International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.6 (2014), pp.347-360
http://dx.doi.org/10.14257/ijmue.2014.9.6.33

SSiCP: a new SVM based Recursive Feature Elimination Algorithm for Multiclass Cancer Classification

Xiaobo Li[1], Xue Gong[2], Xiaoning Peng[3] and Sihua Peng[4, *]

[1]*Department of Computer Science and Technology, College of Engineering, Lishui University, Lishui 323000, China*
[2]*Department of Microbiology and Immunology, School of Medicine, Stanford University, Stanford, CA 94305-5101, USA*
[3]*Department of Internal Medicine, School of Medicine, Hunan Normal University, ChangSha 410006, China*
[4]*Department of Biological Technology, School of Fisheries and Life Science, Shanghai Ocean University, Shanghai 201306, China*

[1]*oboaixil@126.com*, [2]*xuegong@stanford.edu*, [3]*pxiaoning@hunu.edu.cn*,
[4] *shpeng@shou.edu.cn*

Abstract

An extremely crucial step in the diagnosis of cancers is to select a small number of informative genes for accurate classification. This issue has become a hot focus in the data mining of gene expression profiles. Especially for data with a large number of cancer types, many conventional classification methods show very poor performance. Here, we proposed a new approach for gene selection and multi-cancer classification based on step-by-step improvement of classification performance (SSiCP). The SSiCP gene selection algorithms were evaluated over the NCI60 and GCM benchmark datasets, with accuracy of 96.6% and 95.5% in 10-fold cross-validation, respectively. Furthermore, the SSiCP outperformed recently published algorithms when applied to another two multi-cancer data sets. Computational evidence indicated that SSiCP can avoid overfitting effectively. Compared with various gene selection algorithms, the implementation of SSiCP is simple and many of the selected genes by SSiCP are shown to be closely related to cancers.

Keywords: *Multiclass cancer classification; gene expression profile; machine learning; data mining; gene selection*

1. Introduction

Cancer classification is an extremely crucial step for diagnosis and treatment of cancers. Without the correct identification of cancer types, it is almost impossible to achieve a satisfactory therapeutic effect. Based on the DNA microarray technology for cancer identification and classification, many in-depth studies have been done [1, 2]. As for classification with two classes, such as classification between normal and tumor tissues [3], or classification between one subtype and another of a tumor [4], molecular classification using microarray data has obtained a fairly high degree of accuracy. For classification of multiple tumor types, however, the accuracy is yet to be improved [5-8]. Because of the high

*Corresponding author: Sihua Peng, shpeng@shou.edu.cn, +86-21-61900491 (Tel), No.999, Huchenghuan Rd , Nanhui New City, Shanghai 201306, P.R. China.

dimension of the feature space, the excessive noise, and the relatively small sample sizes in DNA microarray data, this issue is under active research in the data mining of gene expression profiles. Especially for data with a large number of cancer types, many conventional classification methods show very poor performance [9], such as the NCI60 data set (9 types of cancer) [5], and the GCM data set (14 types of cancer) [6].

Many researchers proposed some other new methods [10-16]. However, obtaining higher classification accuracy for choosing fewer genes is possible by using more powerful data mining algorithms. In this paper, we proposed a new approach of gene selection and multiclass cancer classification based on step-by-step improvement of classification performance (SSiCP). SSiCP, which is neither support vector machine recursive feature elimination (SVM-RFE) algorithm nor the expansion of SVM-RFE [17], is a new support vector machine (SVM) based implementation of recursive feature elimination (RFE) feature selection methodology. The results show that our strategy is very effective, with a fast calculation procedure.

2. Materials and Methods

2.1. Data Sets

NCI60 dataset [5]. The dataset can be downloaded at: http://wwwgenome. wi.mit.edu/mpr/NCI60/NCI_60.expression.scfrs.txt. There are 60 samples in this dataset, which express 7,129 genes in nine types. Since it contains only two samples, the prostate cancer type was excluded from this study.

GCM dataset [6, 7]. There are 198 samples in the original GCM dataset, which express 16,063 genes in 14 classes of cancers [6]. A subset of the original GCM dataset, which contains 89 samples belonging to 11 classes of tumors and normal samples (a total of 12 categories), was employed in this study, and the dataset was downloaded at the website: http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode= view&paper_id=114.

Human Carcinomas Dataset (HCD174) [18]. The HCD174 dataset contains 174 samples belonging to 11 classes. Each sample has 12,533 gene expressions. All samples were used. The dataset was obtained at the website: http://public.gnf.org/cancer/epican/.

Central Nervous System Embryonal Tumors dataset (CNS) [19]. The CNS dataset contains 42 samples with 7,129 gene probes and can be downloaded at: http://www.broad.mit.edu/mpr/CNS/. All samples were used.

2.2. Gene Pre-selection

Without feature pre-selection, the computation becomes a time-consuming task because of the very high dimensions in feature space. After gene pre-selection, we can obtain a few dozen or hundreds of differentially expressed genes. Based on this reduced gene subset, the second step of gene selection was carried out, with the calculation burden being greatly reduced. As our algorithm is based on the Weka (http://www.cs.waikato.ac.nz/ml/weka/) platform, we tested several feature selection methods on Weka. The chi-squared ($\chi2$) method was chosen as our gene pre-selection algorithm. The Chi-Squared method evaluates each feature individually by computing the $\chi2$ statistic with respect to the classes. The χ^2 value of each feature is calculated as

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(A_{ij} - E_{ij})^2}{E_{ij}}, \qquad (1)$$

where n is the number of classes, k is the number of intervals, A_{ij} is the observational frequency in the ith interval, jth class, and E_{ij} is the expected frequency of A_{ij}. After the χ2 values of all considered features being measured, the values are sorted in descending order, and the larger the χ2 value, the more important is the feature.

2.3. RFE: Recursive Feature Elimination

RFE is an iterative procedure and each iteration can be described as follows:

- *Training the classifier.*
- *Calculating the ranking score for all features.*
- *Eliminating the feature with smallest ranking score.*

In the algorithm of SVM-RFE proposed by guyon et al., each iteration is described as follows [17]:

- *Training the SVM classifier*
- *Calculating the weight vector* $w = \sum_k \alpha_k y_k x_k$ *, where* x_k *is the gene expression vector of a sample* k *,* $y_k \in [-1,+1]$ *encodes the class label of sample* k *, and parameter* α_i *is calculated from the training set.*
- *Calculating the ranking score:* $c_i = (w_i)^2$
- *Seeking the feature with smallest ranking score:* $f = \arg\min(c_i)$
- *Eliminating the feature with* $c = f$.

2.4. Feature Selection Methodology

Step by step feature reduction. SSiCP algorithm is not a kind of wrapper algorithm [20], which searches for an optimal feature subset by using a heuristic function to guide the search procedure. In SSiCP, we do not utilize a search method. Consequently, we do employ an evaluation function to guide the eliminate features step by step.

To some extent, SSiCP is similar to SVM-RFE in two aspects. Both of the algorithms are SVM based algorithm, and both of them employ the recursive feature elimination (RFE) methodology. Nevertheless, they are completely different algorithms. The innovation of our algorithm is the feature elimination criteria. Briefly, we eliminate a feature at a time. If the classifier yields a higher (or equal to the original value) classification accuracy without this feature, this feature is removed forever, otherwise this feature is restored back to the feature set. So SSiCP did not rank the features by some ranking criteria. The key steps of the algorithm proposed were as follows. The proposed SSiCP method in this study includes the following major steps:

Step 1. Train the classifier with n features (genes), and compute the accuracy k with m-fold cross-validation.

Step 2. Eliminate a feature f temporarily, and compute the accuracy $k^{'}$ with m-fold cross-validation.

Step 3. If $k \leq k^{'}$, remove the feature f, and if $k > k^{'}$, restore the feature f. If all the retained features were restored once without increasing $k^{'}$, a local maxima value of accuracy is obtained. In this case, we make $k = k^{'}$.

Step 4. If n=2, stop the calculation. If n>2 go to Step 2.

The above steps are the key points of our algorithm, and the details are shown in Figure 1. The SSiCP software package including the JAVA source code can be downloaded at the website: http://code.google.com/p/ssicp/.

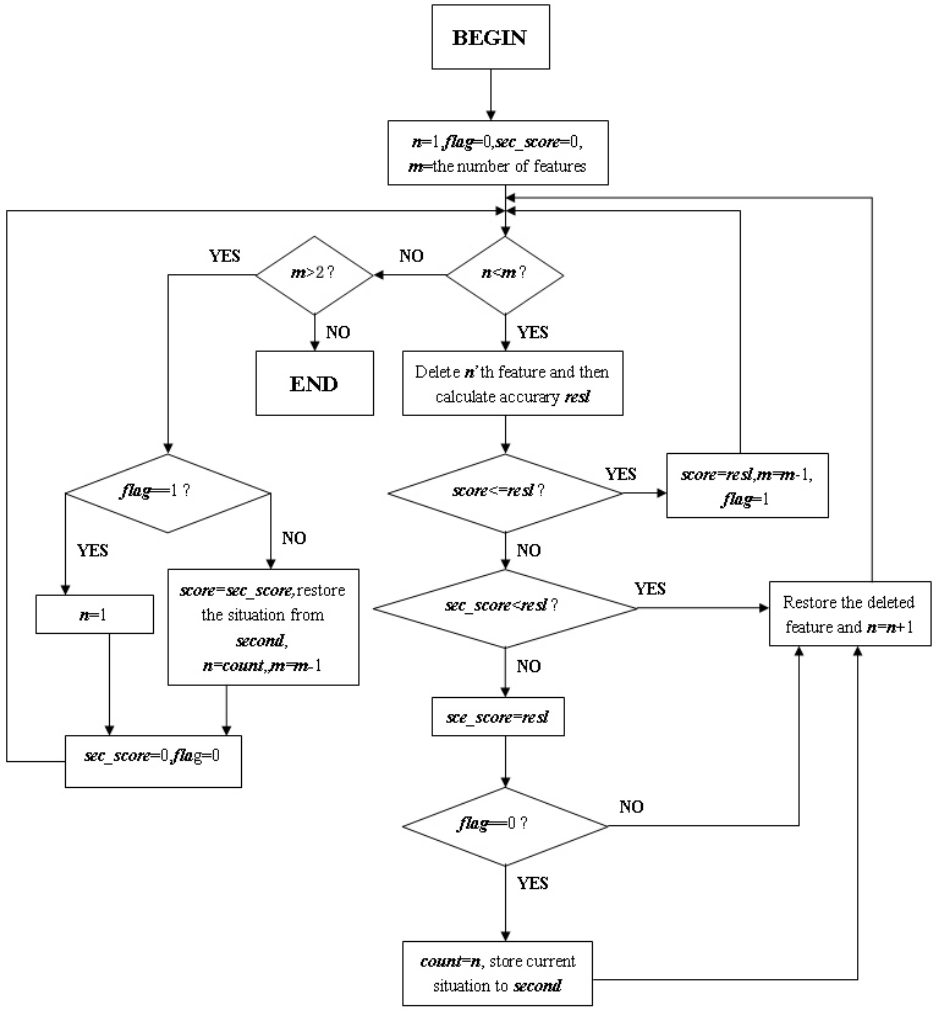

Figure 1. Schematic Map of SSiCP Algorithm

N denotes features in searching for the minimal feature number. The flag is used to determine the direction of the search. The sec score is used to determine the search efficiency.

2.5. Over-fitting Evaluation of SSiCP Algorithm

Overfitting is an important issue in machine learning model. Of the four datasets, there are more samples in HCD174 (174 samples) dataset than that of GCM, NCI60, and NCS. Therefore, to test the overfitting risk of SSiCP algorithm, HCD174 dataset is partitioned into two parts: training set and test set. A classification model is trained by running the SSiCP algorithm on the training dataset, with a ten-fold cross-validation accuracy denoted by x_1. The classifier model is then tested by the independent test dataset, with an accuracy denoted by x_2 . If the difference between x_1 and x_2 is very small, we conclude that SSiCP can avoid overfitting effectively.

2.6. Confirmation of Classification Algorithm in the Second Step of Feature Selection

By comparing the seven classification algorithms including the Naive Bayes classifier, the Bayes Network classifier, sequential minimal optimization algorithm for training a support vector classifier (SMO), KStar, logistic model trees (LMT), J48, and classifier for building linear logistic regression models (SL) (Weka: http://www.cs.waikato.ac.nz/~remco/weka.pdf), we determined the classification algorithm which provided the best performance.

By using the seven classification algorithms on the GCM and NCI60 data sets, the optimal algorithm was selected. Subsequent calculation results showed that SMO outperformed all of the other six algorithms. As a fast algorithm of support vector machines (SVM), the SMO algorithm implements John Platt's sequential minimal optimization algorithm for training a support vector classifier. In SMO, multi-class problems are solved using pair-wise classification.

2.7. Parameter Selection on Weka

When SVM was used to do the classification task, the choice of the kernel function of SVM was a key factor to obtain better performance. For the classification of the microarray dataset, a relatively better classification performance was achieved by using the polynomial kernel function [8]. After testing the four kernel functions (Normalized PolyKernel, PolyKernel, RBFKernel, and StringKernel) on Weka, it was also clear that the best results were achieved by using “PolyKernel”. On Weka, it is necessary to choose the "FilterType" parameter to do the data normalization or standardization, which is always a necessary procedure for data transformation in data mining tasks. Because the data sets we used were normalized, we selected the "Standardization" option, with which better results can be obtained. The other parameters on Weka were set to default values.

3. Results

3.1. Initial Noise Removal and Comparison of Classification Algorithms

The NCI60 and GCM datasets are generally considered as benchmark datasets in the microarray data mining problem, so they are always used to test the performance of a new algorithm. Therefore, seven classification algorithms which are commonly used in data mining issues were employed with these two datasets. First, we obtained the computational results with and without feature pre-selection (using the $\chi2$ test-based feature selection algorithm) (Table 1). The results suggested that after initial pre-selection of the features, the classification performance improved considerably, indicating that the noisy genes in the microarray datasets were removed to a certain extent. The results shown in Table 1 also indicated that when using both NCI60 data and GCM data, the SMO algorithm was superior to the other algorithms. After features (genes) pre-selection, 208 genes were selected from NCI 60 data set and 150 genes from GCM data set.

Table 1. Performance Comparison of Multi-class Classification using the Seven Algorithms on NCI60 and GCM Datasets (%)

		J48	LMT	KStar	SL	SMO	BNet	NB
NCI60	All features (7129)	38.3	53.3	15.0	53.3	**60.0**	55.0	38.3
	208 features	36.2	67.2	15.5	67.2	**84.5**	74.1	67.2
GCM	All features (16063)	42.7	70.8	23.6	70.8	**75.6**	/	28.1
	150 features	56.2	77.5	71.9	77.5	**84.3**	82.0	66.3

3.2. Gene Selection based on Step-by-step Improvement of Classification Performance

By calling the main package of Weka to run our algorithm, the computations were carried out using the NCI60 and GCM datasets, and the gene selection results of the above seven algorithms were obtained (Figure 2 and Figure 3). Clearly, the SMO algorithm also outperformed the other six algorithms in the second step of feature selection. By repeatedly calculating three times using the 208 genes from NCI 60 data set or the 150 genes from GCM data set, 24 genes (Table 2) or 28 genes (Table 3) can be obtained.

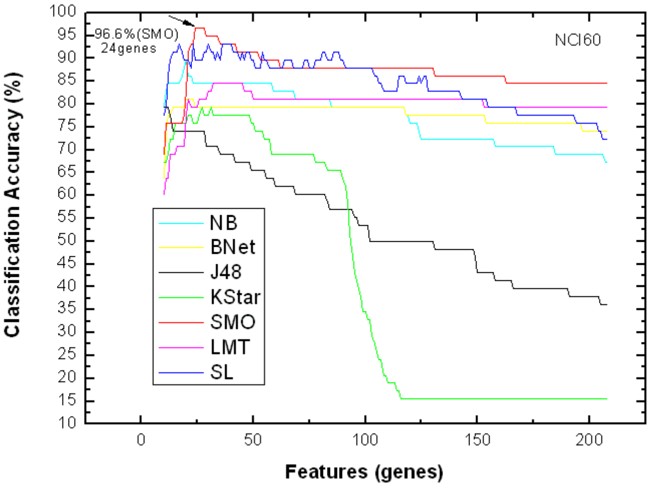

Figure 2. Classification performance comparisons of the seven algorithms by using NCI60 data set. The maximal accuracy of 96.6% was obtained by using SMO algorithm with the 24 genes (red)

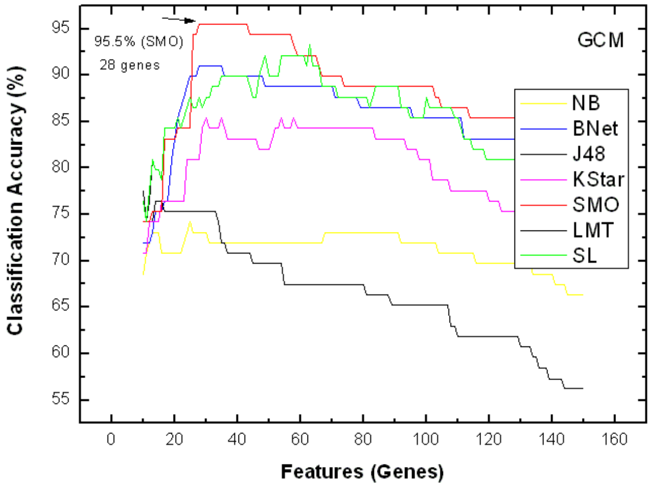

Figure 3. Classification performance comparisons of the seven algorithms by using GCM data set. The maximal accuracy of 95.5% was obtained by using SMO algorithm with the 28 genes (red)

Table 2. The 24 Genes selected from NCI60 Data Set

Probe Set	Gene Symbol	Gene Title
AD000684_cds1_at	LSR	lipolysis stimulated lipoprotein receptor
D31883_at	ABLIM1	actin binding LIM protein 1
D42073_at	RCN1	reticulocalbin 1, EF-hand calcium binding domain
D78611_at	MEST	mesoderm specific transcript homolog (mouse)
HG174-HT174_at	-	-
L41349_at	PLCB4	phospholipase C, beta 4
M14949_at	RRAS	related RAS viral (r-ras) oncogene homolog
U41813_at	HOXA9	homeobox A9
X54232_at	GPC1	glypican 1
X91247_at	TXNRD1	thioredoxin reductase 1
Y00503_at	KRT19	keratin 19
Y08999_at	ARPC1A	actin related protein 2/3 complex, subunit 1A, 41kDa
Z49989_at	SMTN	smoothelin
D13631_s_at	ARHGEF6	Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6
S69231_s_at	DCT	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
HG2815-HT2931_at	-	-
X70940_s_at	EEF1A2	eukaryotic translation elongation factor 1 alpha 2
Z19554_s_at	VIM	vimentin
M24766_s_at	COL4A2	collagen, type IV, alpha 2
M28213_s_at	RAB2A	RAB2A, member RAS oncogene family
M36653_s_at	POU2F2	POU class 2 homeobox 2
X57348_s_at	SFN	stratifin
U02566_s_at	TYRO3	TYRO3 protein tyrosine kinase
U28488_s_at	C3AR1	complement component 3a receptor 1

Table 3. The 28 Genes selected from GCM Data Set

Probe Set	Gene Symbol	Gene Title
K03195_at	SLC2A1	solute carrier family 2 (facilitated glucose transporter), member 1
M18728_at	CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)
M35252_at	TSPAN8	tetraspanin 8
M64099_at	GGTLA1	gamma-glutamyltransferase-like activity 1
U40434_at	MSLN	mesothelin
U48959_at	MYLK	myosin, light chain kinase
U60666_at	LRRC6	leucine rich repeat containing 6
U85193_at	NFIB	nuclear factor I/B
X04828_at	GNAI2	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2
X57766_at	MMP11	matrix metallopeptidase 11 (stromelysin 3)
X58079_at	S100A1	S100 calcium binding protein A1
X63187_at	WFDC2	WAP four-disulfide core domain 2
X71345_f_at	PRSS3	protease, serine, 3 (mesotrypsin)
Y00486_rna1_at	APRT	adenine phosphoribosyltransferase
D82373_at	SART1	squamous cell carcinoma antigen recognized by T cells
RC_AA022884_at	RAB6IP1	RAB6 interacting protein 1
RC_AA116036_at	TPX2	TPX2, microtubule-associated, homolog (Xenopus laevis)
RC_AA147646_s_at	METTL7A	methyltransferase like 7A
RC_AA148516_at	PRLR	Prolactin receptor
RC_AA164851_at	-	-
RC_AA227934_at	-	-
RC_AA233257_at	TGFB1	transforming growth factor beta 1 induced transcript 1
RC_AA284721_s_at	LEPRE1	leucine proline-enriched proteoglycan (leprecan) 1
RC_AA450351_at	-	Transcribed locus
RC_AA454581_at	PACS2	phosphofurin acidic cluster sorting protein 2
RC_AA458578_at	NEDD4L	neural precursor cell expressed, developmentally down-regulated 4-like
RC_AA488074_at	RGS5	regulator of G-protein signaling 5
U03891_at	APOBEC3B	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B

3.3. Comparison of Computational Results using Four Data Sets

Through the above comparisons, the SMO algorithm was selected as the classifier embedded in our algorithm. This SMO-based algorithm was then applied to the other two datasets: HCD174, and CNS. In the calculation process, we generally chose the following parameters: “ten-fold cross-validation”, “PolyKernel” kernel function and “standardization” data filter type, with the remaining parameters set to the default values. The results are shown in Table 4.

Table 4. Performance Comparison of Multi-class Classification on the Four Data Sets (%)

	NCI60		GCM		CNS		HCD174	
	Accuracy (%)	Selected genes	Accuracy (%)	Selected genes	Accuracy (%)	Selected genes	Accuracy (%)	Selected genes
SU[18]	85.37	13					92.0	1100
Pomeroy[19]					83.3	7129		
Yeang[21]			81.25	16063				
Peng [8]	87.93	27	85.19	26				
Lin [22]	95	15	84.3	48				
Xu[23]	84.66	79						
Cai [24]					85.7	45	97.3	80
Zhou[25]			83.28	400				
This study	**96.6**	**24**	**95.5**	**28**	**97.6**	**10**	**97.1**	**37**

3.4. Overfitting Evaluation

HCD174 dataset was partitioned into a training dataset of 142 samples and a test dataset of 32 samples. SSiCP was performed on the training set, and a classification model including 49 features was achieved with an accuracy of 95.8% by ten-fold cross-validation. Then the independent test dataset was used to test the classification model with an accuracy of 93.8%. The decrease of accuracies from 95.8% to 93.8% suggests that SSiCP avoids overfitting effectively.

4. Discussion

For the molecular classification of cancers, two issues must be addressed. The first focuses on achieving high classification accuracy in gene expression profiles; the second is to select a gene subset containing the fewest genes. We addressed these issues by adopting a new gene selection strategy based on step-by-step improvement of classification performance.

At the first step of gene selection, the optimal number of pre-selected genes should be determined. Through computational experiments, we concluded that higher classification accuracy can be achieved when more genes are selected. However, by increasing the number of features, the computational workload increases sharply. Generally, we chose dozens, or hundreds of features, thus higher classification accuracy was obtained, with a lower computing burden.

In addition, there is an important issue to be considered: choosing the most suitable classification algorithm which demonstrates the characteristic of no overfitting as well as achieving high classification accuracy. Having applied seven commonly used data mining algorithms to the benchmark NCI60 and GCM datasets, we concluded that SMO outperforms the other six algorithms (Table 1). Many studies show that the SVM classifier is one of the best algorithms, and demonstrates a strong ability to avoid overfitting [8, 26]. Therefore, using SMO (a fast SVM algorithm) as the classifier in SSiCP is an appropriate choice.

In the comparison of the results obtained from the four datasets, our algorithm was superior to all other algorithms in classification accuracy except for the algorithm of Cai et al., which achieved slightly higher accuracy than ours (97.3% versus 97.1%, Table 4), whereas the number of genes we selected was far less than theirs (80 versus 37, Table 4).

In addition, our algorithm had strong robustness. When using others, such as the genetic algorithm-based algorithms, the gene subsets selected in each computational experiment are not the same. Although higher classification accuracy may be achieved by using any one of the gene subsets, it may be difficult to determine which one is better when we need to use these gene subsets to do further biomedical experiments. Whereas by using our algorithm, the computational experiments can be repeated with exactly the same results: the same classification accuracy and the same selected gene subset.

The advantages of wrapper-based techniques for feature selection are well established [20]. So a comparison should be made between the wrapper-based approaches and SSiCP algorithm. First, it has recently been recognized that wrapper-based techniques have the potential to over-fit the training data [27], while SSiCP has shown the ability to overcome over-fitting by computational experiments. Second, wrapper-based techniques must employ a heuristic search method to search subset feature states in a large state space, making a heavy computational burden on the computer. However, instead of searching states in a huge space, SSiCP uses a step by step improvement of classification accuracy to reduce feature space, with a result of fast procedure of computation and simple implementation of the algorithm.

In the 24 genes selected from NCI60 dataset, at least ten genes were found to have direct evidence of the associations with cancers. Many selected genes were in the pathways of some cancers, including ARHGEF6 in the pathway of pancreatic cancer, COL4A2 in the pathways of cancer and pathway of small cell lung cancer, DCT in melanogenesis pathway, PLCB4 in Wnt signaling pathway, RRAS in the MAPK signaling pathway, and SFN in the Cell cycle and p53 signaling pathway. RRAS is an oncogene, which induced cell transformation in fibroblasts but not in other cell types. R-Ras also reportedly induces a more invasive phenotype in breast epithelial cells through integrin activation [28]. RAB2A is one member of RAS oncogene family [29]. Members of the Rab protein family are nontransforming monomeric GTP-binding proteins of the Ras superfamily and Rabs are prenylated, membrane-bound proteins involved in vesicular fusion and trafficking [30]. Evidence shows that ABLIM1 is related to cancers [31]. HOXA9 was reported to be associated with epithelial ovarian cancers [32]. It was observed that both cancers cell–derived and host-derived GPC1 are important for cancer growth, angiogenesis, and metastasis [33].

In the 28 genes selected from GCM data, at least nine genes were found to have direct evidence of the associations with cancers. CEACAM6 was reported to be widely used in tumor markers in serum immunoassay determinations of carcinoma. Wang et al. concluded that CEACAM6 can antagonize the Src signaling pathway, down-regulate cancer cell cytoskeleton proteins, and block adenovirus trafficking to the nucleus of human pancreatic cancer cells [34]. Yun et al. found that SLC2A1 (GLUT1) was 1 of 3 genes consistently up-regulated in cells with KRAS or BRAF mutations [35]. The TSPAN8 gene, encoding a cell surface glycoprotein defined by the monoclonal antibody CO-029, a 27- to 34-kD membrane protein, was reported to express in gastric, colon, rectal, and pancreatic carcinomas but not in most normal tissues [36]. Evidence shows that MSLN is related to ovarian cancer [37, 38]. Drapkin et al. concluded that WFDC2 (HE4) is overexpressed by serous and endometrioid ovarian carcinomas [39]. Hakoda et al. proposed the evidence of APRT gene with carcinogenesis [40]. TPX2 (C20orf2) is differentially expressed between cancerous and noncancerous lung cells [41]. NEDD4L gene plays a pivotal role in the prostate cancer [42]. Hamzah et al. reported that RGS5 is a major gene responsible for the aberrant morphology of tumor vasculature [43].

5. Conclusion

In summary, SSiCP algorithm outperforms many previous published algorithms for cancer gene selection. Many of the selected genes by SSiCP are shown to be cancer related genes, suggesting that SSiCP is an effective tool for the multiclass cancer classification based on gene expression profiles.

Acknowledgements

This study was supported by grants from the National Natural Science Foundation of China (Grant No. 61373057, 81071628), and a grant from the Zhejiang Provincial Natural Science Foundation of China (Grant No. Y1110763).

Disclosure

Part of information included in this chapter/article has been previously published in Computer Science and Information Engineering, 2009 WRI World Congress on March 31 2009-April 2 2009.

References

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, vol. 286, no. 5439, (**1999**), pp. 531-7.

[2] M. Bittner, P. Meitzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward and J. Trent, "Molecular classification of cutaneous malignant melanoma by gene expression profiling", Nature, vol. 406, no, 6795, (**2000**), pp. 536-40.

[3] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data", Bioinformatics vol. 16, no. 10, (**2000**), pp. 906-14.

[4] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. G. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. M. Yang, G. E. Marti, T. Moore, J. Hudson, L. S. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature, vol. 403, no. 6769, (**2000**), pp. 503-11.

[5] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. E. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines", Nature Genetics, vol. 24, no. 3, **(2000)**, pp. 227-35.

[6] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures", Proceedings of the National Academy of Sciences of the United States of America, vol. 98, no. 26, **(2001)**, pp. 15149-54.

[7] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebet, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub, "MicroRNA expression profiles classify human cancers", Nature, vol. 435, no. 7043, **(2005)**, pp. 834-8.

[8] S. H. Peng, Q. H. Xu, X. B. Ling, X. N. Peng, W. Du and L. B. Chen, "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines", Febs Letters, vol. 555, no. 2, **(2003)**, pp. 358-62.

[9] T. Li, C. L. Zhang and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", Bioinformatics, vol. 20, no. 15, **(2004)**, pp. 2429-37.

[10] S. H. Peng, X. M. Zeng, X. B. Li, X. N. Peng and L. B. Chen, "Multi-class cancer classification through gene expression profiles: microRNA versus mRNA", Journal of Genetics and Genomics, vol. 36, no. 7, (**2009**), pp. 409-16.

[11] H. Choi, D. Yeo, S. Kwon and Y. Kim, "Gene selection and prediction for cancer classification using support vector machines with a reject option", Computational Statistics and Data Analysis, vol. 55, no. 5, (**2011**), pp. 1897-908.

[12] D. G. Li, "Gene expression studies with DGL global optimization for the molecular classification of cancer", Soft Computing, vol. 15, no. 1, (**2011),** pp. 111-29.

[13] Y. Cun and H. Fröhlich, "Biomarker Gene Signature Discovery Integrating Network Knowledge", Biology, vol. 1, no. 1, (**2012**), pp. 5-17.

[14] D. A. Salazar, J. Ivan Velez and J. C. Salazar, "Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?", Revista Colombiana De Estadistica, vol. 35, no. 2, (**2012**), pp. 223-37.

[15] H. Y. Wang, H. Y. Zhang, Z. J. Dai, M. S. Chen and Z. M. Yuan, "TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection", Bmc Medical Genomics, vol. 6, (**2013**).

[16] X. Li, S. Peng, J. Chen, B. Lu, H. Zhang and M. Lai, "SVM-T-RFE: a novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles", Biochem Biophys Res Commun, vol. 419, no. 2, (**2012**), pp. 148-53.

[17] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", Machine Learning, vol. 46, nos. 1-3, (**2002**), pp. 389-422.

[18] A. I. Su, J. B.Welsh, L. M. Sapinoso, S. G. Kern, P. Dimitrov, H. Lapp, P. G. Schultz, S. M. Powell, C. A. Moskaluk, H. F. Frierson and G. M. Hampton, "Molecular classification of human carcinomas by use of gene expression signatures", Cancer Research, vol. 61, no. 20, (**2001**), pp. 7388-93.

[19] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression", Nature, vol. 415, no. 6870, (**2002**), pp. 436-42.

[20] R. Kohavi and G. H. John, "Wrapper for feature subset selection", Artificial Intelligence, vol. 97, (**1997**), pp. 273–324.

[21] C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov and T. Golub, "Molecular classification of multiple tumor types", Bioinformatics, vol. 17 Suppl 1, (**2001**), pp. S316-22.

[22] T. C. Lin, R. S. Liu, C. Y. Chen, Y. T. Chao and S. Y. Chen, "Pattern classification in DNA microarray data of multiple tumor types", Pattern Recognition, vol. 39, no. 12, (**2006**), pp. 2426-38.

[23] R. Xu, G. C. Anagnostopoulos and D. C. Wunsch, "Multiclass cancer classification using semi supervised ellipsoid ARTMAP and particle swarm optimization with gene expression data", Ieee-Acm Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, (**2007**), pp. 65-77.

[24] Z. Cai, R. Goebel, M. R. Salavatipour and G. Lin, "Selecting dissimilar genes for multi-class classification", an application in cancer subtyping, Bmc Bioinformatics, vol. 8, (**2007)**.

[25] X. Zhou and D. P. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data", Bioinformatics, vol. 23, no. 9, (**2007**), pp. 1106-14.

[26] V. N. Vapnik, "Statistical learning theory", New York: Wiley, (**1998**).

[27] J. Reunanen, "Overfitting in making comparisons between variable selection methods", Journal of Machine Learning Research, vol. 3, (**2003**), pp. 371-1382.

[28] H. Rincon-Arano, R. Rosales, N. Mora, A. Rodriguez-Castaneda and C. Rosales, "R-Ras promotes tumor growth of cervical epithelial cells", Cancer, vol. 97, no. 3, (**2003**), pp. 575-85.

[29] A. Chi, J. C. Valencia, Z. Z. Hu, H. Watabe, H. Yamaguchi, N. J. Mangini, H. Huang, V. A. Canfield, K. C. Cheng, F. Yang, R. Abe, S. Yamagishi, J. Shabanowitz, V. J. Hearing, C. Wu and E. Appella and D. F. Hunt, "Proteomic and bioinformatic characterization of the biogenesis and function of melanosomes", J Proteome Res, vol. 5, no. 11, (**2006**), pp. 3135-44.

[30] Entrenz gene: RAB2A RAB2A, member RAS oncogene family Gene ID: 5862.

[31] A. C. Kim, L. L. Peters, J. H. M. Knoll, C. VanHuffel, S. L. Ciciotte, P. W. Kleyn and A. H. Chishti, "Limatin (LIMAB1), an actin-binding LIM protein, maps to mouse chromosome 19 and human chromosome 10q25", a region frequently deleted in human cancers. Genomics, vol. 46, no. 2, (**1997**), pp. 291-3.

[32] W. J. Cheng, J. S. Liu, H. Yoshida, D. Rosen and H. Naora, "Lineage infidelity of epithelial ovarian cancers is controlled by HOX genes that specify regional identity in the reproductive tract", Nature Medicine, vol. 11, no. 5, (**2005**), pp. 531-7.

[33] T. Aikawa, C. A. Whipple, M. E. Lopez, J. Gunn, A. Young, A. D. Lander and M. Korc, "Glypican-1 modulates the angiogenic and metastatic potential of human and mouse cancer cells", Journal of Clinical Investigation, vol. 118, no. 1, (**2008**), pp. 89-99.

[34] Y. H. Wang, R. Gangeswaran, X. B. Zhao, P. J. Wang, J. Tysome, V. Bhakta, M. Yuan, C. P. Chikkanna-Gowda, G. Z. Jiang, D. L. Gao, F. Y. Cao, J. Francis, J. X. Yu, K. D. Liu, H. Y. Yang, Y. H. Zhang, W. D. Zang, C. Chelala, Z. M. Dong and N. Lemoine, "CEACAM6 attenuates adenovirus infection by antagonizing viral trafficking in cancer cells", Journal of Clinical Investigation, vol. 119, no. 6, (**2009**), pp. 1604-15.

[35] J. Y. Yun, C. Rago, I. Cheong, R. Pagliarini, P. Angenendt, H. Rajagopalan, K. Schmidt, J. K. V. Willson, S. Markowitz, S. B. Zhou, L. A. Diaz, V. E. Velculescu, C. Lengauer, K. W. Kinzler, B. Vogelstein and N. Papadopoulos, "Glucose Deprivation Contributes to the Development of KRAS Pathway Mutations in Tumor Cells", Science, vol. 325, no. 5947, (**2009**), pp. 1555-9.

[36] S. Szala, Y. Kasai, Z. Steplewski, U. Rodeck, H. Koprowski and A. J. Linnenbach, "Molecular-Cloning of Cdna for the Human Tumor-Associated Antigen Co-029 and Identification of Related Transmembrane Antigens", Proceedings of the National Academy of Sciences of the United States of America, vol. 87, no. 17, (**1990**), pp. 6833-7.

[37] K. Chang and I. Pastan, "Molecular cloning of mesothelin, a differentiation antigen present on mesothelium, mesotheliomas, and ovarian cancers", Proceedings of the National Academy of Sciences of the United States of America, vol. 93, no. 1, (**1996**), pp. 136-40.

[38] N. Scholler, N. Fu, Y. Yang, Z. M. Ye, G. E. Goodman, K. E. Hellstrom and I. Hellstrom, "Soluble member(s) of the mesothelin/megakaryocyte potentiating factor family are detectable in sera from patients with ovarian carcinoma, Proceedings of the National Academy of Sciences of the United States of America vol. 96, no. 20, (**1999**), pp. 11531-6.

[39] R. Drapkin, H. H. von Horsten, Y. F. Lin, S. C. Mok, C. P. Crum, W. R. Welch and J. L. Hecht, "Human epididymis protein 4 (HE4) is a secreted glycoprotein that is overexpressed by serous and endometriold ovarian carcinomas", Cancer Research, vol. 65, no. 6, (**2005**), pp. 2162-9.

[40] M. Hakoda, K. Nishioka and N. Kamatani, "Homozygous Deficiency at Autosomal Locus Aprt in Human Somatic-Cells Invivo Induced by 2 Different Mechanisms", Cancer Research, vol. 50, no. 6, (**1990**), pp. 1738-41.

[41] R. Manda, T. Kohno, Y. Matsuno, S. Takenoshita, H. Kuwano and J. Yokota, "Identification of genes (SPON2 and C20orf2) differentially expressed between cancerous and noncancerous lung cells by mRNA differential display", Genomics, vol. 61, no. 1, (**1999**), pp. 5-14.

[42] H. Qi, J. Grenier, A. Fournier and C. Labrie, "Androgens differentially regulate the expression of NEDD4L transcripts in LNCaP human prostate cancer cells. Molecular and Cellular Endocrinology, vol. 210, nos. 1-2, (**2003**), pp. 51-62.

[43] J. Hamzah, M. Jugold, F. Kiessling, P. Rigby, M. Manzur, H. H. Marti, T. Rabie, S. Kaden, H. J. Grone, G. J. Hammerling, B. Arnold and R. Ganss, "Vascular normalization in Rgs5-deficient tumors promotes immune destruction", Nature, vol. 453, no. 7193, (**2008**), pp. 410-U67.

Authors

Xiaobo Li, He received his B.Sc. in Microelectronics (1990) from Nankai University (China), Master of Engineering (Research) (2004) from The University of Sydney (Australia) and Ph.D. in Pathology and Pathophysiology (2012) from Zhejiang University (China). Now he is a full-time Associate Professor at Department of Computer Science and Technology, College of Engineering, Lishui University, China. His current research interests include different aspects of bioinformatics, machine learning and data mining.

Xue Gong, She received her Ph.D. in Pathology and Pathophysiology (2010) from Harbin Medical University (China). Now she is a Postdoctoral Research fellow at Department of Microbiology and Immunology, School of Medicine, Stanford University. Her current research interests include different aspects of cancer research and bioinformatics.

Xiaoning Peng, He received his M.Med. in Gastroenterology (1995) from University of South China (China), and M.D. in Infectious Diseases (2001) from Central South University (China). Now he is a full-time Professor at Department of Internal Medicine, School of Medicine, Hunan Normal University, China. His current research interests include different aspects of cancer research and bioinformatics.

Sihua Peng, He received his B.E. (1983) from Southeast University (China), Master of Engineering in Process Automation (1989) from Northeast Dianli University (China), and Ph.D. in Control Science and Engineering (2004) from Zhejiang University (China). Now he is a full-time Associate Professor at Department of Biological Technology, School of Fisheries and Life Science, Shanghai Ocean University, China. His current research interests include different aspects of bioinformatics, machine learning and data mining.