ISSN 1870-4069

Prediction of Cancer using Microarrays Analysis by Machine Learning Algorithms

José Luis Velázquez-Rodríguez[1], Yenny Villuendas-Rey[2], Cornelio Yáñez-Márquéz[1]

[1] Instituto Politécnico Nacional, Centro de Investigación de Computación, Mexico
[2] Instituto Politécnico Nacional, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Ciudad de México, Mexico

Abstract. The analysis of microarrays that contain information on biomolecules related to different types of cancer is one of the current issues in international scientific research due to the impact it has on public health worldwide. The advances in this scientific research route have been impressive; the different international research groups have applied sophisticated algorithms for machine learning, data mining and related branches with the aim of finding solutions to this problem. The present article contains a study of several the classification algorithms used in the literature, and their application for the prediction of cancer using microarrays analysis. More in detail, we tested six classification models, over microarrays data. The application of the supervised classification algorithms was done over the Weka 3 Software environment, using the Leave One Out validation scheme. In addition, a nonparametric statistical test (the Friedman test) identified the significant differences in the performance of the algorithms, according to the experimental results obtained. The analysis of the hypothesis tests of the experimental results indicates that the Support Vector Machine models outperform others for the prediction of cancer.

Keywords: classification models, microarrays, cancer prediction, computational intelligence.

1 Introduction

The analysis of microarrays that contain information on biomolecules related to different types of cancer has great potential for medical diagnostic tests [1]. A microarray contains the gene expression of an individual, which is known through clinical diagnoses if it suffers, or not, a certain type of cancer. It is possible to train Machine Learning algorithms in order to learn the relationship between the levels of gene expression of a patient and their condition as to whether or not they suffer from certain type of cancer; in other words, it is possible to diagnose whether an individual suffers, or not, cancer, through the intelligent analysis of the microarray that contains the information of the level of expression of their genes [2, 3].

The advances in this scientific research route have been impressive [4]; the different international research groups have applied sophisticated algorithms for machine learning, data mining and related branches with the aim of finding solutions to this problem. They have been applied from effective algorithms such as random forests [5],

through extensive comparative studies of dozens of gene classifiers [6], to the application of next-generation neuronal models [7], guaranteeing with this intense activity the validity and current status of this scientific research topic.

In this paper, we review several machine learning algorithmic solutions proposed in recent years to analyze microarray data (section 2). We study some of the classification algorithms used in the literature, and their application to the diagnosis of cancer using microarray data (section 3). In addition, we used 10 microarray datasets (section 4) to test all the classification models. The different datasets have imbalanced data, mixed categorical and numerical attributes. Then, the numerical experiments performed in the comparison between the different classification algorithms applied for different types of cancer diagnoses were made. Finally, we offer the conclusion and some lines of future work (section 5).

2 Background

We present some results that have been reported by international research groups, derived from applying different algorithms of machine learning, data mining and related branches, with the purpose of finding solutions to the problem represented by microarray analysis. that contain information on biomolecules related to different types of cancer.

Korucuoglu et al. [8] proposed a method based on Bayesian Networks, where biological routes are taken for a Bayesian Network and each path is qualified for a method equivalent to Bayesian-Dirichlet. The method was tested with datasets of cancer cells microarrays.

Tan et. al. [9] claimed that by using a method to classify microarrays data it is not only necessary to obtain good results, but also that the results can be easily interpreted, and therefore he proposed a method based on Decision Trees for the analysis of carcinogenic samples, which generates precise and easy to interpret rules.

The genome expression patterns generally consist of thousands of genes, and it is necessary to extract the most significant genes. Cho et. al. [10] obtained significant genes from representative vectors, and proposed a set of neural networks that were trained with this set of significant genes to classify a dataset divided into three classes: leukemia, colon and B-cell lymphoma.

Guyon et al. [11] constructed an efficient classifier for genetic analysis and drug use, using training examples for cancer treatment and using data from normal patients. They propose the use of Support Vector Machines for the elimination of recursive characteristics.

Hu et. al. [12] observed that some types of cancer are related to each other while there are others that are very differentiable. They use a dataset with information on different types of cancer and a clustering method.

In the work of Ruiz et. al. [13] a heuristic is proposed for the selection of attributes to datasets of genetic expressions. His method is based on the use of statistical significance.

Despite the above proposals, there is a lack of experimental comparisons of several classification algorithms over multiple microarray datasets, for cancer pre-diagnosis. In this paper, we address this gap.

3 Datasets and Algorithms

In this section, a summary of applied classification algorithms and datasets related to microarray are presented for the classification of cancer.

3.1 Datasets Related to Microarray

The used datasets include information about different types of cancer, such as leukemia, colon cancer, adenocarcinoma, brain tumor, lymphoma, and others. The datasets used in this paper were taken from online information provided by different repositories. We used the microarray data related to standard classification datasets. In the following, a description of each of the datasets used is shown.

Leukemia dataset: It is a dataset based on the monitoring of gene expressions by DNA microarrays in humans with Acute Leukemia. Distinguish between two types of conditions, between Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). It contains 38 samples of which 27 are ALL and 11 AML, each sample in turn contains 3051 expressions [14].

Colon dataset: The gene expression levels in colon tissue samples are shown in this dataset. It contains 62 patients of which 40 have tumors and 22 are normal. Each sample contains 2,000 human genes [15].

Adenocarcinoma dataset: The original dataset contained 16063 Affymetrix chip genes, but they were reduced to 9.868 by the author of [16]. The Adenocarcinoma dataset includes information on 76 patients, of which 64 have primary tumors and 12 metastatic tumors. The data of 9,868 genes are shown for each patient [16].

Brain Tumor dataset: This dataset contains 42 profiles of genetic expressions of microarrays, divided into five different types of tumors of the central nervous system, which are, 10 medulloblastoma, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumor, eight primitive neuroectodermal tumors and four of human cerebellum. Each profile contains 5,597 genes [15].

Breast Cancer 1 dataset: This dataset contains information on 78 people who suffered from breast cancer [17]. It includes 34 patients who developed metastases within the first five years (class 1) and 44 patients who remained disease-free for more than five years (class 2).

Breast Cancer 2 dataset: This dataset complies with the two classes of the Breast Cancer 1 dataset previously described, and a third additional class which includes 18 patients with germline mutations in the BRCA1 gene (class 3). This dataset contains 96 patterns divided into three classes [17].

Lymphoma dataset: This dataset contains the gene expression levels of the three most prevalent adult lymphoid neoplasms: 42 samples of diffuse large B-cell lymphoma (class 1), nine observations of follicular lymphoma (class 2) and 11 cases of lymphocytic leukemia chronic (class 3). The total samples are 62 and each contains 4,026 genetic expressions (attributes) [15].

National Cancer Institute (NCI) dataset: The data come from complementary DNA microarrays. This dataset contains 61 samples that can be divided into eight different tumor types samples: seven of breast cancer, five of central nervous system cancer, seven of colon cancer, six of leukemia, eight of melanoma, nine of lung carcinoma, six of ovarian and nine of kidney tumors. Each sample contains 5,244 genes [18].

Prostate cancer dataset: This dataset contains the samples of 52 prostate tumors and 50 non-tumor samples, giving a total of 102 samples and each sample contains 6,033 genes [15].

SRBCT dataset: This dataset contains gene expression profiles for classifying small round blue cell tumors in childhood (SRBCT) and contains four classes (12 samples of neuroblastoma, 20 of rhabdomyosarcoma, eight of non-Hodgkin lymphoma, and 23 of Ewing tumors family), obtained from microarrays containing 2308 genes [15].

Table 1. Description of the datasets used.

	Datasets	Attributes	Imbalance analysis		Classes
			Instances	IR	
1.	Leukemia	3051	38	2.454	2
2.	Colon	2000	62	1.818	2
3.	Adenocarcinoma	9868	76	5.333	2
4.	Brain tumor	5597	42	2.500	5
5.	Breast cancer 1	4869	77	1.330	2
6.	Breast cancer 2	4869	95	2.444	3
7.	Lymphoma	4026	62	4.666	3
8.	National Cancer Institute (NCI)	5244	61	1.800	8
9.	Prostate cancer	6033	102	1.040	2
10.	SRBCT	2308	63	2.875	4

In table 1, a summary of the description of the datasets is given. The summary includes the number of attributes, the number of instances, the Imbalance Ratio (IR) among majority and minority classes, and the number of classes. For validation purposes, we used the Leave One Out validation (LOO).

3.2 Performance Measures and Statistical Tests

Imbalanced datasets, or otherwise known as class imbalance problems, manifest their presence when one or more class is underrepresented in the dataset. Some classic performance measures produce a bias towards the majority class in this type of imbalanced datasets. Therefore, these measures become inadequate for classification. To evaluate the performance in this type of datasets, the use of the average True Positive Rate for each class have been recently suggested [19].

		Real Class	
		Positive	**Negative**
Predicted	**Positive**	True Positive	False Positive
	Negative	False Negative	True Negative

Fig. 1. Confusion matrix with data distributed in two classes.

The True Positive Rate (TPR) considers the total positive instances correctly classified in relation to the total instances of the positive class for problems of two classes (Fig. 1), in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) is expressed in equation (1). Also, the True Positive Rate is known as recall or sensitivity.

$$TPR = \frac{TP}{TP + FN}. \tag{1}$$

For a problem with k classes, the sensitivity of the classification in class j calculates the probability of correctly classifying an instance of class j, taking the total of correctly classified instances of class j in relation with the total instances of class j. For the calculation of the sensitivity of the classification is shown in the following equation:

$$S_j = \frac{n_j}{t_j}, \tag{2}$$

where S_j is the sensitivity (also recall or True Positive Rate) of the classification of class j, n_j is the number of instances correctly classified and t_j is the total number of instances in class j.

The minimum classification sensitivity is a measure that allows us to handle multiple classes, and is expressed as follows [20]:

$$Minimum = \{S_j\}. \tag{3}$$

Although minimum classification sensitivity only considers the lowest of all the classification sensitivities calculated. On the other hand, the average classification sensitivity per class [21] is a measure that gives the same weight to each of the classes, regardless of the number of instances that each class has. For this reason, this paper will use this measure of performance, and is denoted as follows:

$$Average\ sensitivity = \frac{1}{k}\sum_{i=1}^{k} S_j, \tag{4}$$

where k is the class total and S_j is the classification sensitivity for j-th class. This measure of performance allows us to consider all classes without there being a preferential bias to any class.

Below we provide an example with data distributed in three classes, and where the average classification sensitivity and the minimum classification sensitivity are calculated:

		Real Class		
		X	**Y**	**Z**
Predicted Class	**X**	3	3	9
	Y	8	8	2
	Z	6	5	7

Fig. 2. An example with data distributed in three classes.

$$S_X = \frac{3}{15} = 0.2,\ S_Y = \frac{8}{18} = 0.444,\ S_C = \frac{7}{18} = 0.388,$$

$$Average = \frac{0.2 + 0.444 + 0.388}{3} = 0.3444,$$

$$Minimum = min\{0.2, 0.444, 0.388\} = 0.2.$$

In order to recognize which of the classification algorithms obtained the best experimental result, statistical hypothesis tests are used, which perform an analysis to evaluate whether there is or not a significant difference in the performance obtained by the models in the proposed datasets. For this paper the Friedman test, which is a non-parametric test, will be used, since it is highly recommended for the type of the analyzed data [22, 23].

Deepening, the Friedman [24] test can be viewed as an extension of the Wilcoxon test to include data recorded in more than two time periods or groups of three or more matched subjects. The test examines the ranges of the data generated in each time period to determine if the variables share the same continuous distribution of their origin.

3.3 Classification Algorithms

Some of the most recognized classification algorithms in the state of the art are: the decision trees, Nearest Neighbor, Naïve Bayes, Random Forest, Support Vector Machines (SVMs) and Sequential Minimal Optimization, which is a case of SVMs. In the next section, these algorithms were used to evaluate the datasets proposed.

Starting with the descriptions, a decision tree (DT) [25] is a map of the possible outcomes of a series of related decisions. The approach based on decision trees is a method commonly used in machine learning. The objective of this approach is to create a model that predicts the value of a target variable based on various input variables. Learning based on this type of approach is the construction of a decision tree from training columns, each labeled with its corresponding class. Some of the most used DT is C4.5, which generates a decision tree from the data through partitions performed recursively. C4.5 is a decision tree which is an extension of the ID3 algorithm.

Random Forests [26] are a combination of decision tree predictors, so that each tree depends on the values of a random vector sampled independently and with the same distribution. They are efficient for large data sets and can handle a large number of features.

The Statistical-Probabilistic approach if often based on the Bayes theorem, from which different methods have been inspired, such as Naïve Bayes [25]. The Bayes theorem tries to obtain the posterior probability in relation to the belonging of an instance to a certain class. For this, a priori probability of a given class is evaluated against the other classes.

Nearest Neighbor (NN) [25] or k nearest neighbor algorithm, is a supervised classification method based on metrics, whose training phase consists of storing the vectors of the training set. While in the test phase the distance between the stored vectors and the test vector is calculated, finally the "k" closest instances are selected; the class that is repeated more times is selected.

Support Vector Machines address the principle of risk minimization structure, that is why we can provide a generalized performance independent of the distribution of patterns. The central idea of the Support Vector Machines is the adjustment of a discrimination function that optimally uses the information of the patterns that separate the classes [25].

Finally, SMO [27] or Sequential minimal optimization, was an implementation of SVMs by John Platt. This model solves the problem of quadratic programming that had the Support Vector Machines during the training process. Another advantage that SMO

has is that the amount of memory used in training is linear, so it allows the use of large amounts of data.

4 Experimental Results and Discussion

In this section we present the experimental results of the classification models described above in the microarrays data for the classification of different types of cancer. We include k nearest neighbor (kNN using k = 3), a decision tree algorithm (C4.5), a Random Forest with 10 base trees, a Support Vector Machine version (SVM with polynomial grade 3 base) and SMO algorithm with polynomial grade 1 base. We used the default parameter values offered in the WEKA software package.

For the division of data between the training set and the test for the classifiers, Leave-one-out cross-validation (LOO) was used.

The results obtained are shown in the table 2, highlighting with bold the best results.

Table 2. Average True Positive Rate obtained by the classification algorithms.

Datasets	C4.5	kNN	Naïve Bayes	Random Forest	SVM (polynomial)	SMO
Leukemina	0.936	0.981	0.863	0.818	**1.00**	**1.00**
Colon	0.755	0.741	0.811	0.859	**0.869**	0.834
Adenocarcinoma	0.479	0.526	0.541	0.500	**0.700**	0.684
Brain tumor	0.635	0.680	0.585	0.760	0.810	**0.855**
Breast cancer 1	0.454	0.526	0.579	0.594	0.587	**0.632**
Breast cancer 2	0.413	0.588	**0.653**	0.537	0.544	0.595
Lymphoma	0.727	0.992	0.858	0.925	0.962	**1.00**
NCI	0.329	0.625	0.455	0.534	0.492	**0.644**
Prostate cancer	0.862	0.862	0.635	0.853	**0.902**	0.882
SRBCT	0.781	0.869	0.947	0.955	**0.989**	**0.989**
Times Best	0	0	1	0	5	6

The Friedman test was carried out in order to find the best classification method in a more appropriate way.

Table 3. Average Rankings of the algorithms (Friedman).

No.	Algorithm	Ranking
1	SMO	1.6
2	SVM	2.2
3	kNN	3.75
4	Random Forest	3.8
5	Naïve Bayes	4.3
6	C4.5	5.35

Table 3 shows the average of ranges obtained by each classifier in the Friedman test, where the best model was SMO. The Friedman test gives a probability of $p = 0.000052$, lower than the 0.05 significance value. But to validate if this model has a significant statistical difference, p-values are obtained by applying the Post Hoc comparison, where the Holm's procedure rejects those hypotheses that an unadjusted value of $p \leq 0.05$.

Table 4. Post Hoc comparison Table (Friedman).

No.	Algorithm	p
1	SVM	0.473289
2	kNN	**0.010177**
3	Random Forest	**0.008551**
4	Naïve Bayes	**0.00125**
5	C4.5	**0.000007**

In Table 4, the p-values that are less than or equal to 0.05 are marked (bold letters), which are the values that the test rejects the hypothesis, in other words, that there is a significant difference of the best ranked method (SMO) with respect to kNN, Random Forest, Naïve Bayes and C4.5 models. In contrast, the SVM model has a p-value greater than 0.5, so the test does not reject the null hypothesis. Therefore, SMO has no significant difference with SVM.

5 Conclusions and Future Work

In this document, we aim at evaluating several Machine Learning algorithms for breast cancer pre-diagnosis, based on microarray data. The evaluations made allow us to determine the best performing algorithms for this task.

The results that were obtained when analyzing the selected algorithms with microarray dataset, the SMO model was the one that obtained the best results. But when performing the Friedman test and the Holm post hoc test, it was shown that since there is no significant difference with the SVM model, it is just as feasible to use SVM to obtain good results. Considering these results, the best classifiers for this type of datasets were the SMO and SVM models.

As a future work, we would propose an algorithm related to the models that had better results (SVM and SMO) or a new algorithm different from those already evaluated, and that competes with the evaluated classifiers, obtaining a better result and obtaining a significant difference with the others. This new algorithm must be proposed to manage dataset with a large amount of data, such as microarrays data.

Acknowledgements. The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, CIDETEC, and CIC), the CONACyT, and SNI for their economic support to develop this work

References

1. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10), 906–914 (2000)
2. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439), 531–537 (1999)

3. Quackenbush, J.: Computational analysis of microarray data. Nature Reviews Genetics 2, 418–427 (2001)
4. Bhaskar, H., Hoyle, D.C., Singh, S.: Machine learning in bioinformatics: A brief survey and recommendations for practitioners. Computers in biology and medicine 36(10), 1104–1125 (2006)
5. Díaz-Uriarte, R., De Andres, S.A.: Gene selection and classification of microarray data using random forest. BMC bioinformatics 7(1), 3 (2006)
6. Birnbaum, D.J., Finetti, P., Lopresti, A., Gilabert, M., Poizat, F., Raoul, J.L., Delpero, J.R., Moutardier, V., Birnbaum, D., Mamessier, E., Bertucci, F.: A 25-gene classifier predicts overall survival in resectable pancreatic cancer. BMC Medicine 15(1), 170 (2017)
7. Dwivedi, A.K.: Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications 29(12), 1545–1554 (2018)
8. Korucuoglu, M., Isci, S., Ozgur A., Otu H.H.: Bayesian Pathway Analysis of Cancer Microarray Data. PLOS ONE 9(7), e102803 (2014)
9. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., Geman, D.: Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21, 3896– 50 3904 (2005)
10. Cho, S.B., Won, H.H.: Cancer classification using ensemble of neural networks with multiple significant gene subsets. Applied Intelligence 26(3), 1573–7497 (2007)
11. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning 46(1-3), 389–422 (2002)
12. Hu, Z., Fan, C., Oh, D.S., Marron, J.S., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L., Nobel, A., Parker, J., Ewend, M.G., Sawyer, L.R., Wu, J., Liu, Y., Nanda, R., Tretiakova, M., Orrico, A.R., Dreher, D., Palazzo, J.P., Perreard, L., Nelson, E., Mone, M., Hansen, H., Mullins, M., Quackenbush, J.F., Ellis, M.J. Olopade, O.I., Bernard, P.S., Perou, C.M.: The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics 7(1), 96 (2006)
13. Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J.S.: Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition 39(12), 2383–2392 (2006)
14. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97(457), 77–87 (2002)
15. Dettling, M., Buhlmann, P.: Supervised clustering of genes. Genome Biology 3 (12), 0069.1–0069.15 (2002)
16. Ramaswamy, S., Ross, K.N., Lander, E.S. Golub, T.R.: A molecular signature of metastasis in primary solid tumors. Nature Genetics 33, 49–54 (2003)
17. Díaz-Uriarte, R., De Andres, S.A.: Gene selection and classification of microarray data using random forest. BMC bioinformatics 7(1), 3 (2006)
18. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., de Rijn, M.V., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O.: Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics 24 (3), 227–235 (2000)
19. Fernández, A., López, V., Galar, M., Del Jesus, M.J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-based systems 42, 97–110 (2013)
20. Fazekas, M.: Analysing Data of Childhood Acute Lymphoid Leukaemia by Seasonal Time Series Methods. JUCS (Journal for Universal Computer Science) 12(9), 1190– 1195 (2006)
21. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, 113–141 (2013)

22. Demšar, J.: Statistical comparisons of classifiers over multiple datasets. Journal of Machine Learning Research 7, 1–30 (2010)
23. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180(10), 2044–2064 (2010)
24. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Annals of Mathematical Statistics 11, 86–92 (1940)
25. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley, New York (1973)
26. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
27. Platt, J.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft Research. Technical Report MSR-TR-98-14 (1998)