Deterministic classifiers accuracy optimization for cancer microarray data

Vânia Rodrigues[1] and Sérgio Deusdado[2]

[1] USAL – Universidad de Salamanca, 37008 Salamanca, España
2 CIMO – Centro de Investigação de Montanha, Instituto Politécnico de Bragança, 5301-855 Bragança, Portugal
sergiod@ipb.pt

Abstract. The objective of this study was to improve classification accuracy in cancer microarray gene expression data using a collection of machine learning algorithms available in WEKA. State of the art deterministic classification methods, such as: Kernel Logistic Regression, Support Vector Machine, Stochastic Gradient Descent and Logistic Model Trees were applied on publicly available cancer microarray datasets aiming to discover regularities that provide insights to help characterization and diagnosis correctness on each cancer typology. The implemented models, relying on 10-fold cross-validation, parameterized to enhance accuracy, reached accuracy above 90%. Moreover, although the variety of methodologies, no significant statistic differences were registered between them, at significance level 0.05, confirming that all the selected methods are effective for this type of analysis.

Keywords: Classification, cancer, microarray, datamining, machine learning.

1 Introduction

Accurate prediction and prognostic risk factor identification are essential to offer appropriate care for patients with cancer. Therefore, it is necessary to find biomarkers for the identification of different cancer typologies. Currently, with the evolution of microarray technology, it is possible for researchers to classify the types of cancer based on the patterns of gene activity in the tumor cells. For this purpose, statistical methods and machine learning techniques can be employed, such as classification methods to allow the assignment of class labels to samples with unknown biological condition, feature selection to identify informative genes and, additionally, clustering methods to discover classes of related biological samples. Detailed reviews on the technology and statistical methods often used in microarray analysis are presented in [1–3]. The objective of this work was to employ machine learning algorithms to analyze and classify gene expression data from cancer tissue samples provided by microarrays. The developed work included the use of three publicly available gene microarray datasets, described in the methodology section, on which the methods were tested and the performance assessed in order to compare the results with the best achievements published in the literature.

This paper has been structured as follows. After a brief introduction, section 2 describes the context and the state of art. Section 3 explains the methodology followed in this study, the procedures, the gene microarray datasets, the classification methods implemented as well as the optimal parameters adopted, concluding with the performance assessment of the classification methods. Experimental work using WEKA datamining workbench, the obtained results are discussed in Section 4 and the conclusions are presented in Section 5.

2 Background

2.1 Microarray Technology

In the last two decades microarrays were widely used to study gene expression. Main microarray technology includes Affymetrix [4] and Illumina [5] platforms. Other important microarray manufacturers are Exiqon [6], Agilent [7] or Taqman [8]. Gene microarray technology rest on the ability to deposit many (tens of thousands) different DNA sequences on a small surface, often referred to as a "chip". The different DNA fragments are arranged in rows and columns, in order to identify the location and distinguish the level of expression of each fragment on the array. Microarrays allow the measurement at expression level of a large simultaneous number of genes. Initially, the gene expression values are obtained by means of microscopic DNA spots attached to a solid surface which have followed a hybridization process [9], then it is possible to read the expression values with a laser, and subsequently store the quantification levels in a file.

Microarray technology has been extensively used by the scientific community. Accordingly, there has been a lot of data generation related to gene expression. This data is scattered and not easily available for public use. The National Center of Biotechnology Information (NCBI) organized, integrated and made available microarray data through a web service, the Gene Expression Omnibus or GEO. GEO is a data repository facility which includes data on gene expression from various sources.

Microarray technology possesses extensive applications in the medical field, mainly regarding diagnostics and prognostics. In this context, microarrays are widely used to know the state of a disease, type of tumor and other important factors for the patient treatment [10].

Considering disease diagnosis, it allows researchers to study and gather knowledge about many diseases such as mental illness, heart diseases, infectious disease and particularly the study of cancer [11]. They are also used in pharmacology response, which consists of the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients, and additionally in the toxicological research to establish a correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to toxicants [12].

2.2 Deterministic Classifiers Overview

Kernel Logistic Regression (KLR) model is a statistical classifier [13] that generates a fit model by minimizing the negative log-likelihood with a quadratic penalty using the Broyden-Fletcher-Goldfard-Shanno (BFGS) optimization [14].

Support Vector Machine (SVM) algorithm is a discriminative classifier that tries to find an optimal hyperplane with maximal margin [15, 16]. SVM was developed for binary classification problems, although extensions to the technique have been made to support multi-class classification and regression problems [9]. This classifier is a state of the art classification system. In Cao et al. [17] SVM was applied in two-class datasets (Leukemia and colon Tumor) and also in multi-class datasets, proposing a novel fast feature selection method based on multiple SVDD (Support Vector Data Description). [18] focused on supervised gene expression analysis of cancers microarrays: prostate cancer, lymphoma and breast cancer. SVM algorithm is implemented in practice using selectable kernel functions. The kernel defines the similarity or a distance measure between new data and support vectors. The dot product is the similarity measure used for linear SVM or a linear kernel because the distance is a linear combination of the inputs. Other kernels can be used to transform the input space into higher dimensions such as Polynomial Kernel and a Radial Kernel. WEKA includes a derivative of SVM, the SMO implementation using sequential minimal optimization, described in [19].

The Stochastic Gradient Descent (SGD) algorithm implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties for classification. Available loss functions include the Hinge loss (linear support vector classifier), Log loss (logistic regression), Squared loss (least squares linear regression), Epsilon-insensitive loss (support vector regression) and Huber loss (robust regression).

Decision tree classifiers recursively partition the instance space using hyperplanes that are orthogonal to axes. The model is built from a root node which represents an attribute and the instance space split is based on function of attribute values (split values are chosen differently for different algorithms), most frequently using its values. Then, each new sub-space of the data is split into new sub-spaces iteratively until an end criterion is met and the terminal nodes (leaf nodes) are each assigned a class label that represents the classification outcome (the class of all or majority of the instances contained in the sub-space). Setting the right end criterion is very important because trees that are too large can be overfitted and small trees can be underfitted, suffering a loss in accuracy in both cases. Most of the algorithms have a mechanism built in that deals with overfitting; it is called pruning.

Each new instance is classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path [20, 21].

Although there are several methodologies to implement decision tree classifiers, for instance: SimpleCart, BFTree, FT, J48, LADTree, LMT and REPTree, the literature refers Logistic Model Trees (LMT) as the most efficient to classify microarray datasets [22].

A Logistic Model Trees (LMT) is a classification algorithm that integrates decision tree induction with logistic regression, building the logistic regression (LR) models at the leaves by incrementally refining those constructed at higher levels in the tree [23]. In the logistic variant, the LogitBoost algorithm [24] is used to produce an LR model at every node in the tree; the node is then split using the C4.5 criterion. Boosting works by sequentially applying a classification algorithm to reweighted versions of the training data and then taking a weighted majority vote of the sequence of classifiers thus produced. For many classification algorithms, this simple strategy results in a dramatic improvement in performance.

3 Methods

3.1 Experimental procedures

The experimental work was based on the WEKA, version 3.8.3, a datamining workbench publicly accessible at: www.cs.waikato.ac.nz/ml/weka/. After data preparation and method selection (considering accuracy above 90%), using the explorer module, the module experimenter was used to automate experiments to achieve multiple classifiers comparison, testing with Paired T-Tester (Corrected). Prior to the experimental analysis, the microarray datasets were pre-processed and normalized on the interval [0,1]. Successively, an external ten-fold cross-validation was performed, which randomly divides each dataset into ten equal parts. In each validation, one of them is taken as the testing set, and the others nine parts are used as the training set. The training and test data do not overlap each other to assure an unbiased comparison. Three functions based classifiers (KLR, SMO and SGD algorithms) and one decision tree classifier (LMT algorithm) were used as base learners. To compare classification performance, we created an experiment that ran 10 times several schemes (all classifications methods used) against each dataset with 10-fold cross-validation. Subsequently, we used literature mining analysis results to compare the performance of the methods applied in these microarrays datasets. These set of experiments were conducted on a computer with an Intel Core i7-5500U CPU 2.40 GHz processor, with 8.00 GB RAM.

3.2 Datasets

Three publicly available microarray datasets from different cancer typologies were used to test the classification methods, namely Leukemia, Lymphoma and Prostate datasets. The Leukemia datasets were obtained online from http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63, and were published as part of the experimental work in [25]. The Lymphoma and Prostate datasets were obtained online from http://ico2s.org/datasets/microarray.html, and were published as part of the experimental work in [26, 27].

All of them are two-class datasets. In Leukemia (a) dataset are present two types of leukemia: Acute Lymphoblastic (ALL) and Acute Myeloid Leukemia (AML). The leukemia dataset was analyzed in two different versions, the original composed by 52 samples and 12582 genes and a reduced version, composed by 28 samples keeping the same features. The goal for this subdivision was to test if the number of samples influences the prediction results.

Lymphoma dataset consists of 58 Diffuse large B-cell lymphoma samples vs. 19 follicular lymphoma samples.

Prostate cancer datasets consist of 52 tumor samples vs. 50 controls.

The composition details of the used datasets are shown below in Table 1.

Table 1. Used datasets characterization.

Dataset	2 Classes	Genes	References
Leukemia(a)	24 – 28 (ALL –AML)	12582	[25]
Leukemia(b)	14 – 14 (ALL –AML)	12582	[25]
Diffuse large B-cell lymphoma	58 – 19	2647	[26]
Prostate cancer	52 - 50	2135	[27]

3.3 Classification methods parameterization

KLR method

The parameters optimized were support vector with different types of kernel function. The penalty parameter λ with smaller values conjugated with different types of kernel functions was tested. The linear kernel function with λ=0.001 presented the smaller mean absolute error.

SVM method

We used the SMO classifier, a specific efficient optimization algorithm used to enhance the SVM performance. The model contains the complexity parameter C that influences the number of support vectors, we set C to 0.5. If C is lower, the more sensitive the algorithm becomes to training data, leading to higher variance and lower bias. With a higher C, the algorithm becomes less sensitive to the training data, in this case we obtain lower variance and higher bias. We tested polynomial functions of different degrees with different filters types without good results and, consequently a polynomial kernel without filter was selected, having set the exponent to 0.5.

SGD method

SGD is an optimization method for unconstrained optimization problems. It approximates the true gradient by considering a single training example at a time. The algorithm works iteratively over the training examples and for each example updates the model parameters. The learning rate parameter was optimized setting a small value (0.0001) affecting the learning binary class SVM.

LMT method

LMT consists of a tree structure that is made up of a set of inner or non-terminal nodes N and a set of leaves or terminal nodes T. Considering S the whole instance space, spanned by all attributes that are presented in the dataset. Then the tree structure gives a disjoint subdivision of S into regions S_t, and every region is represented by a leaf in the tree:

$S = \cup_{t \in T} S_t \quad S_t \cap S_{t'} = \emptyset$ for $t \neq t'$

The model represented by whole LMT is given by $F_j(x) = \alpha_0^j + \sum_{k=1}^{m} \alpha_{v_k}^j . v_k$

If $\alpha_{v_k}^j = 0$ for $v_k \notin V_t$. The model of LMT is then given by

$$f(x) = \sum_{t \in T} f_t(x) . I(x \in S_t)$$

Where $I(x \in S_t)$ is 1 if $x \in S_t$ and 0 otherwise. Considering the WEKA implementation of LMT, we used the fast regression heuristic that avoids cross-validating the number of LogitBoost iterations at every node [23]. LMT employs the minimal cost-complexity pruning mechanism to produce a compact tree structure.

3.4 Performance Evaluation

In this study, we trained the classifiers to predict outcomes of cancer microarray datasets contained positive samples and control samples. The evaluation measures to evaluate the classifiers [28, 29], include classification accuracy (ACC), *i. e.*, the ratio of the true positives and true negatives obtained by the classifier over the total number of instances in the test dataset, defined as:

$$ACC = \frac{TN + TP}{TP + FP + FN + TN}$$

Kappa (k) coefficient is a statistical measure for qualitative (categorical) items as given by:

$$k = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy}$$

Kappa coefficient is interpreted using the guidelines outlined by Landis and Koch (1977), where strength of the k is interpreted in the flowing manner: 0.01-0.20 slight; 0.21-0.40 fair; 0.41-0.60 moderate; 0.61-0.80 substantial; 0.81-1.00 almost perfect [30].

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of prediction, without considering their direction [31]. It is given by:

$$MAE = \frac{\sum_{i=1}^{n} |predicted_i - \text{actual}_i|}{total\ predictions}$$

Precision (PRE), it is also called the Positive Predictive Values (PPV), is the proportion of the true positives against the true positives and false positives, as given by equation:

$$PRE = \frac{TP}{TP + FP}$$

Recall (REC) also called sensitivity and hit rate, is the proportion of the true positives against true positives and false negatives, as given by the equation:

$$REC = \frac{TP}{TP + FN}$$

F-measure, it is also called F score, is the harmonic mean of precision and recall which is given by the equation:

$$f_{measure} = \frac{2 * PRE * REC}{PRE + REC}$$

ROC stands for Receiver Operating Characteristic. It´s created by plotting the True Positives rates vs False Positives rates. It is also exploited to evaluate the performance of classifiers as Area Under ROC.

4 Results and Discussion

For each dataset in this study, the results of the classifiers estimation performance are presented in Table 2. These results are expressed on average, considering the 10 times that each test was repeated.

Table 2. Results achieved by algorithms with 10-fold cross-validation.

Dataset	Classifier	ACC(%) (st. dev.)	k (st. dev.)	MAE (st. dev.)	Recall (st. dev.)	F-Measure (st. dev.)	Area Under ROC (st. dev.)
Leukemia (a)	KLR	100	1	0.01(0.01)	1	1	1
	SVM	100	1	0.00	1	1	1
	LMT	97.33(6.67)	0.94(0.14)	0.13(0.08)*	0.94(0.14)	0.96(0.09)	1
	SGD	100	1	0.00	1	1	1
Leukemia (b)	KLR	98.17(8.17)	0.95(0.20)	0.02(0.06)	1	0.99(0.06)	1
	SVM	96.67(10.86)	0.93(0.24)	0.03(0.11)	1	0.97(0.09)	0.96(0.12)
	LMT	100	1	0.14(0.04)*	1	1	1
	SGD	96.33(11.26)	0.92(0.26)	0.04(0.11)	1	0.97(0.09)	0.96(0.13)
Diffuse large B-cell lymphoma	KLR	95.50(6.90)	0.87(0.22)	0.05(0.06)	0.97(0.07)	0.97(0.04)	0.98(0.05)
	SVM	98.70(3.94)	0.97(0.09)	0.01(0.04)	0.98(0.05)	0.99(0.03)	0.99(0.03)
	LMT	92.25(10.35)	0.77(0.31)	0.09(0.09)	0.96(0.09)	0.95(0.07)	0.94(0.15)
	SGD	98.20(4.84)	0.96(0.11)	0.02(0.05)	0.98(0.06)	0.99(0.04)	0.99(0.04)
Prostate cancer	KLR	89.18(8.61)	0.78(0.17)	0.11(0.08)	0.89(0.13)	0.89(0.09)	0.96(0.06)
	SVM	92.33(8.27)	0.85(0.17)	0.08(0.08)	0.96(0.09)	0.93(0.08)	0.92(0.08)
	LMT	90.76(8.83)	0.81(0.18)	0.15(0.07)	0.92(0.12)	0.91(0.09)	0.95(0.07)
	SGD	90.18(8.21)	0.80(0.16)	0.10(0.08)	0.93(0.11)	0.90(0.08)	0.90(0.08)*

* Statistically different at significance level 0.05

The experiment was configured using KLR as the referential for all datasets, the results registered in Table 2 correspond to the comparison between the different classifiers considering the used evaluation measures.

Leukemia

On leukemia (a) dataset, the prediction results of KLR, SVM, and SGD are 100% ACC followed by LMT with ACC of ≈97%. Kappa coefficient results of KLR, SVM and SGD indicates a perfect agreement (1) between the classification and the true classes, having a LMT result almost perfect (≈0.94). F-measure and Area under ROC presents results nearly 1 on all methods, which indicates the good performance of the classification models used. These results are similar because there are not differences statistically significant between them. On the contrary, MAE is statistically better in KLR than LMT but not statistically significant differences on SVM and SGD, the same is verified on leukemia (b) dataset. On the leukemia (b) dataset, LMT achieves 100% ACC. On the contrary, SVM and SGD achieved ACC ≈96.67% and ACC≈96.33%, respectively, but they do not present differences statistically significant as well. In comparison, the cross-validation results reported in literature for this

datasets [17], presented results of SVM methods using kernel functions achieving results of average recall equal to 93.93%. In the cited work, the authors optimized the method to achieve the best results equal to 96.43% of average recall, however their study used a smaller number of features. In [32], leukemia datasets with smaller number of features presented the maximum ACC results equal to 97.43% using the RBF Network classifier.

Diffuse Large B-cell Lymphoma

For the Diffuse Large B-cell lymphoma dataset, the ACC result of SVM is 98.70% and SGD is 98.20%, followed by KLR with 95.50% and then LMT with 92.25%. K results of LMT revealed a substantial agreement (0.77) between the classifications and the true classes, whereas the other classifiers presented almost a perfect agreement. On the lymphoma dataset there are no statistical differences on MAE among the four classifiers. F-measure and Area under ROC indicates an excellent prediction of the classification methods ($\geq$0.9). In literature, the results reported for this dataset [18] achieved 95% ACC using a higher number of features. The same datasets were analyzed in [32] and presented the best outcome prediction having ACC equal to 92.45%. In [33] was achieved 95.7% ACC, also with high number of features.

Prostate Cancer

For the prostate cancer dataset, the best ACC result was 92.33% and was obtained with SVM. LMT, SGD and KLR achieved very close results, respectively 90.76%, 90.18% and 89.18%. KLR and SGD Kappa coefficient results indicate substantial agreement $\approx$0.78, 0.80, respectively, between these classifiers and the true classes, while in SVM and LMT presented almost perfect agreement, with k equal to 0.85 and 0.81, respectively. However, all k do not present differences statistically significant. Analyzing the results of Area under ROC, there is a significant statistical difference between KLR (0.96) and SGD (0.90). On the contrary, LMT (0.95) and SVM (0.92) are not statistically different. F-measure results were very close to 1, which means good performance of all classifiers implemented. Comparatively with our work, in [18] was used a higher number of features in the cross-validation results for this dataset, achieving 94% ACC. In the papers published by [34] and [33] were obtained 94.6% and 93.4% ACC, respectively. In [32] the best outcome prediction measured by ACC was equal to 95.20% using the SVM classifier.

5 Conclusions

All the classifiers involved in this study (KLR, SVM, LMT, SGD) presented good performance in gene expression analysis on cancer microarrays data, proving to be effective and reliable in this type of data. The classifiers performance, except for the measures MAE and Area under ROC, in some schemes, are not statistically different. The developed experimental work achieved better or close-to-best performance by comparison with other methods applied on the same datasets in the literature.

References

1. Allison, D.B., Cui, X., Page, G.P., Sabripour, M.: Microarray data analysis: from disarray to consolidation and consensus. Nat. Rev. Genet. 7, 55–65 (2006).
2. Hoheisel, J.D.: Microarray technology: beyond transcript profiling and genotype analysis. Nat. Rev. Microbiol. 7, 200–210 (2006).
3. Quackenbush, J.: Computational analysis of microarray data: Computational genetics. Nat. Rev. Genet. 2, 418–427 (2001).
4. Talloen, W., Göhlmann, H.: Gene Expression Studies Using Affymetrix Microarrays. Chapman and Hall/CRC (2009).
5. Illumina: Illumina Genes Expression arrays, (2009).
6. Exiqon: Exiqon Genes Expression arrays, (2009).
7. Zahurak, M., Parmigiani, G., Yu, W., Scharpf, R.B., Berman, D., Schaeffer, E., Shabbeer, S., Cope, L.: Pre-processing Agilent microarray data. BMC Bioinformatics. 8, 142 (2007).
8. Taqman: Taqman Genes Expression arrays, (2009).
9. Castillo, D., Gálvez, J.M., Herrera, L.J., Román, B.S., Rojas, F., Rojas, I.: Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling. BMC Bioinformatics. 18, (2017).
10. Kaliyappan, K., Palanisamy, M., Govindarajan, R., Duraiyan, J.: Microarray and its applications. J. Pharm. Bioallied Sci. 4, 310 (2012).
11. Raghavachari, N.: Microarray Technology: Basic Methodology and Application in Clinical Research for Biomarker Discovery in Vascular Diseases. In: Freeman, L.A. (ed.) Lipoproteins and Cardiovascular Disease. pp. 47–84. Humana Press, Totowa, NJ (2013).
12. Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., Weinstein, J.N.: A gene expression database for the molecular pharmacology of cancer. Nat. Genet. 24, 236–244 (2000).
13. Wahba, G., Gu, C., Wang, Y., Chappell, R.: Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. In: Computational Learning Theory and Natural Learning Systems. pp. 133–162. MIT Press (1995).
14. Smith, B., Wang, S., Wong, A., Zhou, X.: A Penalized Likelihood Approach to Parameter Estimation with Integral Reliability Constraints. Entropy. 17, 4040–4063 (2015).
15. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. pp. 144–152. ACM Press, Pittsburgh, Pennsylvania, United States (1992).
16. Vapnik, V.N.: Statistical learning theory. Wiley, New York (1998).
17. Cao, J., Zhang, L., Wang, B., Li, F., Yang, J.: A fast gene selection method for multi-cancer classification using multiple support vector data description. J. Biomed. Inform. 53, 381–389 (2015).
18. Glaab, E., Bacardit, J., Garibaldi, J.M., Krasnogor, N.: Using Rule-Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data. PLoS ONE. 7, e39932 (2012).
19. Schölkopf, B., Burges, C.J.C., Smola, A.J. eds: Advances in kernel methods: support vector learning. MIT Press, Cambridge, Mass (1999).
20. Polaka, I., Tom, I., Borisov, A.: Decision Tree Classifiers in Bioinformatics. Sci. J. Riga Tech. Univ. Comput. Sci. 42, 118–123 (2010).
21. Rokach, L., Maimon, O.: Data mining with decision trees: theory and applications. World Scientific, Hackensack, New Jersey (2015).

22. Li, Y., Wang, N., Perkins, E.J., Zhang, C., Gong, P.: Identification and Optimization of Classifier Genes from Multi-Class Earthworm Microarray Dataset. PLoS ONE. 5, e13715 (2010).
23. Landwehr, N., Hall, M., Frank, E.: Logistic Model Trees. Mach. Learn. 59, 161–205 (2005).
24. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann. Stat. 28, 337–407 (2000).
25. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30, 41–47 (2001).
26. Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 8, 68–74 (2002).
27. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 1, 203–209 (2002).
28. Saito, T., Rehmsmeier, M.: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 10, e0118432 (2015).
29. Tharwat, A.: Classification assessment methods. Appl. Comput. Inform. (2018).
30. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Int. Biom. Soc. 33, 159–174 (1977).
31. Sammut, C., Webb, G.I. eds: Encyclopedia of Machine Learning. Springer US, Boston, MA (2010).
32. Dagliyan, O., Uney-Yuksektepe, F., Kavakli, I.H., Turkay, M.: Optimization Based Tumor Classification from Microarray Gene Expression Data. PLoS ONE. 6, e14579 (2011).
33. Wessels, L.F.A., Reinders, M.J.T., Hart, A.A.M., Veenman, C.J., Dai, H., He, Y.D., Veer, L.J. v.: A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics. 21, 3755–3762 (2005).
34. Li Shen, Eng Chong Tan: Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data. IEEE/ACM Trans. Comput. Biol. Bioinform. 2, 166–175 (2005).