Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2013) 21: 2357 – 2366

doi:10.3906/elk-1203-138

Hybrid SPR algorithm to select predictive genes for effectual cancer classification

Aruna SUNDARAM,[1,*] **Nandakishore Lellapalli VENKATA,**[2]
Rajagopalan Sarukai PARTHASARATHY[3]

[1]Department of Computer Applications, Dr. M.G.R Educational and Research Institute University, Maduravoyal, Chennai, Tamil Nadu, India

[2]Department of Mathematics, Dr. M.G.R Educational and Research Institute University, Maduravoyal, Chennai, Tamil Nadu, India

[3]Professor Emeritus, Dr. M.G.R Educational and Research Institute University, Maduravoyal, Chennai, Tamil Nadu, India

Received: 31.03.2012 • **Accepted:** 04.07.2012 • **Published Online:** 30.10.2013 • **Printed:** 25.11.2013

Abstract: Designing an automated system for classifying DNA microarray data is an extremely challenging problem because of its high dimension and low amount of sample data. In this paper, a hybrid statistical pattern recognition algorithm is proposed to reduce the dimensionality and select the predictive genes for the classification of cancer. Colon cancer gene expression profiles having 62 samples of 2000 genes were used for the experiment. A gene subset of 6 highly informative genes was selected by the algorithm, which provided a classification accuracy of 93.5%.

Key words: Cancer classification, filters, wrappers, correlation feature selection, sequential backward search, support vector machines, DNA microarray

1. Introduction

Biomedical informatics is an emerging field applying information technologies in medical care. It is the science of using system-analytic tools to develop algorithms for the management, process control, decision-making, and scientific analysis of medical knowledge [1]. It merges the data knowledge and the necessary tools in the decision-making process. Its focus is mainly on algorithms that are needed to manipulate and acquire knowledge from the information. This makes biomedical informatics different from other medical disciplines. It has a wide variety of applications in health care. Clinical decision support systems (CDSSs) are among those recently developing in the medical domain. The CDSS was built based on 2 approaches, namely rule-based and machine-learning (ML) algorithms. ML-based systems are more preferred than interactive rule-based systems because they gain knowledge from the data. These systems are pervasive over a range of medical areas, such as cancer or dermatology.

Cancer is a group of diseases in which the cells in the body grow, change, and multiply out of control [2]. Cancer detection using DNA microarrays has become a significant area of research. The classifications of different tumor types are more important in diagnosis and drug discovery because only the malignant tumors are cancerous. Conventional cancer classification methods based on clinical methods are reported to have several limitations [3] in their diagnostic ability. Specifications of therapies according to tumor types differentiated by pathogenic patterns may maximize the efficacy of the patients [4–11]. The recent advent of microarray

*Correspondence: arunalellapalli@yahoo.com

technology has allowed the simultaneous monitoring of thousands of genes, which accelerated the development of cancer classification using gene expression data [12–14].

Successful microarray classification using ML algorithms is an extremely challenging task because of its high dimensionality (usually having thousands to tens of thousands of genes) and very small data set size (less than 100), and most genes are not related to cancer classification. Predictive accuracy is an important criterion for these algorithms. Biologists are more interested in classifiers not only giving high accuracy but also providing biological relevancy, which helps them to discover new drugs, achieve effective management of the disease, and reduce the toxicity among patients. Recent research [15] has shown that a small number of genes are enough for an accurate diagnosis of most diseases, even though the number of genes varies greatly between different diseases. These genes are the marker, predictive, candidate, or highly informative genes. Using a small set of genes for classification gives better diagnostic accuracy. Moreover, it provides an opportunity to analyze the nature of the disease further and the genetic mechanisms responsible for it.

Dimensionality reduction techniques play an important role in identifying predictive genes from gene expression data. Feature selection, which is one among them, selects a good subset of genes from the gene expression set. It follows 2 approaches for selecting a feature subset, namely neural networks and statistical pattern recognition (SPR) techniques [16]. SPR techniques yield optimum or suboptimum solutions based on their search criteria. SPR algorithms follow the filter approach if they select a gene subset based on a discriminating factor independent of the learning algorithm. SPR algorithms follow the wrapper approach if they use the learning algorithms for the subset selection. Filters have high computational competence. Wrappers give high predictive accuracy. To combine the advantages of both methods, hybrid algorithms are of recent research interest.

In this paper, we propose a hybrid SPR algorithm that combines a correlation feature selection (CFS) with ranking and sequential backward selection (SBS) using support vector machines (SVMs) to select predictive genes from gene expression data. A colon cancer microarray gene expression data set was used to experiment with the algorithm. The rest of this paper is structured as follows. Section 2 explains the hybrid SPR algorithm and Section 3 gives the results obtained. The concluding remarks are given in Section 4 to discuss further research issues.

2. Hybrid SPR algorithm

The hybrid SPR algorithm is a greedy algorithm. It follows the divide-and-conquer approach with the search criterion of a constrained search. The algorithm combines filters and wrappers to select candidate genes. CFS, with a ranking of the genes by the SVM, acts as a filter to remove redundant and irrelevant genes. SBS with SVM acts as a wrapper to select the set of genes with high predictive accuracy and biological relevancy. The aim of this algorithm is to achieve a minimum gene subset with highly informative genes. The stages in the algorithm are shown in Figure 1.

2.1. Algorithm

Input: Full training set F_T containing all of the genes.

Output: Feature subset Fs containing predictive genes.

Step 1: Select a subset F_C from F_T using the CFS method.

Step 2: Rank the genes in set F_C using SVM by calculating the square of the weights of the genes.

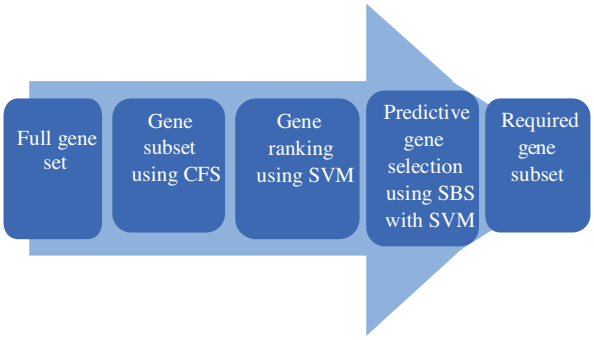


Figure 1. Stages in the hybrid SPR algorithm.

Step 3: Push the genes into a stack S according to their ranks in ascending order, such that the genes with the lowest rank will be in the top of the stack.

Step 4: Compute the classification accuracy A_g for the genes in S using SVM.

Step 5: (While n $\geq$ 3) do steps 6–8, where n is the total number of genes in S.

Step 6: POP n/3 genes from stack S and PUSH them into the stack T.

Step 7: Compute the classification accuracy A_l for the genes in S using SVM.

Step 8: If ($A_l < A_g$), then PUSH genes from T to S. Exit else, update n as n = number of genes in stack S, $A_g = A_l$, and empty stack T.

Step 9: End while.

Step 10: POP genes in S to set F_S.

Step 11: End.

Initially, the correlation coefficients for all of the genes are computed using Eq. (1), where C is the correlation between the summed feature subsets and the class variable, n is the number of subset features, k_{ac} is the average of the correlations between the subset features and the class variable, and k_{ai} is the average intercorrelation between the subset features. Gene subsets F_C having a high-class correlation and low intercorrelation are selected.

$$C = (nk_{ac})/(\sqrt{(n + n(n-1)k_{ai})}) \tag{1}$$

Next, the weight of every gene $w(t)$in F_C is calculated from Eq. (2), where $x_i(t)$ is the value of the tth gene of the ith sample:

$$W(t) = \sum_{SVs} y_i a_i x_i(t). \tag{2}$$

Genes are ranked by the square of the weight assigned by the SVM. The parameters used for ranking the genes using the SVM are C = 1.0, epsilon = 1.0E-25, and tolerance parameter = 1.0E-10. Finally, SBS with the SVM is used to select the candidate gene subset F_S from F_C.

The algorithm is an appropriate mix of filters and wrappers for selecting the required gene subset. There are redundant and irrelevant genes in the data set. CFS [17] selects genes having high class-correlation and low intercorrelation. Intercorrelated genes are redundant in nature. Hence, CFS acts as a redundancy filter to eliminate redundant genes. The selected gene subset still contains irrelevant genes. To identify them, gene ranking with SVM is done. Gene ranking acts as an irrelevancy filter. SBS with the SVM is then used to select the required predictive gene subset. SVM is a class of learning algorithms that are based on the principle of structural risk minimization (SRM) and have a number of advanced properties, including the ability to handle

a large feature space, effective avoidance of over fitting, and information condensing for the given data set [18]. SBS [19] starts with a full set of genes and sequentially removes the gene with the least importance to locate a better subset. SBS works best when the feature set has a large number of features. Only 1 feature is usually removed in SBS at each step. Since the data are high-dimensional, the SBS algorithm is modified to remove one-third of the genes at each step. Furthermore, SBS has no backtracking option. In this algorithm, in the final stage (Step 8), if the local accuracy becomes lower than the global accuracy, then backtracking is done to avoid eliminating useful genes. For the reason stated above, SBS with the SVM acts as an effective wrapper combination.

3. Results

Experiments were conducted in WEKA [20] with 10-fold cross-validation. Ten-fold cross-validation was proven to be statistically good enough in evaluating the performance of the classifier [21].

3.1. Data set description

The colon cancer data set [22] consisting of broad patterns of gene expressions revealed by a clustering analysis of tumor and normal colon epithelial tissues probed by the Affymetrix oligonucleotide array was used in the experiment. The cancer biopsies were collected from tumors, and the normal biopsies were collected from healthy parts of the colons of the same patients. Table 1 gives details of the colon cancer data set.

Table 1. Colon cancer data set.

Total no. of samples	No. of genes	Class-wise samples	
		Tumor	Normal
62	2000	40	22

3.2. Predictive gene selection

From 2000 genes, after removing the redundant genes using correlation coefficients, CFS selected a subset of 26 genes. These genes were ranked by the SVM and pushed into a stack in such a way the genes with the least rank will be on the top of the stack. Using SBS with the SVM, informative genes were selected from the gene subset in the stack. The main parameter values used for the SVM classifier are reported in Table 2.

Table 2. Parameter values used for the SVM classifier.

C	Epsilon	Exponent	Cache size	T	CV
1.0	1.0E-12	1.0	250,007	0.0010	10

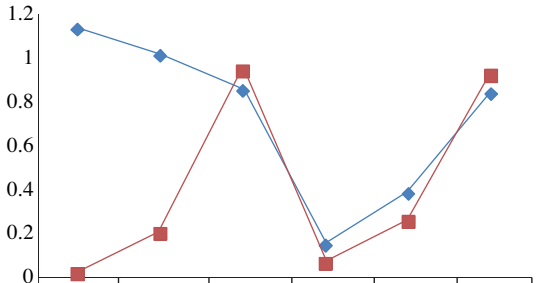
Table 3 shows the GenBank accession numbers (GANs) and the description of the predictive gene subset selected by the hybrid SPR algorithm having biological relevance to colon cancer.

Table 3. GANs and descriptions of the predictive genes selected by the hybrid SPR algorithm.

GAN	Description
M63391	*Homo sapiens* desmin gene
Z50753	*Homo sapiens* mRNA for GCAP-II/Uroguanylin precursor
M76378	*Homo sapiens* cysteine rich protein (CRP) gene, exons 5,6
J02854	Human 20-kDa myosin light chain mRNA
H08393	NiB *Homo sapiens* cDNA clone image: mRNA sequence
D14812	*Homo sapiens* KiAA0026 mRNA

A number of studies available in the literature have identified the genes selected by our algorithm as biomarkers relevant to colon cancer. For instance, the M63391 desmin gene, which encodes a muscle-specific class III intermediate filament, has been discovered to be downregulated in colon cancer [23]. This has been verified by biological experiments [24]. In adult striated muscle, they form a fibrous network connecting myofibrils to each other and the plasma membrane [25]. The Z50753, M76378, and J02854 genes are among the genes in a subset selected in [26] having a high classification accuracy. The association between the M76378 CRP gene and colon cancer was mentioned in [27]. The H08393 gene was found to be the biomarker for accessing colon cancer [28]. The M76378 and H08393 genes were selected as the informative genes for colon cancer in [29]. The D14812 gene was among the top-ranked genes in the experiments conducted by the authors in [30,31].

Figure 2 summarizes the storage space utilized by the data set: the model build time, accuracy, mean absolute and root mean squared error, and area under the receiver operating characteristic curve (AUC) for the colon cancer data set before and after the application of the hybrid SPR algorithm.



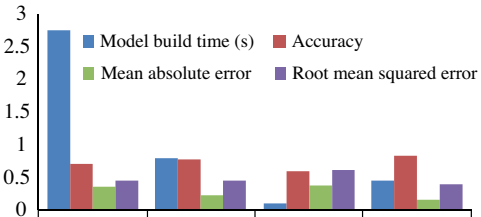
	Storage space (Mb)	Model build time (s)	Accuracy	Mean absolute error	Root mean squared error	AUC
All	1.13	1.01	0.85	0.1452	0.381	0.836
Six genes	0.016	0.2	0.94	0.0645	0.254	0.919

Figure 2. Empirical comparison before and after hybrid SPR algorithm.

Even though the SVM gives good classification accuracy with high-dimensional data, the computational cost, learning time, predictive ability, and memory requirements will be greater because of the curse of dimensionality. Its generalization ability can be improved by gene selection. This can be proven by the empirical comparison of the results in Figure 2. The results imply that storage space, model building time, mean absolute error and root mean squared error decrease, and the accuracy and AUC increase after gene subset selection by the hybrid SPR algorithm. Though 6 genes were selected from a group of 2000 genes, the small subset selected by our algorithm contains significant characteristics of the data domain. It yields a better classification performance and is biologically relevant to colon cancer.

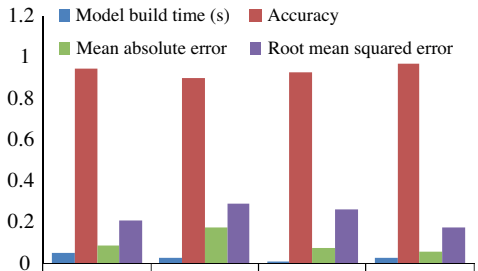
3.3. Cancer classification

The data-mining algorithms Simple Cart, radial basis function (RBF) network, naïve Bayes, and J48 were used to classify the colon cancer data set with all of the genes (2000) and with the optimum gene subset (6) selected by the proposed algorithm. Figure 3 shows the results for data-mining algorithms with 2000 genes. Figure 4 shows the results for data-mining algorithms with the 6 candidate genes selected by the hybrid SPR algorithm. The comparison of the results shows that the gene subset selected by the algorithm improves the predictive accuracy of all of the data-mining algorithms used for classification. The time taken to build the model and the error rate has been decreased to a great extent with the selected gene subset.



	Simple Cart	RBF network	Naïve bayes	J48
Model build time (s)	2.77	0.81	0.11	0.45
Accuracy	0.71	0.79	0.61	0.84
Mean absolute error	0.3692	0.2332	0.3871	0.1736
Root mean squared error	0.4531	0.4532	0.6222	0.3951

Figure 3. Colon data classification for 4 data-mining algorithms with 2000 genes.



	Simple Cart	RBF network	Naïve bayes	J48
Model build time (s)	0.05	0.03	0.01	0.03
Accuracy	0.95	0.9	0.93	0.97
Mean absolute error	0.0888	0.1738	0.0732	0.0614
Root mean squared error	0.2107	0.2955	0.2648	0.1753

Figure 4. Colon data classification for 4 data-mining algorithms with 6 genes.

3.4. Related work

Classification accuracies of the hybrid SPR algorithm and some studies from the literature of colon cancer data sets are summarized in Table 4. In [32], a locally linear embedding method was used for selecting informative genes. The authors selected 50 genes and obtained a classification accuracy of 85%. In [33], the evolutionary algorithm was used for selecting 50 candidate genes. The accuracy obtained was 75.8%. In [34], the authors used feature selection techniques, namely relief-F, which selected 4 genes, and CFS, which selected 26 genes, with a classification accuracy of 85.4% and 88.7%, respectively. In [35], for 6 predictive genes, an accuracy of 83.9% was obtained using random forest, and for a geometric representation-based classification algorithm, the accuracy was 87.1%. In [29], clustering and rough sets attribute reduction selected 6 genes. For the k-nearest neighbor (k-NN) algorithm, the accuracy was 79%, for naïve Bayes it was 82.2%, and for C5.0 it was 90.3%. In [36], the authors used hybrid gene selection algorithms and selected 3 marker genes yielding an accuracy of 92.0%. With feature selection using t-statistics, the authors selected 10 genes [37]. For the SVM-RBF method, the accuracy obtained was 85.4%.

In [38], the authors applied the breadth-first heuristic search algorithm based on a neighborhood rough set and obtained a gene subset of 6 genes with a classification accuracy of 85%. Ding and Peng [39] proposed a minimum redundancy-maximum relevancy (MRMR) method for gene selection. For 20 genes, they obtained a classification accuracy of 91.9%. In [40], the authors used normalized mutual information with the greedy method for gene selection. They presented an entropy-based iterative algorithm for selecting genes. They reported a classification accuracy of 91.9% for 9 genes. In [41], the authors used an adaptive neuro fuzzy

inference system (ANFIS) model for the classification of microarray data. They used information gain (IG) and signal-to-noise ratio (SNR) for gene selection. With these, they reported a classification accuracy of 93.5% and 90.3%, respectively, for 4 genes. In [42], using the based Bayes error filter (BBF) for the k-NN algorithm, for a gene subset of 12 genes, the authors obtained a classification accuracy of 90.3%.

Table 4. Performance comparison of our method with other approaches for colon data.

Algorithm	No. of genes	Accuracy (%)
Locally linear embedding, SVM-RBF [32]	50	65.0
Evolutionary algorithm [33]	50	75.8
Relief [34]	4	85.4
CFS [34]	26	88.7
Random forest [35]	6	83.9
Geometric representation-based classification algorithm [35]	6	87.1
Clustering and rough set attribute reduction k-NN [29]	6	79.0
Clustering and rough set attribute reduction naïve Bayes [29]	6	82.2
Clustering and rough set attribute reduction C5.0 [29]	6	90.3
HykGene rgk-NNs, SVMs, C4.5, naïve Bayes [36]	3	92.0
SVM-RBF [37]	10	85.4
Breadth-first heuristic search algorithm based on a neighborhood rough set [38]	6	85.0
MRMR [39]	20	91.9
Greedy [40]	9	91.9
ANFIS-SNR [41]	4	90.3
ANFIS-IG [41]	4	93.5
BBF [42]	12	90.3
Our method, the hybrid SPR algorithm	6	93.5

4. Conclusion

Cancer research is one of the major research areas in the medical field. The prediction of different tumor types has great value in providing better treatment and toxicity minimization for the patients. The role of microarray gene expressions in cancer diagnosis is very significant. Accurate classification from the DNA microarray is a difficult task because of its high dimensionality and low sample data. In this paper, a hybrid SPR gene selection algorithm has been proposed to reduce the dimensionality of the data set and select predictive genes of biological relevance for effective cancer classification. This algorithm is a combination of filters and wrappers. In this algorithm, a SVM is used to rank the subset of genes selected by the CFS method. SBS with the SVM is then used to select the predictive genes from the gene subset. Gene ranking by the SVM and CFS acts as a filter, and SBS with the SVM acts as a wrapper to select the optimal candidate genes. The algorithm was used in experiments on colon cancer gene data having 62 samples of 2000 genes. The algorithm yielded an informative gene subset of 6 genes and an accuracy of 93.5%. The data-mining algorithms Simple Cart, RBF network, naïve Bayes, and J48 were used to classify the colon cancer with marker genes selected by the algorithm. The gene subset improved the predictive accuracy of all of the classifiers. In this work, the algorithm was used in experiments on a colon cancer data set. In the future, this algorithm will be used in experiments on various gene expression data sets, which will provide a broader experimental evaluation and further improvement of the algorithm.

References

[1] H.E. Shortliffe, Biomedical Informatics: Computer Applications in Health Care and Biomedicine, Berlin, Springer, 2006.

[2] D. West, P. Mangiameli, R. Rampal, V. West, "Ensemble strategies for a medical diagnosis decision support system: a breast cancer diagnosis application", European Journal of Operation Research, Vol. 162, pp. 532–551, 2005.

[3] A. Azuaje, "Interpretation of genome expression patterns: computational challenges and opportunities", IEEE Engineering in Medicine and Biology, Vol. 19, p. 119, 2000.

[4] J. DeRisi, L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, J.M. Trent, "Use of a cDNA microarray to analyze gene expression patterns in human cancer", Natural Genetics, Vol. 4, pp. 457–460, 1996.

[5] W. Dubitzky, M. Granzow, D. Berrar, S. Bulashevska, C. Conrad, D. Gerlich, R. Eils, "Comparing symbolic and subsymbolic machine learning approaches to classification of cancer and gene identification", Methods of Microarray Data Analysis, pp. 151–165, 2002.

[6] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, Vol. 286, pp. 531–537, 1999.

[7] S.L. Pomeroy, P. Tamayo, M. Gassenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, Y.H. Kim, L.C. Goumnerova, M. Black, C. Lau, 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, T.R. Golub, "Prediction of central nervous embryonal tumor outcome based on gene expression", Nature, Vol. 415, pp. 436–442, 2002.

[8] T. Sørile, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P.E. Lønning, A.L. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclass with clinical implications", Proceedings of the National Academy of Science of the USA, Vol. 98, pp. 10869–10874, 2001.

[9] L.J. Van't Veer, D. De Jong, "The microarray way to tailored cancer treatment", Nature Medicine, Vol. 8, pp. 13–14, 2002.

[10] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer", Nature, Vol. 415, pp. 530–536, 2002.

[11] D.A. Zajchowski, M.F. Bartholdi, Y. Gong, L. Webster, H.L. Liu, A. Munishkin, C. Beauheim, S. Harvey, S.P. Ethier, P.H. Johnson, "Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells", Cancer Research, Vol. 61, pp. 5168–5178, 2001.

[12] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, W.C. Chan, D. Botstein, P.O. Brown, "Gene shaving as a method for identifying distinct sets of genes with similar expression patterns", Genome Biology, Vol. 1, pp. 1–21, 2000.

[13] S. Lakhani, A. Ashworth, "Microarray and histopathological analysis of tumors: the future and the past?", Nature Reviews Cancer, Vol. 1, pp. 151–157, 2001.

[14] D. Slonim, P. Tamayo, J. Mesirov, T. Golub, E. Lander, "Class prediction and discovery using gene expression data", Proceedings of the 4th International Conference on Computational Molecular Biology, pp. 263–272, 2000.

[15] Y. Saeys, I. Inza, P. Larranaga, "A review of feature selection techniques in Bioinformatics", Bioinformatics, Vol. 23, pp. 2507–2517, 2007.

[16] D. Zongker, A. Jain, "Algorithms for feature selection: an evaluation", Proceedings of the 13th International Conference on Pattern Recognition, Vol. 2, pp. 18–22, 1996.

[17] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection", Journal of Machine Learning Research, Vol. 3, pp. 1157–1182, 2003.

[18] A.A. Goshtasby, Image Registration: Principles, Tools and Methods, Berlin, Springer, 2012.

[19] N.V. Vapnik, Statistical Learning Theory, New York, Wiley, 1998.

[20] Weka: A multi-task machine learning software. Available at http://www.cs.waikato.ac.nz/ml/weka.

[21] P. Baldi, S. Brunak, Y. Chauvin, F. Anderson, H. Nielsen, "Assessing the accuracy of prediction algorithms for classification and overview", Bioinformatics, Vol. 16, pp. 412–424, 2000.

[22] National Center for Biotechnology Information, Colon Cancer Data, U.S. National Library of Medicine, available at http://www.ncbi.nlm.nih.gov.

[23] G.M. Groisman, S. Polak-Charcon, H.D. Appelman, "Fibroblastic polyp of the colon: clinicopathological analysis of 10 cases with emphasis on its common association with serrated crypts", Histopathology, Vol. 48, pp. 431–437, 2006.

[24] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Cell Biology, Vol. 95, pp. 6745–6750, 1999.

[25] GeneCards, DES Gene GeneCards - DESM Protein - DESM Antibody, available at http://www.genecards.org/cgi-bin/carddisp.pl?gene=DES.

[26] Z. Chen, J. Li, L. Wei, "A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue", Artificial Intelligence in Medicine, Vol. 41, pp. 161–175, 2007.

[27] P. Mahata, K. Mahata, "Selecting differentially expressed genes using minimum probability of classification error", Journal of Biomedical Informatics, Vol. 40, pp. 775–786, 2007.

[28] S. Barnhill, I. Guyon, J. Weston, US Patent US20050165556 - Colon Cancer Biomarkers, 2005. Available at http://www.google.com/patents/US20050165556.

[29] L. Sun, D. Miao, H. Zhang, "Gene selection with rough sets for cancer classification", 4th International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 3, pp. 167–172, 2007.

[30] J. Li, X. Tang, J. Liu, J. Huang, Y. Wang, "A novel approach to feature extraction from classification models based on information gene pairs", Pattern Recognition, Vol. 41, pp. 1975–1984, 2008.

[31] X. Li, S. Rao, T. Zhang, Z. Guo, Q. Zhang, K. Moser, E. Topol, "An ensemble method for gene discovery based on DNA microarray data", Science in China Series C, Vol. 47, pp. 396–405, 2004.

[32] C. Shi, L. Chen, "Feature dimension reduction for microarray data analysis using locally linear embedding", Asia Pacific Bioinformatics Conference, pp. 211–217, 2005.

[33] T.J. Umpai, S. Aitken, "Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes", Bioinformatics, Vol. 6, pp. 168–174, 2005.

[34] L. Yu, H. Liu, "Redundancy based feature selection for microarray data", Department of Computer Science and Engineering of Arizona State University, Technical Report, 2004.

[35] S. Kim, "Spectral methods for cancer classification using microarray data", International Conference on Computational Sciences and Optimization, Vol. 1, pp. 588–592, 2009.

[36] Y. Wang, F.S. Makedon, J.C. Ford, J. Pearlman, "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data", Bioinformatics, Vol. 21, pp. 1530–1537, 2005.

[37] S.M. Alladi, P.S. Santosh, V. Ravi, U.S. Murthy, "Colon cancer prediction with genetic profiles using intelligent techniques", Bioinformation, Vol. 3, pp. 130–133, 2008.

[38] M.L. Hou, S.L. Wang, X.L. Li, Y.K. Lei, "Neighborhood rough set reduction based gene selection and prioritization for gene expression profile analysis and molecular cancer classification", Journal of Biomedicine and Biotechnology, Vol. 2010, pp. 1–12, 2010.

[39] C. Ding, H. Peng, "Minimum redundancy feature selection from microarray gene expression data", Proceedings of Computational Systems Bioinformatics, pp. 185–205, 2003.

[40] X. Liu, A. Krishnan, A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data", BMC Bioinformatics, Vol. 6, p. 76, 2005.

[41] Z. Wang, V. Palade, Y. Xu, "Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis", Proceedings of the 2nd International Symposium on Evolving Fuzzy System, pp. 241–246, 2006.

[42] G. Zhang, H.W. Deng, "Gene selection for classification of microarray data based on the Bayes error", BMC Bioinformatics, Vol. 8, p. 370, 2007.