first test, we used datasets, particularly the Leukemia data (two-class) of [22] and the Lymphoma data (three classes) of [16], to examine our technique with previous works with admire to the chosen genes. For the Leukemia records and Lymphoma records, we collected numerous most important genes (in table 2) that have been published in several papers. It could easily be visible that 3 probes, X95735_at, M27891_at and M23197_at were reported with the aid of 5 published papers, and their ranking through our technique are 4th, 17th and 8th, respectively.

Table 5 Top ranked 10 features for Lymphoma data

1.	GENE1622X	6.	GENE1647X
2.	GENE2403X	7.	GENE1610X
3.	GENE653X	8.	GENE2402X
4.	GENE1644X	9.	GENE1648X
5.	GENE1607X	10.	GENE1643X

For Leukemia data and Lymphoma data, the top-ranked 10 functions acquired through our system are shown in table 4 and 5 respectively in which genes are in columns from 1 to 10. There's a worthwhile result performed by way of our method, it obtained the genes with the very best weight.

Table 6 Comparison of SVM-kernel PLS RFE with MRMR and four other models of svm on two class dataset

Dataset	MRMR		Svm RFE		kPLS		SVM-kPLS with MRMR	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
AML ALL	97.5	100	96.3	100	94.6	100	96.1	100
Breast	68.0	69.2	69.9	67.5	72.2	71.5	72.7	75.4
Lung	77.4	81.5	72.1	76.5	76.8	77.6	77.4	82.6
DLBCL	94.8	99.2	94.8	99.2	93.4	98.3	97.5	100
Medulloblastoma	71.7	72.9	70	73.1	70	77.2	73.3	82.7
Prostate	96.0	97.5	96.0	96.7	95.1	98.7	97.3	97.9
Avg.	84.2	86.7	83.2	85.5	83.7	87.2	85.7	89.8

Table 7 Comparison of SVM-kernel PLS RFE with MRMR and four other models of svm on multi class dataset

Dataset	MRMR		Svm RFE		kPLS		SVM-kPLS with MRMR	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
St	88.9	0.866	86.4	0.851	86.8	0.834	89.9	0.876
Ly	**100**	**1**	96.7	0.933	**100**	**1**	**100**	**1**
Lu	76.9	0.399	74.6	0.382	74.5	0.36	79.2	0.532
ML	93.2	0.884	87.8	0.801	90.3	0.834	95.8	0.92
SR	98.8	0.983	97.6	0.964	98.9	0.983	97.6	0.964
Avg.	91.6	0.826	88.6	0.786	90.1	0.802	92.7	0.86

Table 8 Running time of 5 feature filtering algorithm for binary class & multi class

Class	Dataset	MRMR	SVMRFE	PLS	kPLS	kPLS with MRMR
Binary class	AML ALL	5.1510	52.5854	210.4046	12.1222	0.0891
	Breast	5.1496	88.6176	>1e+003	10.6423	0.1092
	Lung	7.5420	52.898	693.1857	16.8629	0.2410
	DLBCL	5.5614	53.109	221.2261	12.0526	0.097
	Medulloblastoma	5.1343	51.997	421.8250	19.2384	0.268
	Prostate	18.1848	65.108	>1e+003	64.2148	0.6010
Multi-class	St	34.0030	67.5321	>1e+003	>1e+003	**2.1180**
	Ly	2.733	5.7846	217.2568	27.9456	**0.2361**
	Lu	10.253	9.7816	>1e+003	17.8940	**0.5500**
	ML	6.643	8.7484	791.0244	98.8890	**0.259**
	SR	1.8230	5.8336	87.6536	8.8784	**0.1714**

Table 6 &7 authenticates the excessive overall performance by means of SVM-kernel PLS with MRMR over different techniques for SVM classifier. Here one could see that SVM-kernel PLS with MRMR provide outperforming results for all datasets by way of attaining accuracies and coefficients values advanced than all other strategies. As an end the overall excessive average Acc and AUC values in both tables display the effectiveness and significance of our method as compare to different popular techniques.

Both Acc and AUC values of our technique have higher values among others and eventually the average consequences likewise are nice. Despite the fact that for few datasets our results are just like their outcomes but in these instances time taken by our approach is extensively smaller than different techniques. As an instance in table 7 for AMLALL dataset, along with our technique, the AUC is 100% for lots strategies but time consumed up via our method is most effective 0.0891 s even as the time taken by way of other techniques, mRMR, SVMrfe and PLS, kPLS are approximately 5 s, 52 s, 210 s and 12 s, respectively (see in table 8). So time intake by means of our algorithm is regularly less than others which depict standard well overall performance of our method.

Table 9 Performance statistics with other classifiers

Classifiers	Overall accuracy	Average accuracy
NN	0.639024	0.928705
AP	0.634146	0.926829
SVM	0.64878	0.929756

CONCLUSION

Best of literature studied there numerous feature selection methods exist which emphasis on redundancy and sometime they even discard features those are mutually attached. So here has been used MRMR approach with kPLS-RFE. Moreover, stability of algorithm also increases as changing in training set less likely to affect the performance. The approach has also dealt with the major difficulty of high dimensionality even in small sample size and accuracy maintained even after increasing number of classes. For classification, state-of-art classifier svm has discovered accomplishment in an assortment of regions. Here the Linear SVM classifier utilized with filter choice technique.

In this paper, described an effective multivariate-based feature filter method for cancer classification, namely, kernel PLS RFE with MRMR filter method. It showed that gene-gene interactions cannot be ignored in feature selection techniques to improve classification performance. In other words the nonlinear relationship of gene-gene interactions is a vital concept that can be taken into account to enhance accuracy. To capture these nonlinear relations of interaction between genes here used kernel method because kernel method can be used to reveal the intrinsic relationships that are hidden in the raw data. In order to capture the reasonable number of components, it makes use of the relationship between PLS and linear discriminant analysis to determine the number of components in kernel space based on kernel linear discriminant analysis. To verify the importance of gene-gene interactions also compared our feature selector with other multivariate-based feature selection methods by using classifier SVM. Experimental results, expressed as both accuracy (Acc) and area under the ROC curve (AUC), showed that our method leads to promising improvement in ACC and AUC. The conclusion is that the gene-gene interactions, nonlinear relationships of gene-gene interactions are core interactions that can improve classification accuracy, efficiently.

REFERENCES

1.[J. Hammon, November 2013]. “Optimization combinatoire pour la sélection de variables en régression en grande dimension”: Application en génétiqueanimale.

2.[RandallWald, Taghi M. Khoshgoftaar]. “Optimizing Wrapper-Based Feature Selection for Use on Bioinformatics Data”, Amri Napolitano Florida Atlantic University.

3.[T. M. Phuong, Z. Lin et R. B. Altman, 2005]. “Choosing SNPs using feature selection. Proceeding, IEEE Computational Systems Bioinformatics Conference, pages 301-309.

4.[B. Duval, J.-K. Haoet J. C. Hernandez Hernandez, 2009]. “A memetic algorithm for gene selection and molecular classification of an cancer”. In Proceedings of the 11th Annual conference on Genetic and evolutionary computation, GECCO '09, pages 201-208, New York, NY, USA.

5.[Yvan Saeys, IñakiInza and Pedro Larrañaga]. “Review of feature selection techniques in bioinformatics”.

6.Zhang, Min-Ling, José M. Peña, and Victor Robles. "Feature selection for multi-label naive Bayes classification." Data Sciences 179, no. 19 (2009): 3218–3229

7.Dietterich, Thomas G., Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles." Artificial Intelligence 89, no. 1-2 (1997): 31–71.

8.Hsu, C.-W., and C.-J. Lin, C.-J., 2002, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks, 13, 415-425.JAMES, G., 1998

9.http://link.springer.com/article/10.1007/s10462-009-9114-9 A review on the combination of binary classifiers in multiclass problems

10. X. Chen, X. Zeng, and D. van Alphen. Multi-class feature selection for texture classification. Pattern Recognition Letters, 27(14):1685{1691, 2006
11. G. Madazrov and D. Gjorgjevikj. Evaluation of distance measures for multi-class classification in binary svm decision tree. In Artificial Intelligence and Soft Computing: 10th International Conference, (ICAISC), 2010.
12. A.C.Lorena, A.C.Carvalho, J.M.Gama, are view on the combination of binary classifiers in multi-class problems, Artificial Intelligence Review30 (1–4) (2008)19–37.
13. F. Aiolli and A. Sperduti. An efficient SMO-like algorithm for multiclass SVM. In Proceedings of IEEE workshop on Neural Networks for Signal Processing, pages 297–306, 2002
14. http://www.stjuderesearch.org/data/ALL1
15. http://www-genome.wi.mit.edu/mpr/CNS
16. http://llmpp.nih.gov/lymphoma
17. http://www.biomedcentral.com/content/supplementary/1471-2105-7-228-S4.tgz.
18. Wold H (1966) Estimation of principal features and related models byiterative least squares. Multivariate Analysis. New York: Academic.
19. Ra´nnar S, Lindgren F, Geladi P, Wold S (1994) A pls kernel algorithm for datasets with many variables and fewer objects. Part 1: Theory and algorithm.Journal of Chemometrics 8: 111–125.
20. Wold S, Ruhe A, Wold H, Dunn IW (1984) thecollinearity problem in linearregression. The partial least squares (pls) approach to generalized inverses. SIAM. Journal of Scientific and Statistical Computations 5: 735-743.
21. Gutkin M, Shamir R, Dror G (2009) Slimpls: a method for feature selection ingene expression-based disease classification. Plos One 4.
22. http://wwwgenome.wi.mit.edu/cgi-bin/cancer/datasets.cgi
23. http://math.bu.edu/people/sray/software/prediction
24. http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
25. http://wwwgenome.wi.mit.edu/mpr/prostate
26. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999)Molecular classification of cancer: class discovery and class prediction by geneexpression monitoring. Science 286: 531–537.
27. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. (2000) Distinct typesof diffuse large b-cell lymphoma identified by gene expression profiling. Nature403: 503–511.
28. Dramiski M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, et al. (2008)Monte carlo feature selection for supervised classification. Bioinformatics 24:110–117.