Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ROBUST FEATURE SELECTION FROM MICRO ARRAY DATA USING LINEAR KERNEL SVM-RFE ALLIED WITH BOOTSTRAPPING

[1]G.VICTO SUDHA GEORGE, [2]Dr. V.CYRIL RAJ
[1] Research Scholar and Asst.Professor, Department Of Computer Science and Engineering,
[2] Addnl. Dean Engineering and Technology,
Dr.M.G.R Educational and Research Institute University,
Chennai-600095, Tamil Nadu, India
Email: [1]sudhajose72@gmail.com

ABSTRACT

In biomedical applications of computational biology, feature selection is considered to be an important issue as succeeding biological validations may be influenced by the robustness of the selected features significantly. More over a robust set of features may strengthen an expert's confidence in the result of a selection method. But the stability or robustness of the selection algorithms has received attention lately.

Though SVM-RFE is proved to be one of the most successful feature selection methods, it selects features using the gene expression data only without using any other biological information of the genes. Hence there is no guarantee that the selected features will have biological relevance. To overcome this issue, a framework for reliable feature selection has been proposed in this paper which illustrates that bootstrapping and leave-one-out cross-validation when used along with SVMRFE, increases its robustness significantly .

Keywords: *Feature Selection; Biomarker Detection; Stability; Machine Learning, Bootstrapping.*

1. INTRODUCTION

In accordance with machine learning, biomarker selection is known as feature selection. The Feature selection is used for finding a small set of features (markers) that best explains the difference between the diseased and the control samples[1][2]. The benefits of feature selection algorithms are aiding the construction of powerful classification models by eliminating irrelevant or noisy features [3], reducing the time complexity of the classification models and the ability to concentrate on a subset of relevant features, which intern can be used for finding new knowledge [4]. Further it can also bring down the cost of future clinical tests.

Based on how they interact with the estimation of the classification model, feature selection techniques are grouped into three classes[5][6]. Filter methods work independently of the classifier design, and carry out feature selection by seeing the intrinsic properties of the data [7][8][9].On the other hand, wrapper and embedded methods perform feature selection by using a specific classification model. While wrapper methods perform a search approach in the space of probable feature subsets, guided by the predictive performance of a classification model [10][11][12], internal parameters of the classification model is used by the embedded methods for feature selection[13][14][15]. In case of high-dimensional spaces embedded methods usually show a better computational complexity than wrapper methods.

One of the problems with existing feature selection algorithms is the instability of the selected feature subset in the successive runs (for same feature selection procedure or different procedure even for the same data) that can achieve the same or similar predictive accuracy [16][16][18][19]. In fact high reproducibility of feature selection and high classification accuracy should be given equal importance [20].It is widely believed that a study that cannot be repeated has little value [21]. Ultimately, the instability of feature selection results will reduce the confidence in discovered markers.

The stability or robustness of the selection algorithms with respect to sampling variation have received attention very recently. In this paper therefore, a general framework has been proposed for reliable feature selection. Since the SVM RFE feature selection algorithm has outperformed most of its counterparts in the field of cancer classification it is used as the baseline.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1.1 Reasons for Instability

There are three reasons for instability of the selected features. Firstly, Algorithm design without considering stability. Traditional feature selection methods focus on selecting a minimum subset of features to construct a classifier of the best predictive accuracy [16]. Secondly, Existence of multiple sets of true markers: The real data set may have many sets of potential true markers, and different ones may be selected under different settings [16][22]. Thirdly, Small number of samples in high dimensional data: In the analysis of gene expression data and proteomics data, there are typically only hundreds of samples but thousands of features. It has been experimentally verified that the relatively small number of samples in high dimensional data is one of the main
sources of the instability problem in feature selection[22][23]. To understand the nature of the instability of selected feature subset, Kim has developed a new mathematical model and concluded that at least thousands of samples are needed to achieve stable feature selection [24].

In this paper we are addressing the third cause of instability i.e., small number of samples in high dimensional data. We have used n-CV technique to generate more number of sample from the existing small number of samples. Each time few samples are randomly removed from the original data set which can be used for testing while the remaining samples are used to train SVM-RFE to generate the gene ranking list. The above process is repeated 100 times. Finally the genes that have occupied the top 10 positions in more than 85 gene ranking lists are claimed as the top ranking genes. We have also compared the result of our proposed method with the standard method called GEO2R.

1.2 Organization of the Paper

The remainder of this paper is organized as follows. In Section 2, we have discussed about the data set used in this study, the data preprocessing stage which convert the raw data to required format, the initial irrelevant genes elimination stage that reduce the complexity of further computation, the discussion about working SVM-RFE algorithm and finally we gave given our proposed frame work and the re-ranking strategy adopted to generate more stable feature set . In section 3, we have discussed about the results generated by our proposed method. To show the biological relevance of the genes selected by our method from literature we have given proof and from the gene Ontology Consortium we have shown the biological process and the functions of the top 10 genes selected. Apart from that we have also compared the result of our method with a standard method GEO2R and shown that seven out of the ten genes selected by method are the same. Finally, section 4 concludes with a discussion of the contributions of our proposed work.

2. MATERIALS AND METHODS

2.1 Data Source

The global expression profiling of 29 samples (GSM241999 to GSM242027) with dataset ID as GSE9574 is obtained from GEO (http://www.ncbi.nlm.nih.gov/geo/). The samples are taken from histologically normal micro dissected breast epithelium. Out of the 29 samples 14 samples are from epithelium adjacent to a breast tumor and 15 samples were obtained from patients undergoing reduction mammoplasty without apparent breast cancer. The Affymetrix platform GPL96 [HG-U133A] Affymetric Human Genome U133A Array has been used to generate this dataset. The description of the samples is given in table 1.

2.2 Data Preprocessing

The microarray data is normalized using RMA (Robust Multi-chip Averages), a quintile-based (**Irizarry** et al, 2003) method. RMA is an algorithm used to create an expression matrix from Affymetrix data. The raw intensity values are background corrected, log2 transformed and then quantile normalized. Next a linear model is fit to the normalized data to obtain an expression measure for each probe set on each array. We conducted the RMA analysis using Bioconductor, which is an open-source tool for bioinformatics using the R statistical programming language.

2.3 Elimination of the irrelevant Genes

The normalized microarray data contains the expression value of all the genes (in ten thousands). So the next step is to filter out the irrelevant genes which are not responsible for breast cancer. Here we have used T-test with a P value of 0.001 and 500 genes passed out the test. Still it is numerous for biomarker discovery. Hence in this paper a stable feature selection algorithm is proposed to select few numbers of important features.

Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.4 SVM-RFE for Feature Selection

RFE (Recursive Feature Elimination) is an iterative procedure of SVM classifier. A cost function J computed on training samples is used as an objective function. Expanding J in Taylor series to the second order using the OBD algorithm [26], and neglecting the first order term at the optimum of J, yields

$$DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2$$

Here $(w_i)^2$ was used as the ranking criterion and we used LIBSVM (Library for Support Vector Machines) [27] with a linear kernel.

Though SVMRFE has outperformed many other feature selection algorithms in cancer classification domain there is no surety that all the resultant features will be biologically relevant, the reason being that RFE is purely a statistical method. More over the main problem of micro array data is Curse of Dimentionality (Small number samples and more number of genes),which brings out instability in feature selection. To address these two issues we have used bootstrapping method along with SVMRFE to improve the stability of the selected features.

2.5 Stable Feature Selection Frame Work

The problem with micro array data is lack of experimental samples that can be solved by using bootstrapping a resembling technique. So bootstrapping along with SVM RFE is used here to select stable features (genes). In order to make good use of limited data for predicting breast cancer biomarkers, the generation of training set is a key factor. For this intention, the dataset was split into n subsets of approximately equal size at random, then one subset was eliminated, and the left behind samples formed the training set. Every time a different subset was chosen so that all the samples had an equal chance to be elected as the training data. We call it n-CV (see Figure 1) , where n could be equal to 12, 6, 4, or 3, respectively.

In the selected dataset, there are 29 samples described in Table 1, and a 10-CV is used for each subset with three samples. Then the Leave One out Cross Validation (LOOCV) is used for training SVM (see Figure 2). After that we performed 100 times of 10-CV and each 10-CV ran 10 times of the SVM-RFE procedure.

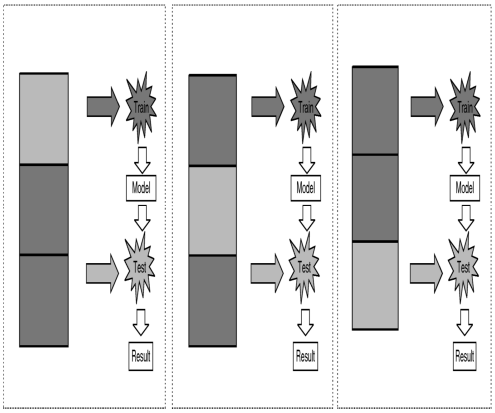

Figure 1. Procedure of three-fold cross-validation

Table1. Sample Description (Noncancerous (1) and Cancerous (0))

Sl No.	Accession No.	Title	Class Label
1	GSM241999	Reduction mammoplasty patient	1
2	GSM242000	Reduction mammoplasty patient	1
3	GSM242001	Reduction mammoplasty patient	1
4	GSM242002	Reduction mammoplasty patient	1
5	GSM242003	Reduction mammoplasty patient	1
6	GSM242004	Reduction mammoplasty patient	1
7	GSM242005	Reduction mammoplasty patient	1
8	GSM242006	Reduction mammoplasty patient	1
9	GSM242007	Reduction mammoplasty patient	1
10	GSM242008	Reduction mammoplasty patient	1
11	GSM242009	Reduction mammoplasty patient	1
12	GSM242010	Reduction mammoplasty patient	1
13	GSM242011	Reduction mammoplasty patient	1
14	GSM242012	Reduction mammoplasty patient	1
15	GSM242013	Reduction mammoplasty patient	1
16	GSM242014	Breast cancer patient	0
17	GSM242015	Breast cancer patient	0
18	GSM242016	Breast cancer patient	0
19	GSM242017	Breast cancer patient	0
20	GSM242018	Breast cancer patient	0
21	GSM242019	Breast cancer patient	0
22	GSM242020	Breast cancer patient	0
23	GSM242021	Breast cancer patient	0
24	GSM242022	Breast cancer patient	0
25	GSM242023	Breast cancer patient	0
26	GSM242024	Breast cancer patient	0
27	GSM242025	Breast cancer patient	0
28	GSM242026	Breast cancer patient	0
29	GSM242027	Breast cancer patient	0

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

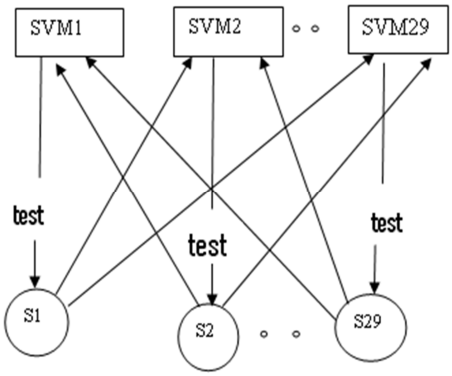

Figure 2. LOOCV for twenty nine samples.

2.6 The Re-ranking Strategy

The input for this step is 100 ranked lists of the 500 genes. Re-ranking measure is framed to form a final ranking of all genes by taking into account the occurrence and original ranking of every gene .The Steps to be followed are given as below.

Step1. Find the number of occurrences of each gene (Og) in the top 100 position.

Step2. Select the genes where Og>= 85

Step3. Calculate the ranking of the selected genes

$$W_g = 1 - (Rankg-1) * 0.002 \quad (1)$$

Step4. Calculate the overall weight of each gene

$$OW_g = \sum_{N=1}^{k} W_g \quad 1 \leq N \leq K \text{ Where } O_g \geq 85 \quad (2)$$

Step5. Based on OW_g select the Top 10 genes as the Biomarker genes.

Here $Rank_g$ is the rank of the gene in N^{th} iteration K=100, OW_g is over all Weight of a gene and W_g is Weight of the gene in the N^{th} iteration.

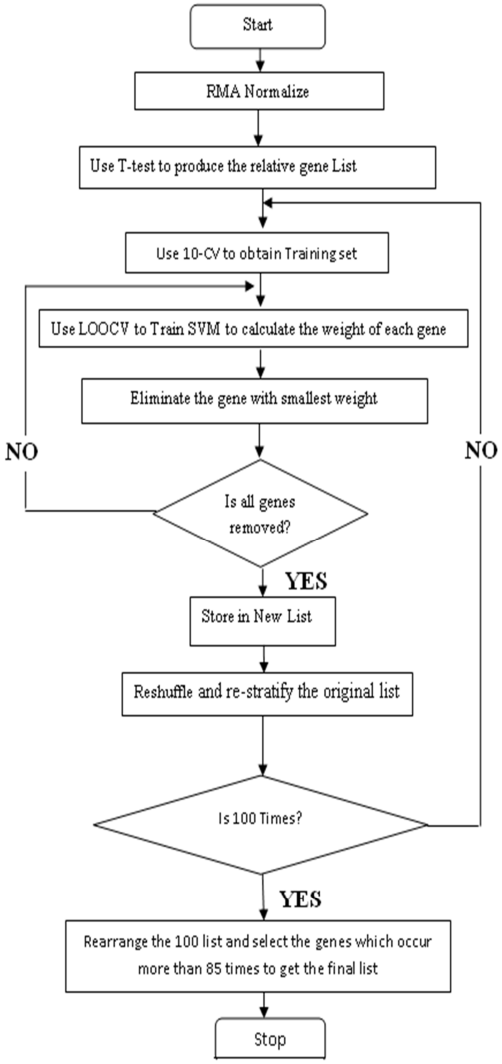

Figure 3. Flowchart of Stable Feature Selection

Figure 3 describes the overall flow of the stable feature selection algorithm. Since SVMRFE is a statistical method there is a possibility for a irrelevant gene to be selected as a top ranked gene . To overcome this issue we have considers only the genes which have occurred in the top 100 list for at least 85 times of the iterations.

3. RESULTS AND DISCUSSION

Figure 4 shows the occurrence of these genes in top-10 list when conducting 100 times of 10-CV. It illustrates that the occurrence of these genes in the top list is very high and consistent. We checked the functional annotations using GO (http://www.geneontology.org/) and the literatures and obtained related information shown in Table 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

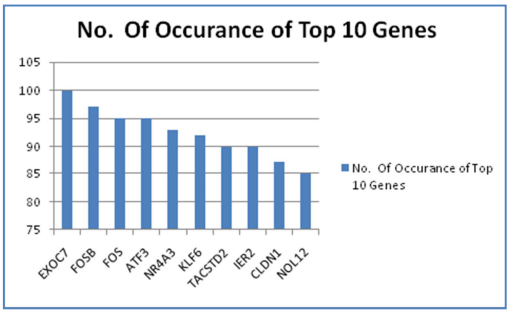

Figure 4. The occurrence the top-10 genes in 100 times of 10-CV

3.1 Literature proof of the Top 10 Genes

For the top 10 genes selected, we referred the literature to prove that the selected genes play vital role in progression of breast cancer. Out of the ten genes selected by our method, we have found that seven genes have strong biological relevance in the initiation and progression of the disease. A multigene HRneg/ Tneg signature linked to immune/ inflammatory cytokine regulation was identified from pooled expression microarray data in [28] and they have identified a novel set of 14 prognostic gene candida and EXOC 7 is proved to be one of them. A study was made about FosB protein in [29] and they have found a strong nuclear FosB immunoreactivity in epithelial cells of normal lobules and ducts, and carcinomas frequently showed loss of FosB expression or weak immunostaining.

To investigate the role of Fos family members in regulating breast cancer cell growth, antisense experiments have been performed in [30] and demonstrated that cFos is a critical regulator of breast cancer cell growth. The expression of ATF3 by immunoblot in human breast tumors is being examined to study the potential relevance of ATF3 in human breast cancer and it is shown that ARF3 plays an oncogenic role in malignant breast carcinoma cells [31].To identify the role of KLF6 inhibition in breast cancer development, a study has been carried out [32]and it has been found that KLF6 mediates cell growth in ERalpha-positive breast cancer cells through interaction with the c-Src protein. It has been proved [33]that the expression of TACSTD2 is turned OFF in most normal samples but ON in almost all of the breast cancer samples independent of the subtypes. IT is demonstrated in [34] that the significant loss of CLDN1 protein in breast cancer cells may play a role in invasion and metastasis. The loss of CLDN4 expression in areas of apocrine metaplasia and in the majority of grade 1 invasive carcinomas also suggests a particular role for this protein in mammary glandular cell differentiation and carcinogenesis.

3.2 Comparison of the Result with GEO2R Results

GEO2R is an interactive web tool that allows users to compare two or more groups of samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions. In Table 2 we have compared the top 10 genes selected by our proposed method and the result given by the standard tool GEO2R.It is observed that seven of the genes FOSB, FOS, ATF3, ,NR4A3,TACSTD2, IER2,NOL12 selected by our method are same as that of by GEO2R.From the literature it is shown these genes play major role in the progression of Breast Cancer. In the literature proof the functionality of these genes are clearly discussed. It is noticed that the gene occupying the top most position is not the same and the positions of these genes in the two results are also not the same.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2. Comparison of the result of our method with Output of GEO2R

Rank	Top 10 Genes Selected By our Method		Top 10 Genes Selected using GEO2R	
	Symbol	**Name**	**Symbol**	**Name**
1	EXOC7	exocyst complex component 7	PTP4A3	protein tyrosine phosphatase type IVA, member 3
2	FOSB	FBJ murine osteosarcoma viral oncogene homolog B	IER2	immediate early response 2
3	FOS	FBJ murine osteosarcoma viral oncogene homolog	FOSB	FBJ murine osteosarcoma viral oncogene homolog B
4	ATF3	activating transcription factor 3	NR4A3	nuclear receptor subfamily 4, group A, member 3
5	NR4A3	nuclear receptor subfamily 4, group A, member 2	ATF3	activating transcription factor 3
6	KLF6	Kruppel-like factor 6	BTG2	BTG family, member 2
7	TACSTD2	tumor-associated calcium signal transducer 2	NOL12	nucleolar protein 12
8	IER2	immediate early response 2	FOS	FBJ murine osteosarcoma viral oncogene homolog
9	CLDN1	claudin 1	TACSTD2	tumor-associated calcium signal transducer 2
10	NOL12	nucleolar protein 12	H3F3B	H3 histone, family 3B (H3.3B)

Table 3. Gene Ontology related details of the Top 10 Genes

Rank	Gene Name	Description	GO Function	Go Process	GO Component
1	EXOC7	Exocyst complex component 7	Docking of exocytic vesicles with fusion sites on the plasma membrane.	Exocytosis Protein transport	Cytoplasm Membrane Cell membrane
2	FOSB	FBJ murine osteosarcoma viral oncogene homolog B	sequence-specific DNA binding, ,transcription factor binding	negative regulation of transcription from RNA polymerase II promoter	Nucleus
3	FOS	FBJ murine osteosarcoma viral oncogene homolog	double-stranded DNA binding, transcription factor activity,	female pregnancy, positive regulation of transcription from RNA polymerase II promoter,	cytosol, ,transcription factor complex
4	ATF3	activating transcription factor 3	identical protein binding, transcription factor activity.	negative regulation of transcription,	nucleolus,nucleoplasm ,nucleus
5	NR4A3	nuclear receptor subfamily 4, group A, member 3	DNA binding, transcription factor activity involved in positive regulation of transcription	gene expression, positive regulation of cell cycle,	nucleoplasm,nucleus,transcription factor complex
6	KLF6	Kruppel-like factor 6	DNA binding, metal ion binding	Transcription,Transcription regulation	
7	TACSTD2	tumor-associated calcium signal transducer 2	receptor activity	negative regulation of cell motility, positive regulation of stem cell differentiation,	cytosol,extracellular space,lateral plasma membrane,nucleus
8	IER2	immediate early response 2			
9	CLDN1	claudin 1	identical protein binding ,structural molecule activity	calcium-independent cell-cell adhesion, cell-cell junction organization	Cell junction Cell membrane
10	NOL12	nucleolar protein 12	rRNA binding		

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. CONCLUSION

A common problem with existing feature selection methods is that the selected genes by the same method often vary with some variations of the samples in the same training set. The instability of the gene signatures bring out uncertainties about the reliability of the selected genes as biomarker candidates and thwart biologists from deciding candidates for further validations. Though there are many reasons for the insatiability of the selection process, we have addressed the main problem faced by Micro Array Data, small number of samples and large number of genes. To address this issue in this paper a robust feature selection frame work has been proposed. For the breast cancer dataset from literature and gene Ontology it is proved that 7 out of the top 10 genes selected by the proposed method have very good biological relevance for the progression of Breast Cancer.

4.1Limitations and Future work

To improve the stability of the selected features, in the proposed work we have iterated the feature selection process 100 times which took considerable computation time. More over the proposed method could able to find only seven out of the top 10 genes with good biological relevance.

In future we are planning to formulate still stronger re-ranking strategy to reduce the computation time and to improve the accuracy and stability of the selected features.

REFERENCES

[1] M. Hilario and A. Kalousis, "Approaches to Dimensionality Reduction In Proteomic Biomarker Studies", Briefings in Bioinformatics, vol. 9, no. 2, 2008, pp.102–118.

[2] F. Azuaje, Y. Devaux, and D. Wagner, "Computational Biology for Cardiovascular Biomarker Discovery", Briefings in Bioinformatics, vol. 10, no.4, 2009, pp. 367–377.

[3] Krishnapuram,B. et al. "Gene Expression Analysis: Joint Feature Selection and Classifier Design". In B.Schölkopf, et al. (eds) Kernel Methods in Computational Biology. MIT Press, Cambridge, MA, 2004, pp. 299–317.

[4] Guyon,I. and Elisseeff,A. "An Introduction to Variable and Feature Selection". J. Mach. Learn. Res., 3, 2003, 1157–1182.

[5] Saeys,Y. et al. "A Review of Feature Selection Techniques in Bioinformatics". Bioinformatics, 23, 2007, 2507–2517.

[6] G. Bontempi, "A Blocking Strategy to Improve Gene Selection for Classification of Gene Expression Data", IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 2, Apr.-June 2007, pp. 293-300.

[7] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression", Proc. Nat'l Academy of Sciences USA, vol. 99, 2002, pp. 6567-6572.

[8] J. Devore and R. Peck, "Statistics: The Exploration and Analysis of Data", third ed. Duxbury Press, 1997.

[9] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", J. Am. Statistical Assoc., vol. 97, 2002, pp. 77-87.

[10] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection", Artificial Intelligence, vol. 97, nos. 1/2, 1997, pp.273-324.

[11] C. Huang and C. Wang, "A GA-Based Feature Selection and Parameter Optimization for Support Vector Machines", Expert Systems with Applications, vol. 31, 2006, pp. 231-240.

[12] Y. Tang, Y.-Q. Zhang, and Z. Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis", IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 3, July-Sept. 2007, pp. 365-381.

[13] G. Fung and O.L. Mangasarian, "Data Selection for Support Vector Machine Classifiers", Proc. ACM SIGKDD, 2000, pp.64-70.

[14] L. Shen and E.C. Tan, "Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data", IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 2, no. 2, Apr.-June 2005, pp. 166-175.

[15] B. Krishnapuram, L. Carin, and A.J. Hartemink, "Joint Classifier and Feature Optimization for Cancer Diagnosis Using Gene Expression Data", Proc. Seventh Ann. Int'l. Conf. Computational Molecular Biology, 2003.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[16] L. Yu, C. Ding, and S. Loscalzo, "Stable Feature Selection Via Dense Feature Groups", in Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'08), 2008, pp. 803–811.

[17] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome Signature Genes in Breast Cancer: Isthere a Unique Set?" Bioinformatics, vol. 21, no. 2, 2005, pp. 171–178.

[18] S.Michiels, S. Koscielny, and C. Hill, "Prediction of Cancer Outcome with Microarrays: a Multiple Random Validation Strategy", Lancet, vol. 365, no. 9458, 2005, pp. 488–492.

[19] M. Zucknick, S. Richardson, and E. A. Stronach, "Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods", Statistical Applications in Genetics and Molecular Biology, vol. 7, no. 1, 2008, p. 7.

[20] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, "Algebraic Stability Indicators for Ranked Lists in Molecular Profiling", Bioinformatics, vol. 24, no. 2, 2008, pp. 258–264.

[21] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D.Wang, C.Wang, and Z. Guo, "Evaluating Reproducibility of Differential Expression Discoveries in Microarray Studies by Considering Correlated Molecular Changes", Bioinformatics, vol. 25, no. 13, 2009, pp.1662–1668.

[22] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li, and X. Li, "Apparently Low Reproducibility of True Differential Expression Discoveries in Microarray Studies", Bioinformatics, vol. 24, no. 18, 2008, pp. 2057–2063.

[23] S. Loscalzo, L. Yu, and C. Ding, "Consensus Group Stable Feature Selection", in Proceeding of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'09), 2009, pp.567–575.

[24] S.-Y. Kim, "Effects of Sample Size on Robustness and Prediction Accuracy of a Prognostic Gene Signature", BMC Bioinformatics, vol. 10, 2009, p. 147.

[25] Irizarry RA, Hobbs B, Collin F, Beazer YD, Antonellis KJ, et al. "Exploration, Normalization and Summaries of High Density oligonucleotide array probe level data". Biostatistics 4(2): 2003, 249–264.

[26] Irizarry Y, Denker JS, Solla SA, "Optimal Brain Damage". In DS.Touretzky, ed. Advances in Neural Information Processing Systems II, 1990, 598–605, San MateoCA: Morgan Kaufman.

[27] Chang CC, Lin CJ "LIBSVM: a Library for Support Vector Machines", 2001, Software available at http://www.csie.ntu.edu.tw/,cjlin/libsvm.

[28] Christina Yau, Laura Esserman, Dan H Moore, Fred Waldman, John Sninsky, Christopher C Benz1, "A Multigene Predictor of Metastatic Outcome in Early Stage Hormone Receptor-Negative and Triple-Negative Breast Cancer", Breast Cancer Research 2010, 12:R85

[29] Milde-Langosch K, Kappes H, Riethdorf S, Löning T, Bamberger AM, "FosB is Highly Expressed in Normal Mammary Epithelia, But Down-Regulated in Poorly Differentiated Breast Carcinomas", Breast Cancer Res Treat. 2003 Feb; 77(3):265-75.

[30] Chunhua Lu, Qiang Shen1, Elizabeth DuPre´, Heetae Kim, Susan Hilsenbeck and Powel H Brown, "cFos is Critical for MCF-7 Breast Cancer Cell Growth", Oncogene, 2005, 24, 6516–6524.

[31] X Yin, JW DeWille and T Hai, "A Potential Dichotomous Role of ATF3, an Adaptive-Response Gene in Cancer Development", Oncogene, 2008, 27, 2118–2127.

[32] Liu J[1], Du T, Yuan Y, He Y, Tan Z, Liu Z. "KLF6 Inhibits Estrogen Receptor-Mediated Cell Growth in Breast Cancer Via a c-Src-Mediated Pathway". Mol Cell Biochem, 2010 Feb; 335(1-2):29-35.

[33] Ming Wu, Li Liu and Christina Chan1, "Identification of Novel Targets for Breast Cancer by Exploring Gene Switches on a Genome Scale", BMC Genomics 2011, 12:547.

[34] Anna-Mária Tőkés, Janina Kulka, Sándor Paku, Ágnes Szik, Csilla Páska, Pál Kaposi Novák,László Szilák, András Kiss, Krisztina Bögi and Zsuzsa Schaff, "Claudin-1, -3 and -4 Proteins and mRNA Expression in Benign and Malignant Breast Lesions", Breast Cancer Res. 2005; 7(2): R296–R305.