Identifying Informative Genes for Prediction of Breast Cancer Subtypes

Iman Rezaeian[1], Yifeng Li[1], Martin Crozier[2], Eran Andrechek[3], Alioune Ngom[1], Luis Rueda[1], and Lisa Porter[3]

[1] School of Computer Science, University of Windsor,
401 Sunset Avenue, Windsor, Ontario, N9B 3P4, Canada
{rezaeia,li11112c,lrueda,angom}@uwindsor.ca
[2] Department of Biological Sciences, University of Windsor,
401 Sunset Avenue, Windsor, Ontario, N9B 3P4, Canada
{mcrozier,lporter}@uwindsor.ca
[3] Department of Physiology, Michigan State University,
567 Wilson Rd, East Lansing, MI, 48824, United States
andrech1@msu.edu

Abstract. It is known that breast cancer is not just one disease, but rather a collection of many different diseases occurring in one site that can be distinguished based in part on characteristic gene expression signatures. Appropriate diagnosis of the specific subtypes of this disease is critical for ensuring the best possible patient response to therapy. Currently, therapeutic direction is determined based on the expression of characteristic receptors; while cost effective, this method is not robust and is limited to predicting a small number of subtypes reliably. Using the original 5 subtypes of breast cancer we hypothesized that machine learning techniques would offer many benefits for feature selection. Unlike existing gene selection approaches, we propose a tree-based approach that conducts gene selection and builds the classifier simultaneously. We conducted computational experiments to select the minimal number of genes that would reliably predict a given subtype. Our results support that this modified approach to gene selection yields a small subset of genes that can predict subtypes with greater than 95% overall accuracy. In addition to providing a valuable list of targets for diagnostic purposes, the gene ontologies of selected genes suggest that these methods have isolated a number of potential genes involved in breast cancer biology, etiology and potentially novel therapeutics.

Key words: breast tumor subtype, gene selection, classification

1 Introduction

Despite advances in treatment, breast cancer remains the second leading cause of cancer related deaths among females in Canada and the United States. Previous studies have revealed that breast cancer can be categorized into at least five subtypes, including basal-like (Basal), luminal A, (LumA), luminal B (LumB), HER2-enriched (HER2), and normal-like (Normal) types [1, 2]. These subtypes have their own genetic signatures, and response to therapy varies dramatically from one subtype to another. The

variability among subtypes holds the answer to how to better design and implement new therapeutic approaches that work effectively for all patients. It is clinically essential to move toward effectively stratifying patients into their relevant disease subtype prior to treatment.

Techniques such as breast MRI, mammography, and CT scan, can examine the phenotypical mammary change, but provide little effective information to direct therapy. Genomic techniques provide high-throughput tools in breast cancer diagnosis and treatment, allowing clinicians to investigate breast tumors at a molecular level. The advance of microarray approaches have enabled genome-wide sampling of gene expression values and/or copy number variations. The huge amount of data that has been generated has allowed researchers to use unsupervised machine learning approaches to discover characteristic "signatures" that have since established distinct tumor subtypes [1]. Tumor subtyping has explained a great deal about some of the mysteries of tumor pathology [3], and has begun to enable more accurate predictions with regard to response to treatment [4]. While offering enormous opportunity for directing therapy, there are some challenges arising in the analysis of microarray data. First, the number of available samples (e.g. patients) is relatively small compared to the number of genes measured. The sample size typically ranges from tens to hundreds because of costs of clinical tests or ethical constraints. Second, microarray data is noisy. Although the level of technical noise is debatable [5], it must be carefully considered during any analysis. Third, due to technical reasons, the data set may contain missing values or have a large amount of redundant information. These challenges affect the design and results of microarray data analysis.

This current study focuses on identifying a minimal number of genes that will reliably predict each of the breast cancer subtypes. Being a field of machine learning, pattern recognition can be formulated as a feature selection and classification problem for multi-class, high-dimensional data using two traditional schemes. The first applies a multi-class "feature selection" method directly followed by a classifier to measure the dependency between a particular feature and the multi-class information. A well-known example of the feature selection method is the minimum redundancy maximum relevance (mRMR) method proposed in [6] and [7]. The second traditional scheme is the most common of the two and treats the multi-class feature selection as multiple binary-class selections. Methods using multiple binary class selections differ in how to bisect the multiple classes. The two most popular ways to solve this problem are one-versus-one and one-versus-all [8]. In this paper, we propose a novel and flexible hierarchial framework to select discriminative genes and predict breast tumor subtypes simultaneously. The main contributions of this paper can be summarized as follows:

1. We implement our framework using *Chi2* feature selection [9] and a *support vector machine (SVM) classifier* [10] to obtain biologically meaningful genes, and to increase the accuracy for predicting breast tumor subtypes.
2. We Use a novel feature selection scheme with a hierarchial structure, which learns in a cross-validation framework from the training data.
3. We establish a flexible model where any feature selection and classifier can be embedded for use.

4. We discover a new, compact set of biomarkers or genes useful for distinguishing among breast cancer types

2 Related Work

Using microarray techniques, scientists are able to measure the expression levels for thousands of genes simultaneously. Finding relevant genes corresponding to each type of cancer is not a trivial task. Using hierarchical clustering, Perou and colleagues developed the original 5 subtypes of breast cancer based on the relative expression of 500 differentially expressed genes [1]. It has since been demonstrated that combining platforms to include DNA copy number arrays, DNA methylation, exome sequencing, microRNA sequencing and reverse-phase protein arrays may define these subtypes even further [2]. It is postulated that there are, indeed, upward of over 10 different forms of breast cancer with differing prognosis [25]. Other groups have tailored analysis toward refining the patient groups based on relative prognosis, reducing the profile for one subtype to a 14-gene signature [26]. Given any patient subtype, obtained through one or several platforms, we hypothesize that machine learning approaches can be used to more accurately determine the number of genes required to reliably predict a subtype for a given patients.

On the other hand, modeling today's complex biological systems requires efficient computational techniques designed in articulated model, and used to extract valuable information from existing data. In this regard, pattern recognition techniques in machine learning provide a wealth of algorithms for feature extraction and selection, classification and clustering. A few relevant approaches are briefly discussed then.

An entropy-based method for classifying cancer types was proposed in [16]. In entropy-classed signatures, the genes related to the different cancer subtypes are selected, while the redundancy between genes is reduced simultaneously. Recursive feature addition (RFA) has been proposed in [17], which combines supervised learning and statistical similarity measures to select relevant genes to the cancer type. A mixture classification model containing a two-layer structure named as mixture of rough set (MRS) and support vector machine (SVM) was proposed in [18]. This model is constructed by combining rough sets and SVM methods, in such a way that the rough set classifier acts as the first layer to determine some singular samples in the data, while the SVM classifier acts as the second layer to classify the remaining samples. In [19], a binary particle swarm optimization (BPSO) was proposed. BPSO involves a simulation of the social behavior in organisms such as bird flocking and fish schooling. In BPSO, a small subset of informative genes is selected where the genes in the subset are relevant for cancer classification. In [20], a method for selecting relevant genes in comparative gene expression studies was proposed, referred to as *recursive cluster elimination* (RCE). RCE combines k-Means and SVM to identify and score (or rank) those gene clusters for the purpose of classification. k-Means is used initially to group the genes into clusters. RCE is then applied to iteratively remove those clusters of genes that contribute the least to classification accuracy. In the work described in this paper we used the original five breast cancer subtypes to determine whether our proposed hierarchial tree-based scheme could reduce the gene signature to a reliable subset of relevant genes.

3 Methods

First, we describe the training phase for gene selection and breast cancer subtyping, and then we describe how the model can be used in predicting subytpes in a clinical setting. The complete gene profile of each breast cancer subtype is compared against the others. Each subtype varies in the genes that are associated with it, and in the accuracy with which those genes predict that specific subtype. The subtypes are then organized by two main criteria. The first criterion is the level of accuracy with which the selected genes identify the given subtype. The second criterion is the number of genes identified. Clearly applying two or more gene selection criteria is a multi-objective problem in optimization [21]. In this study, we use the rule that select the smallest subset of genes that yields the highest accuracy. Therefore, a subtype that is predicted with 95% accuracy by five genes is ranked higher than a subtype for which 20 genes are required to acquire the same accuracy. The subtype that is ranked highest is removed and the procedure is repeated for the remaining subtypes comparing each gene profile against the others. The highest ranked subtype is again removed and becomes a leaf on the hierarchical tree (see Fig. 1). Therefore, each leaf on the tree becomes a distinct subtype outcome.

3.1 Training Phase

We give an example of such a tree to illustrate our method in Fig. 1. Suppose there are five subtypes, namely $\{C_1, \cdots, C_5\}$. The training data is a $m \times n$ matrix $\boldsymbol{D} = \{\boldsymbol{D}_1, \cdots, \boldsymbol{D}_5\}$ corresponding to the five subtypes. $\boldsymbol{D}_i$, of size $m \times n_i$, is the training data for class C_i. m is the number genes and n_i is the number of samples in subtype C_i. $n = \sum_{i=1}^{5} n_i$ is the total number of training samples from all five classes. First of all, feature selection and classification are conducted, in a cross-validation fashion, for each class against the other classes. For example, suppose subtype C_3 obtains the highest rank based on accuracy and the number of genes contributing to that accuracy. We thus record the list of the particular genes selected and create a leaf for that subtype. We then remove the samples of the subtype, which results in $\boldsymbol{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_4, \boldsymbol{D}_5\}$ and continue the process in the same fashion. Thus, at the second level, subtype C_5 yields the highest rank, and hence its gene list is retained and a leaf is created. Afterward the training data set becomes $\boldsymbol{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_4\}$ for the third level. We repeat the training procedure in the same fashion until there is no subtype to classify. At the last level, two leaves are created, for C_4 and C_2, respectively.

3.2 Prediction Phase

Once the training is complete, we can apply the scheme to predict breast cancer subtypes. Given the gene expression profile of a new patient, a sequence of classification steps are performed by tracing a path from the root of the tree toward a leaf. At each node in the path, only the genes selected in the training phase are tested. The process starts at the first level (root of the tree), in which case only the genes selected for C_3, namely G_3 are tested. If the patient's gene profile is classified as a positive sample, then the prediction outcome is subtype C_3, and the prediction phase terminates. Otherwise, the sequence of classification tests is performed in the same fashion, until a leaf

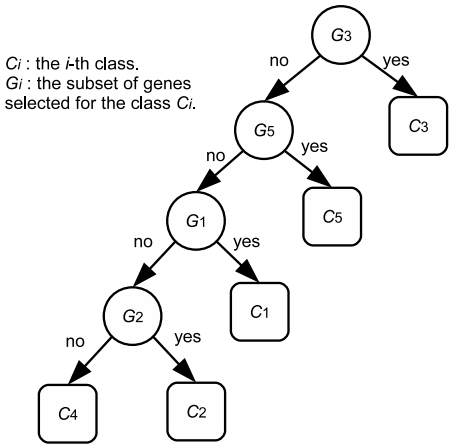

Fig. 1. Determining breast cancer type using selected genes.

is reached, in which case the prediction outcome is the subtype associated with the leaf that has been reached.

3.3 Characteristics of The Method

Our structured model has the following characteristics. First, it involves a greedy scheme that tries the subtype which obtains the most reliable prediction and the smallest number of genes first. Second, it conducts feature selection and classification simultaneously. Essentially, it is a specific type of decision tree for classification. The differences between the proposed model and the traditional decision tree includes: i) each leaf is unique, while one class usually has multiple leaves in the later; ii) classifiers are learned at each node, while the traditional scheme learns decision rules; and iii) multiple features can be selected, while in the traditional scheme each node corresponds to only one feature. Third, the proposed model is flexible as any feature selection method and classifier can be embedded. Obviously, a classifier that can select features simultaneously also applies, (e.g. the l_1-norm SVM [11]).

3.4 Implementation

In this study, we implement our model by using Chi2 feature selection [9] and the state-of-the-art SVM classifier [10]. These two techniques are briefly described briefly next. Chi2 is an efficient feature selection method for numeric data. Unlike some traditional methods which discretize numeric data before conducting feature selection, Chi2 *automatically* and *adaptively* discretizes numeric features and selects features as well. It keeps merging adjacent discrete statuses with the lowest χ^2 value until all χ^2 values exceed their confidence intervals determined by a decreasing significant level, while keeping consistency with the original data. If, finally, a feature has only one discrete

status, it is removed. The χ^2 value of a pair of adjacent discrete statuses or intervals is computed by the χ^2 statistic, with 1 degree of freedom, as follows:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{1}$$

where n_{ij} is the number of samples in the i-th interval and j-th class, and e_{ij} is the expected value of n_{ij}. e_{ij} is defined as $r_i \frac{c_j}{n}$ where $r_i = \sum_{j=1}^{k} n_{ij}$, $c_j = \sum_{i=1}^{2} n_{ij}$, and n is the total number training samples.

Based on these selected genes, the samples are classified using SVM [10]. Soft-margin SVM is applied in our current study. SVM is a linear maximum-margin model with decision function $d(\boldsymbol{x}) = \text{sign}[f(\boldsymbol{x})] = \text{sign}[\boldsymbol{w}^{\text{T}}\boldsymbol{x} + b]$ where $\boldsymbol{w}$ is the normal vector of the separating hyperplane and b is the bias. Soft-margin SVM solves the following problem in order to obtain the optimal $\boldsymbol{w}$ and b:

$$\begin{aligned} &\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \boldsymbol{C}^{\text{T}}\boldsymbol{\xi} \\ &\text{s.t. } \boldsymbol{Z}^{\text{T}}\boldsymbol{w} + b\boldsymbol{y} \geq \mathbf{1} - \boldsymbol{\xi} \\ &\quad \boldsymbol{\xi} \geq 0, \end{aligned} \tag{2}$$

where $\boldsymbol{\xi}$ is a vector of slack variables, $\boldsymbol{C}$ is a vector of constant that controls the trade-off between the maximum margin and the empirical error, $\boldsymbol{y}$ is a vector that contains the class information (either -1 or +1), and $\boldsymbol{Z}$ contains the normalized training samples with its i-th column defined as $\boldsymbol{z}_i = y_i\boldsymbol{x}_i$ [13]. Since optimization of the SVM involves inner products of training samples, by replacing the inner products by a kernel function, we can obtain a kernelized SVM.

For the implementation, the Weka machine learning suite was used [14]. A gene selection method based on the χ^2 feature evaluation algorithm was first used to find a subset of genes with the best ratio of accuracy/gene number [9]. For classification, LIBSVM [15] in Weka is employed. The *Radial basis function* (RBF) kernel is used with the LIBSVM classifier without normalizing samples and with default parameter settings.

4 Computational Experiments and Discussions

4.1 Experiments

In our computational experiment, we analyzed Hu's data [12]. Hu's data (CEO accession number GSE1992) were generated by three different platforms including Agilent-011521 Human 1A Microarray G4110A (feature number version) (GPL885), Agilent-012097 Human 1A Microarray (V2) G4110B (feature number version) (GPL887), and Agilent Human 1A Oligo UNC custom Microarrays (GPL1390). Each platform contains 22,575 probesets, and there are 14,460 common probesets among these three platforms. We used SOURCE [22] to obtain 13,582 genes with unique unigene IDs in order to merge data from different platforms. The dataset contains 158 samples from

five subtypes of breast cancer(13 Normal, 39 Basal, 22 Her2, 53 LumA and 31 LumB). The sixth subtype Claudin is excluded from our current analysis as the number of samples of this class is too few (only five). However, we will investigate this subtype in our future work.

To evaluate the accuracy of the model, 10-fold cross-validation is used. As shown in Table 2, using all genes decreases the overall accuracy of the model, since many of the genes are irrelevant or redundant. For example, using all 13,582 genes, the overall accuracy is just 77.84%; while using a ranking algorithm and taking the top 20 genes for prediction brings the accuracy up to 86.70%. Table 1 shows the top 20 genes ranked by the Chi-Squared attribute evaluation algorithm to classify samples as one of the five subtypes. Using the proposed hierarchical decision-tree-based model, makes the prediction procedure more accurate. While the accuracy of prediction between LumA and LumB is relatively low compared to the other classes. This is because of the very high similarity and overlap between samples of these two classes. The overall accuracy of the model, as shown in Table 2, is 95.11%. This is very interesting since only 18 genes are used to predict the subtypes that the patient belongs to. As a matter of fact, our method is able to increase its accuracy from around 86% to 95% by using a new subset of genes based on the proposed method containing only 18 genes.

Table 1. Top 20 genes ranked by the Chi-Squared attribute evaluation algorithm to classify samples as one of the five subtypes.

Rank	Gene Name	Rank	Gene Name	Rank	Gene Name	Rank	Gene Name
1	FOXA1	6	THSD4	11	DACH1	16	ACOT4
2	AGR3	7	NDC80	12	GATA3	17	B3GNT5
3	CENPF	8	TFF3	13	INPP4B	18	IL6ST
4	CIRBP	9	ASPM	14	TTLL4	19	FAM171A1
5	TBC1D9	10	FAM174A	15	VAV3	20	CYB5D2

Fig. 2 shows the tree learned in the training phase and the set of genes selected at each step. The selected genes are contained in each node, a patient's gene expression profile is used to feed the tree for prediction, each leaf represents a subtype, and the accuracy at each classification step is under the corresponding node. From this figure, we can see that the Basal subtype is chosen first as it obtains the highest accuracy, 99.36% to classify patients from the other subtypes including Normal, Her2, LumA and LumB. Then the samples of Basal are removed for the second level. The Normal subtype is chosen then, since it achieves the highest accuracy (95.79%) to separate samples from the other subtypes, including Her2, LumA and LumB. From previous studies, it is well-known that the subtypes LumA and LumB are very difficult to be identified among all subtypes. This is the reason for why LumA and LumB appear at the bottom of the tree. After removing other subtypes, LumA and LumB can avoid misclassification on the other subtypes. In spite of this drawback, the accuracy for separating LumA and LumB is as high as 88.1%.

As shown in Figure 2, there is no overlap between the genes selected among the different clusters. This result provides interesting new biomarkers for each breast cancer

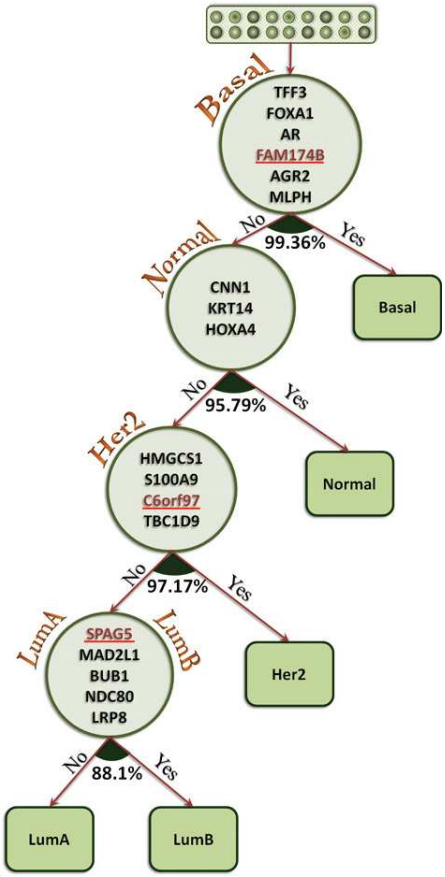

Fig. 2. Determining breast cancer type using selected genes.

subtype. Some of the selected genes have been previously indicated in cancer (highlighted in black in Figure 2), while others have emerged as interesting genes to be investigated. For example, TFF3 and FoxA1 genes are predictably indicated in Basal subtype. Another feature of the proposed hierarchical model is that the number of genes in each node has been optimized to give the best ratio of accuracy and number of selected genes. For this, at first, 10 genes with highest rank have been selected for each node. Then, out of those selected genes, those with lower rank are removed step by step as long as the accuracy of classification using the remaining genes don't get decreased.

4.2 Biological Insight

We used FABLE to determine if the genes selected by our approach are biologically meaningful. Fast Automated Biomedical Literature Extraction (FABLE) is a web-based tool to search through MEDLINE and PubMed databases. The genes that are related

Table 2. Accuracy of classification using LibSVM Classifier

Classification Method	Gene Selection Method	# of Genes	Accuracy	Precision	Recall	F-measure
LibSVM	—	all genes	77.84%	0.802	0.778	0.749
LibSVM	Chi-Squared	20	86.70%	0.866	0.867	0.864
Proposed Method	Proposed Method	18	95.11%	0.951	0.951	0.951

to tumors reported in the literature are highlighted in black in Figure 2. Those not yet reported are underlined and colored in red. We can see that 15 out of 18 genes have been found in the literature. This implies that our approach is quite effective in discovering new biomarkers.

We also explored the reasons for the high performance of our method. First, the subtypes that are easily classified are on the top of the tree, while the harder subtypes are considered only after removing the easier ones. Such a hierarchical structure can remove the disturbance of other subtypes, thereby allowing us to focus on the most difficult subtypes, LumA/B. Second, combining gene selection when building the classifier allows us to select genes that contribute to prediction accuracy. Third, our tree-based methodology is quite flexible; any existing gene selection measure and classification technique can be embedded in our model. This will allow us to apply this model to subtypes as they become more rigorously defined using other platforms such as copy number variation. Furthermore, our method could be applied to groups of patients stratified based on responses to specific treatments. Collectively, having a small, yet reliable number of genes to screen is more cost effective and would allow for subtype information to be more readily applied in a clinical setting.

5 Conclusion and Future Work

In this study, we proposed a novel gene selection method for breast cancer subtype prediction based on a hierarchical, tree-based model. The results demonstrate an impressive accuracy to predict breast cancer types using only 18 genes. Herein, we propose a novel gene selection method for breast cancer subtype prediction based on a hierarchical, tree-based model. The results demonstrate an impressive accuracy to predict breast cancer subtypes using only 18 genes in total. Moreover, Most of the selected genes are shown to be related to breast cancer based on previous studies, while a few are yet to be investigated. As future work, we will validate these results using cell lines that fall within a known subtype. We will determine whether our predicted 18 gene array can accurately denote which subtype each of these cell lines falls under. This hierarchical, tree-based model can narrow down analysis to a relatively small subset of genes. Importantly, the method can be applied to more refined stratification of patients in the future, such as subtypes derived using a combination of platforms, or for groups of patients that have been subdivided based on response to therapy. Using this computational tool we can determine the smallest possible number of genes that need to be screened for accurately placing large populations of patients into specific subtypes of cancer or specified treatment groups. This could contribute to the development of improved screening tools, providing increased accuracy for a larger patient population than that achieved by

Oncotype DX, but allowing for a cost effective approach that could be widely applied to the patient population.

Acknowledgments. This research has been supported by grants from Seeds4Hope (WECCF), CBCRA (#02051), and Canadian NSERC Grants #RGPIN228117-2011 and #RGPIN261360-2009.

References

1. Perou, C.M., et al.: Golecular Portraits of Human Breast Tumours. Nature. 406, 747–752 (2000)
2. Perou, C.M., et al.: Comprehensive Molecular Portraits of Human Breast Tumours. Nature. 490, 61–70 (2012)
3. Chandriani,S., Frengen, E., Cowling, V.H., Pendergrass, S.A., Perou, C.M., Whitfield, M.L., Cole, M.D.: A Core MYC Gene Expression Signatures is Prominent in Basal-Like Breast Cancer but only Partially Overlaps the Core Serum Response. PLOS One. 4(8), e6693 (2009)
4. van't Veer, L.J., et al.: Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. Nature. 415(6871), 530–536 (2002)
5. Klebanov, L., Yakovlev, A.: How High is The Level of Technical Noise in Microarray Data?. Biology Direct. 2, 9 (2007)
6. Ding, C., Peng, H.: Munimun Redundancy Feature Selection from Microarray Gene Expression Data. Journal of Bioinformatics and Computational Biology. 3(2), 185–205 (2005)
7. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(8), 1226–1238 (2005)
8. Li, T., Zhang, C., Ogihata, M.: A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Vased on Gene Expression. Bioinformatics. 20(15), 2429–2437 (2004)
9. Liu, H., Setiono, R.: Chi2: Feature Selection and Discretization of Numeric Attributes. In: IEEE International Conference on Tools with Artificial Intelligence, pp. 388–391. IEEE Press, New York (1995)
10. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
11. Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: 1-Norm Support Vector Machines. In: NIPS, MIT Press, Cambridge, MA (2004)
12. Hu, Z., et al.: The Molecular Portraits of Breast Tumors are Conserved Across Microarray Platforms. BMC Genomics. 7, 96 (2006)
13. R. O. Duda and P. E. Hart and D. G. Stork: Pattern Classification. Wiley-Interscience, New York (2006)
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter. 11(1), 10–18 (2009)
15. Chang, C.-C, Lin, C.-J.: LIBSVM: a Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology. 12, 27:1–27:27 (2011)
16. Liu, X., Krishnan, A., Mondry, A.: An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data. BMC Bioinformatics. 6, 76 (2005)
17. Liu, Q., Sung, A.H., Chen, Z., Liu, J., Huang, X., Deng, Y.: Feature Selection and Classification of MAQC-II Breast Cancer and Multiple Myeloma Microarray Gene Expression Data. PLoS One. 4(12), e8250 (2009)

18. Zeng, T., Liu, J.: Mixture Classification Model Based on Clinical Markers for Breast Cancer Prognosis. Artificial Intelligence in Medicine. 48, 129–137 (2010)
19. Mohamad, M.S., Omatu, S., Deris, S., Yoshioka, M.: Particle Swarm Optimization for Gene Selection in Classifying Cancer Classes. Artificial Life and Robotics. 14(1), 16–19 (2009)
20. Yousef, M., Jung, S., Showe, L., Showe, M.: Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. BMC Bioinformatics. 8, 144 (2007)
21. Li, Y., Ngom, A., Rueda, L.: A Framework of Gene Subset Selection Using Multiobjective Evolutionary Algorithm. LNBI/LNCS. 7632, 38–48 (2012)
22. Diehn, M., et al.: SOURCE: a Unified Genomic Resource of Functional Annotations, Ontologies, and Gene Expression Data. Nucleic Acids Research. 31(1), 219–223, (2003) Available at http://smd.stanford.edu/cgi-bin/source/sourceSearch
23. Sorlie, T., et al.: Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications. PANS. 98(19), 10869–10874 (2001)
24. Sorlie, T., et al.: Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets. PANS. 100(14), 8418–8423 (2003)
25. Curtis, C., et al.: The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. Nature. 486(7403), 346–352 (2012)
26. Hallett, R.M., Dvorkin-Gheva, A., Bane, A., Hassell, J.A.: A Gene Signature for Predicting Outcome in Patients with Basal-Like Breast Cancer. Scientific Reports. 2, 227 (2012)