Capturing Heuristics and Intelligent Methods for Improving Micro-Array Data Classification

Andrea Bosin, Nicoletta Dessì, and Barbara Pes

Università degli Studi di Cagliari, Dipartimento di Matematica e Informatica,
Via Ospedale 72, 09124 Cagliari, Italy
andrea.bosin@dsf.unica.it,{dessi,pes}@unica.it

Abstract. Classification of micro-array data has been studied extensively but only a small amount of research work has been done on classification of micro-array data involving more than two classes. This paper proposes a learning strategy that deals with building a multi-target classifier and takes advantage from well known data mining techniques. To address the intrinsic difficulty of selecting features in order to promote the classification accuracy, the paper considers the use of a set of binary classifiers each of ones is devoted to predict a single class of the multi-classification problem. These classifiers are similar to local experts whose knowledge (about the features that are most correlated to each class value) is taken into account by the learning strategy for selecting an optimal set of features. Results of the experiments performed on a publicly available dataset demonstrate the feasibility of the proposed approach.

1 Introduction

A challenging approach to predict the outcome of certain biological events involves using micro-array experimental data for classifying tumors as either benign or malignant. The general goal of this research line is to turn a qualitative diagnosis by a pathologist into a quantitative diagnosis while leading to a finer understanding of cancer mechanisms.

Currently, each and every micro-array experiment requires measuring several thousand gene fragments and inducing classifiers over them. The classification difficulty depends greatly on the data provided by DNA micro-arrays as they are of high dimensionality while the collected samples are small in number. To simplify, a micro-array dataset is a matrix of m-dimensional observations (i.e. one for each patient) where each dimension is a feature i.e. a measurable property of a specific gene that is designed as gene-expression profile. The class is a label associated to the observation meaning that all samples can be classified as belonging to one of different classes.

While significant progress has been made in the development of machine learning methods for binary classification [1] [2] [3], finding good classifiers is known to be a difficult task when gene expression profiles are used as complex biomarkers defining many different classes of cancer (multi-target classification) and only a small amount of research work has been done on classification involving more than two classes.

Specifically, difficulties arise when the micro-array dataset exhibits one class that is much more clearly characterized than others (such as [4] or [5] which analyze different tumor types). This makes it impossible to determine which genes are responsible of each single pathology (class value) because the feature selection results in a set that might be most representative of only one (or some, not all) class value.

The purpose of this paper is to present a method for micro-array multi-target classification that tries to overcome the above problems by taking into account two major concerns: the learning strategy and the acquisition of knowledge from specialized binary classifiers. The paper separates these concerns by proposing a learning strategy that deals with building a multi-target classifier and takes advantage from well known data mining techniques (Naïve Bayes [6], Support Vector Machines [7], k-Nearest Neighbor [8]). Focused on the feature selection process, this strategy provides an iterative method that determines at each step which features to acquire next in order to find an optimal set of features. A major contribute of this paper is in the second concern that deals with proposing the use of a set of binary classifiers each of ones is devoted to predict a single class of the multi-classification problem. From this point of view, the binary classifiers are similar to local experts whose knowledge is taken into account by the learning process that collects and takes benefit of information concerning the optimal set of features that each single binary classifier selects.

The paper is organized as follows. Section 2 briefly summarizes some related works. Section 3 illustrates our learning strategy, whose validation is asserted by experiments presented in Section 4. Finally, Section 5 presents a brief discussion as well as concluding remarks.

2 Related Work

Building a predictive model (i. e. a classifier that is expected to discriminate normal from cancer tissues or to distinguish among different classes of tumors) presents multiple challenges, because the micro-array data include a large number of gene expression values per experiment (several thousands of features), and a relatively small number of samples (a few dozen of patients). Most features being irrelevant to the problem at hand, the micro-array data exhibit a high degree of noise: a large number of features in input to the learning algorithms may turn them to build inefficient classifiers with the additional charge of memory and time consumption. Moreover, correlation between feature sets results in the counter effect of over-fitting [3]: that is creating classifiers that may not generalize well to new data from the same type and distribution, despite their excellent accuracy on the training set.

A number of studies have shown that accurate classification of micro-array data can be made using a reduced number of genes [3][4][9] indicating that it is more important to explore data and utilize independent features to train classifiers, rather than increase the number of features we use. Additionally, the identification of discriminatory genes is of fundamental and practical interest since medical diagnostic tests may benefit from the examination of a small subset of relevant genes.

The underlying distribution of the features being not known, the feature selection process originates ambiguity in deciding which group of features constitutes an optimal set. The central question of the “minimum informative subset problem” still remains an active focus of micro-array research and is a challenging task because the number of possible feature subsets increases exponentially with the number of features, making exhaustive search impractical.

Along with this central question there is a range of other questions, compactly summarized in [10], and it is not clear how to proceed with feature selection, knowing that under different conditions the selected feature subsets will likely be different depending on sample sizes for the validation process and on different selection methods. Furthermore, the small sample size and high dimensionality of the data constrain the possibility of properly validating the chosen classification model and different classifiers perform differently on micro-arrays, with dataset sparsity as the major contributor for the differences [11]. Selecting simple classifiers that need minimal parameter tuning seems to be the appropriate approach, independently of data complexity and especially for small sample sizes.

3 The learning strategy

A multi-target classification problem can be treated directly or decomposed into several binary classification problems. However, it has been observed [12] that the direct application of a learning strategy to the multi-target problem may result in an over-representation of the abundant and/or easily separable classes. To circumvent the above mentioned limitation, we consider breaking the original M-class problem into a set of binary sub-problems (one for each class) and performing classification by training and combining these binary classifiers with respect to some criterion.

Because we are interested in separating each pathology from all the others, the proposed strategy adopts the one-versus-rest (1-vs-r) classification schema that provides M binary classifiers, each of ones is trained for distinguishing between a given class and the M-1 other remaining classes whose instances are considered as negative examples. According to this schema, the first step of the learning strategy decomposes the original dataset into M sub-datasets, each separating the instances of a given class from the rest of the classes. Then, a filter is applied to each sub-dataset that orders the features by their rank within each class, a high score being indicative of a relevant variable. Specifically, the ranking procedure concentrates on the correlation of each gene with only one class value at a time, against all the others, and originates M sets of ranked genes, each of ones is correlated only with one given pathology (this is much more interesting also from a medical point of view): irrelevant features are excluded from the classification task, thereby reducing both the noise of the dataset as well as the time needed to perform the classification.

This pre-processing step, which is independent from the classification task, is based on some statistical criteria (i.e. MDL, χ^2) and provides a basis for numerically weighting the variables individually, but the next crucial step is the selection of a smaller number of highly specific features, i.e. an “optimal” set of features to employ in learning each binary classifier. The classification process dealing with a large number of variables, it is computationally intractable to search the whole space of feature subsets and one has to settle for approximations of the minimal optimal set of features that significantly improves the learning algorithm's performance. Consequently, a feature selection process defines both a strategy to search the space of possible feature subsets and a measure for assessing the goodness of the selected subset.

Our approach considers each binary class decision as problem instance and selects features for it separately. It results in M-separated processes of feature selection each of ones first considers the N top-ranked attributes for the sub-dataset under examination. The basic idea is to select variables step by step according to their predictive power and using a sequential forward selection that starts the search with an empty variable subset. During each single step, one or more variables are considered for the inclusion in the subset using as criterion the performance of a classifier built with a new subset that is obtained by the inclusion of the considered variables**.**

Table 1 depicts a general schema of the learning strategy that is applied to each single sub-dataset.

Table 1. A general schema of the learning strategy

1	Consider a single sub-dataset R and rank all its features according to a statistical criterion
2	Select the N top-ranked features (e.g. start with N = 1)
3	Build a binary classifier with the N top-ranked features
4	Test classifier accuracy (on an independent test dataset)
5	Extend the feature subset by adding the next k top-ranked features (e.g. k = 1) and put N = N + k
6	Repeat steps 2 to 5 and stop if the accuracy has not increased according to some stopping criterion or after a fixed number of iterations

Because just a small number of variables is needed for separating the micro-array data, a ranking criterion based on classification success rate can distinguish between the top-ranked variables [3]. However, the filter methods and the classification algorithms are not directly relevant for this study: the question we address here is that of comparing the above mentioned nested classifiers and proposing a stopping criterion for halting iterations.

Measured in terms of false positive (FP) classification rate and false negative (FN) classification rate, the classification accuracy is usually still the only measure used for evaluating the performance of micro-array classifiers that are based on data mining techniques. As well, measures such recall and precision are popular metrics employed in data mining applications. Moreover, it is quite common to monitor the tradeoff between true positives (TP) and false positives (FP) by graphical means such as the Receiver Operating Characteristic (ROC) curve that shows FP on the x-axis while TP is plotted along the y-axis. The visualization of the classifier performance is one of the attractive features of ROC analysis that is useful for comparing the relative performance among different classifiers while the area under the ROC curve (AUC) provides another approach for evaluating which model is better on average.

A recent work [13] outlines that ROC curves are inadequate for the needs of data mining research in several significant respects and demonstrate the validity of the cost curves [14] as a graphical mean for overcoming these deficiencies. Based on the point/line duality between ROC space and cost space, this graphical technique (best detailed in [15]) plots, along the x-axis, the *probability cost* $p_C[+]$, defined as

$$p_C[+] = p[+]C[+|-] \,/\, (p[+]C[+|-] + p[-]C[-|+]) \tag{1}$$

and, along the y-axis, the *normalized expected cost* NEC, defined as

$$NEC = FN \cdot p_C[+] + FP \cdot (1 - p_C[+]) \tag{2}$$

By $p[+]$ we denote the probability of the positive class, and by $C[+|-]$ the cost of predicting + when the instance is actually - (and by $C[-|+]$ the reverse). Basically, the probability cost $p_C[+]$ is a distorted version of $p[+]$ based on the cost notion, while the normalized expected cost measures the classifier performance weighting each classification error by the corresponding cost. Both the probability cost and the normalized expected cost range from 0 to 1.

We experimented cost curves as a measure for comparing the performance of a family of nested classifiers and for selecting the classifier that has the lowest cost. Specifically, we modified the proposed strategy by evaluating, at the end of each step, the cost curves of the actual and the previous classifier: the vertical distance between the cost curves directly indicates the performance difference between them. The iteration is carried on as long as the cost difference is greater than a fixed threshold and the variable subset corresponding to the lowest cost classifier is assumed as optimal subset.

This approach assumes that each binary classifier is best inside certain subsets of the whole feature domain. However, the problem is to integrate this information coming from multiple independent binary classifiers that are similar to specialized local experts. To this end, a new approach is here explored that considers to treat directly the multi-class problem by a multi-classifier learnt on a “globally optimal” subset of features that results by joining all the optimal subsets generated by the presented learning strategy.

This “knowledge-based” multi-classifier is a mechanism to combine information received from several sources while the learning strategy is employed just for feature selection i.e. and for achieving knowledge about genes featuring each single class. Because only the genes really relevant to each pathology are involved in multi-classification, it is expected that the accuracy of a “knowledge-based” multi-classifier may be increased by the knowledge of local binary classifiers that are most reliable for specific domains.

4 Experiments

The Acute Lymphoblastic Leukaemia dataset [16] (ALL, from now) has been used as a test-bed for the experiments presented in this section. ALL consists of 327 samples (specifically, 215 training and 112 test samples), each one described by the expression level of 12558 genes. 7 classes are involved in total, i.e. all known ALL sub-types (T-ALL, E2A-PBX1, TEL-AML1, BCR-ABL, MLL, Hyperdip > 50) and a generic class OTHERS, that groups all samples not belonging to any of the previous sub-types.

For all the experiments we carried on, we applied the χ^2 statistics, as a ranking criterion, and the following popular classification algorithms: Naïve Bayes (NB) [6], Support Vector Machines (SVM) [7], k-Nearest Neighbor (k-NN) [8]. All the experiments have been carried out using the Weka machine learning software package [15] and a software tool based on cost curves [13].

Fig. 1 shows the accuracy of SVM, NB and k-NN multi-classifiers as a function of the number of features when the multi-target classification problem is treated directly. This means that feature selection is performed by measuring the correlation of each gene simultaneously with all class values. As can we see, the behavior of all classifiers is similar: the accuracy has some initial oscillation and a large number of features (between 750 and 1500) is necessary to reach a plateau and achieve the maximum accuracy, in agreement with recent literature [17] [18]. Additionally, we cannot discriminate which features are mostly correlated to a specific class value.

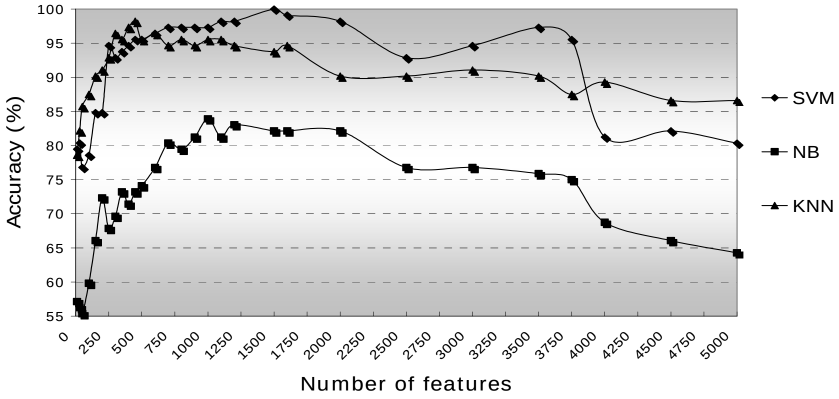

Fig. 1. Accuracy of SVM, NB and k-NN multi-classifiers.

In the following, results are presented using the proposed learning strategy. As a first step, the original training set was decomposed into 6 binary sub-datasets by separating the instances of each single sub-type from the rest whose instances were considered negative examples. After data preparation, we applied the ranking procedure that discriminates, for each sub-dataset, the set of features most strongly related to each target class. This resulted in 6 sets of ranked features ordered by their rank within each class. Then, for each classification algorithm (i.e. SVM, NB and k-NN), we applied the learning strategy described in Section 3.

Fig. 2 shows, in the left panel, the cost curves for two nested binary classifiers learnt with 3 (dashed line) and 9 features (solid line) when the SVM algorithm is applied to the sub-dataset that separates the class BCR-ABL. The cost difference is clearly observable and shows that the accuracy increases while augmenting the number of features. On the contrary, the right panel in Fig. 2 shows an accuracy reduction when a new feature is added in building the family of nested classifiers for the sub-dataset that separates the class T-ALL. Specifically, the cost curves depict the performance of the SVM classifier learnt with the first top-ranked feature (dashed line) and the first two top-ranked features (solid line).

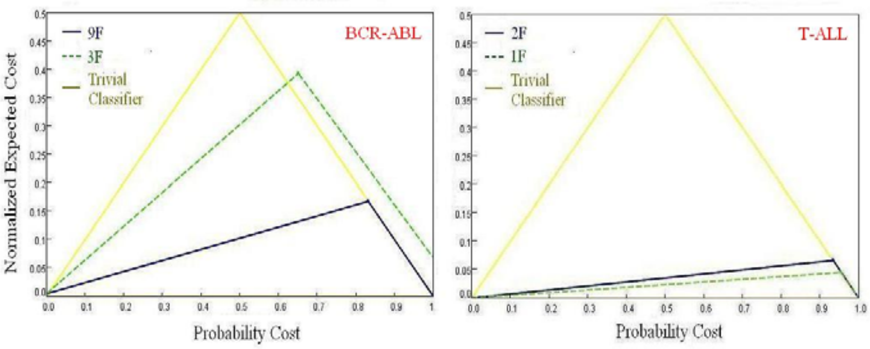

Fig. 2. Cost curves for two SVM nested classifiers learnt on the sub-dataset that separates the class BCR-ABL (left panel) and the class T-ALL (right panel).

Table 2 shows the resulting optimal number of features respective of each single class as well as the accuracy of the binary classifiers built on these optimal subsets (in brackets).

Table 2. Optimal number of features for each ALL sub-type and best accuracy for each binary classifier (in brackets).

	T-ALL	*E2A-PBX1*	*TEL-AML1*	*BCR-ABL*	*MLL*	*Hyperdip>50*
SVM	1 *(100 %)*	1 *(100 %)*	12 *(100 %)*	9 *(99,1 %)*	3 *(100 %)*	5 *(96,4 %)*
NB	1 *(100 %)*	1 *(100 %)*	14 *(100 %)*	10 *(97,3 %)*	16 *(99,1 %)*	12 *(96,4 %)*
k-NN	1 *(100 %)*	1 *(100 %)*	12 *(100 %)*	3 *(97,3%)*	3 *(100 %)*	9 *(95,5 %)*

Results in Table 2 show that binary classifiers achieve their maximum accuracy with very few features, in that outperforming standard multi-classifiers (Fig. 1). In particular, only one gene is sufficient to perfectly discriminate T-ALL and E2A-PBX1 sub-types, irrespective of the adopted classification algorithm. No error occurs in TEL-AML1 classification too, even if a higher number of features (12-14) is required. Most misclassifications occur in discriminating BCR-ABL and Hyperdyp>50, suggesting a less sharp genetic characterization of these sub-types.

Finally, Table 3 summarizes results when the proposed "knowledge-based" approach is applied, i.e. when the multi-class problem is treated directly using a "globally optimal" subset of features that results by joining all the optimal subsets previous generated.

Table 3. Optimal number of features and classification accuracy for SVM, NB and k-NN knowledge-based multi-classifiers.

	Number of features	*Accuracy*
SVM	31	96,4 % (91,1 ÷ 98,6%)
NB	54	86,6 % (79,1 ÷ 91,7%)
k-NN	29	93,8 % (87,7 ÷ 97,0%)

To approximately evaluate the statistical significance of the above accuracy measure, we constructed a binomial confidence interval as suggested in [19]; specifically, last column of Table 3 reports, in brackets, the intervals relative to a 95% confidence level. As shown by these results, the proposed "knowledge-based" approach turns out to be much more accurate than standard multi-classifiers depicted in Fig.1. Indeed, the accuracy of a standard SVM multi-classifier is only 80,4% (72,1 ÷ 86,7%) with 30 features, while a knowledge-based multi-classifier achieves 96,4% (91,1 ÷ 98,6%) with 31 features, as witnessed by confusion matrices in Tables 4. The statistical significance of the observed difference has been further proved by the application of the McNemar'test [20], which is recommended [19][21] for the cases where the learning algorithm is run only once (i.e. without any form of resampling).

Table 4. Confusion matrices of standard (**A**) and knowledge-based (**B**) SVM multi-classifiers.

A. *Standard multi-classification*

		predicted class						
		a	**b**	**c**	**d**	**e**	**f**	**g**
actual class	**a**	1	0	0	0	3	0	2
	b	0	9	0	0	0	0	0
	c	0	0	12	0	10	0	0
	d	0	0	1	4	1	0	0
	e	2	0	1	0	22	0	2
	f	0	0	0	0	0	15	0
	g	0	0	0	0	0	0	27

30 features, accuracy = ***80,4%***

B. *"Knowledge-based" multi-classification*

		predicted class						
		a	**b**	**c**	**d**	**e**	**f**	**g**
actual class	**a**	5	0	1	0	0	0	0
	b	0	9	0	0	0	0	0
	c	0	0	21	0	1	0	0
	d	0	0	0	6	0	0	0
	e	0	0	1	0	26	0	0
	f	0	0	0	0	0	15	0
	g	0	0	0	0	1	0	26

31 features, accuracy = ***96,4%***

a = BCR-ABL, **b** = E2A-PBX1, **c** = Hyperdip>50,
d = MLL, **e** = OTHERS, **f** = T-ALL, **g** = TEL-AML1

Analogously, the performance of NB and k-NN multi-classifiers (whose confusion matrices are omitted for sake of space) greatly improves when the knowledge-based approach is adopted. Specifically, NB achieves 55,4% (46,2 ÷ 64,3%) with 50 features selected in the standard way, while its accuracy is 86,6% (79,1 ÷ 91,7%) when an optimal set of 54 features is selected according to the proposed knowledge-based heuristic. Similarly, the standard k-NN multi-classifier achieves 82,1% (74,0 ÷ 88,1%) with 30 features, while the accuracy of the knowledge-based k-NN multi-classifier is 93,8% (87,7 ÷ 97,0%) with 29 features.

5 Discussion and Concluding Remarks

The presented learning strategy has been validated by different experiments. The first experiment (Fig.1), here referred to as standard multi-classification, highlights the intrinsic weakness of a feature selection process where the correlation of each gene with all class values is simultaneously measured. Related difficulties are also reported by recent literature [17] [18] [22].

The second experiment (Fig. 2 and Table 2) reveals that a 1-vs-r decomposition is much more effective than a single multi-classifier. Our results can be compared with [4] where six different heuristics for feature selection are explored by learning NB, SVM and k-NN binary classifiers on the ALL dataset. In [4] the best feature selection heuristic ranks attributes according to their entropy and selects features "having an entropy value less than 0.1 if these exist, or the 20 features with the lowest entropy values otherwise", up to a maximum of 20 features for each binary classifier. The resulting NB, SVM and k-NN models are then combined according to a hierarchical scheme and respectively misclassify 7, 5 and 4 samples. It is important to observe that in [4] the threshold of 20 used to cut off top-ranked features is an arbitrary number, whereas our learning strategy enables to find, for a given ranking method and a given classification algorithm, the "optimal" cut off point.

The paper goes further in this improving classification by proposing a new approach referred to as "knowledge-based" multi-classification. As shown by the third experiment (Table 3 and Table 4), this approach turns out to be much more accurate than standard multi-classifiers, with the same number of features. [5] is the only work where a similar heuristic is adopted, but on a different case study. However, in [5] the global subset of features is built by taking an equal number of genes for each class, hence requiring more genes than in our approach. Indeed, as witnessed by our analysis, the optimal number of features can be different for different class values, suggesting a different strength in genetic characterization of cancer sub-types.

Our approach is tailored to capture genetic sub-type specificity, enabling to sensibly reduce the total number of features involved in multi-classification and returning only the features that are really relevant. As future extension, we plan to validate our strategy on different multi-cancer datasets, to obtain more insights on genetic mechanisms underlying cancer sub-type differentiation. In particular, the effectiveness of the proposed knowledge-based approach will be further investigated in the context of poorly differentiated cancer-sub-types, where the identification on the genes responsible of each sub-type is of crucial importance.

Acknowledgments

The authors are very grateful to Robert Holte for the useful suggestions and for the freely distribution of the cost curve software tool that has been used for the cost based analysis.

References

1. Golub T.R., et al.., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286, 531-537, 1999.
2. Alizadeh A.A., et al., Distinct Types of Diffuse Large B-cell Lymphoma identified by gene expression profiling, Nature 403, 503-511, 2000.
3. Guyon I., Weston J., Barnhill S., and Vapnik V., Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46 (1-3): 389 – 422, 2002.
4. Liu H. et al., A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns, Genome informatics 13: 51-60, 2002.
5. Piatetsky-Shapiro G., et al., Capturing Best Practice for Microarray Gene Expression Data Analysis, SIGKDD'03 (August 24-27, 2003, Washington, USA).
6. Friedman N., Geiger D., Goldszmidt M., Bayesian Network Classifiers, Machine Learn-ing, 29: 131-161, 1997.
7. Vapnik V., Statistical Learning Theory, Wiley-Interscience, New York, NY, USA, 1998.
8. Cover T.M. and Hart P.E., Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 13:21-27 (1967).
9. Tao L. et al., A Comparative Study on Feature Selection and Multiclass Classification Methods for tissue classification based on gene expression, Bioinformatics, Vol 20 no.15, 2004.
10. Simon H., Supervised analysis when the number of candidate features greatly exceeds the number of cases, SIGKDD Explorations, vol. 5, issue 2, pp. 31-36, 2003.
11. Somorjai R. et al., Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, cavets, cautions, Bioinformatics, vol. 19, no. 12, 2003.
12. Forman G., An extensive empirical study of feature selection metrics for text classification, JMLR, 3:1289-1306, 2003.
13. Drummond C., Holte R.C., Cost Curves: An improved Method for Visualizing Classifier Performance, Machine Learning Journal, Vol. 65, Number 1, October 2006.
14. Drummond C., Holte R.C., Explicitly Representing Expected Cost: An Alternative to ROC Representation, in Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining, pp. 198-207, 2000.
15. Witten I. H., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, second edition, Elsevier, 2005
16. St. Jude Children's Research Hospital. http://www.stjuderesearch.org/data/ALL1/.
17. Mukherjee S., Classifying Microarray Data Using Support Vector Machines, Understanding And Using Microarray Analysis Techniques: A Practical Guide. Kluwer Academic Publishers, Boston, MA, 2003.
18. Statnikov A. et al., A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics, Vol. 21 no. 5 2005.
19. Dietterich T.G., Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Neural Computation, 10:1895-1924, 1998.
20. Everitt B.S., The analysis of contingency tables, Chapman and Hall, London, 1977.
21. Demsar J., Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research 7, 1–30, 2006.
22. Bosin A., Dessì N., Pes B., A Cost-Sensitive Approach to Feature Selection in Micro-Array Data Classification, in Proceedings of WILF'07 (Portofino, Italy, July 2007).