# Ensemble of Dissimilarity based Classifiers for Cancer Samples Classification

Ángela Blanco, Manuel Martín-Merino[1], and Javier de las Rivas[2]

[1] Universidad Pontificia de Salamanca
C/Compañía 5, 37002, Salamanca, Spain.
ablancogo@upsa.es mmartinmac@upsa.es
[2] Cancer Research Center (CIC-IBMCC, CSIC/USAL)
Salamanca, Spain. jrivas@usal.es

**Abstract** DNA Microarray technology allow us to identify cancer samples considering the gene expression levels across a collection of related samples.
Several classifiers such as Support Vector Machines (SVM), $k$ Nearest Neighbors ($k$-NN) or Diagonal Linear Discriminant Analysis (DLDA) have been applied to this problem. However, they are usually based on Euclidean distances that fail to reflect accurately the sample proximities. Several classifiers have been extended to work with non-Euclidean dissimilarities although none outperforms the others because they misclassify a different set of patterns.
In this paper, we combine different kind of dissimilarity based classifiers to reduce the misclassification errors. The diversity among classifiers is induced considering a set of complementary dissimilarities for three different type of models. The experimental results suggest that the algorithm proposed helps to improve classifiers based on a single dissimilarity and a widely used combination strategy such as Bagging.

## 1 Introduction

DNA Microarray technology allow us to monitor the expression levels of thousands of genes simultaneously across a collection of related samples. This technology has been applied particularly to the prediction of three different type of cancer with encouraging results [12].

A large variety of machine learning techniques have been proposed to this aim such as Support Vector Machines (SVM) [10], $k$ Nearest Neighbors [9] or Diagonal Linear Discriminant Analysis (DLDA) [9]. However, the algorithms considered in the literature rely frequently on the use of the Euclidean distance that fails to reflect accurately the proximities among the sample profiles [8,4]. The classifiers mentioned above have been extended to work with non-Euclidean dissimilarities [18]. In spite of this, the resulting algorithms misclassify a different set of patterns and fail to reduce significantly the errors. This can be explained because each dissimilarity reflects different features of the data and they induce different type of errors.

Several authors have pointed out that combining non-optimal classifiers can help to reduce particularly the variance of the predictor [14,20]. In order to achieve this goal, different versions of the classifier are usually built by sampling the patterns or the features [6]. Nevertheless, in our application, this kind of resampling techniques reduce the size of the training set. This may increase the bias of individual classifiers and the error of the combination [20].

In this paper, we build the diversity of classifiers considering three different kinds of models such as SVM, $k$-NN and DLDA. The diversity is increased considering a set of complementary dissimilarities for each model. The classifiers induced will take advantage of the whole sample avoiding the bias introduced by resampling techniques such as Bagging. In order to incorporate non-Euclidean dissimilarities the base classifiers are modified in an appropriate way. Finally, the classifiers are aggregated using a voting strategy [14]. The method proposed has been applied to the prediction of different type of cancer using the gene expression levels with remarkable results.

This paper is organized as follows. Section 2 introduces the dissimilarities considered to build the diversity of classifiers. Section 3 comments how the classifiers can be extended to work from a dissimilarity matrix. In section 4 we present our combination strategy. Section 5 illustrates the performance of the algorithm in the challenging problem of gene expression data analysis. Finally, section 6 gets conclusions and outlines future research trends.

## 2 Dissimilarities for gene expression data analysis

An important step in the design of a classifier is the choice of a proper dissimilarity that reflects the proximities among the objects. However, the choice of a good dissimilarity is not an easy task. Each measure reflects different features of the data and the classifiers induced by the dissimilarities misclassify frequently a different set of patterns. Therefore, no dissimilarity outperforms the others.
In this section, we comment shortly the main differences among several dissimilarities proposed to evaluate the proximity between samples considering the gene expression levels. For a deeper description and definitions see [8,4,11].

The Euclidean distance evaluates if the gene expression levels differ significantly across different samples. An interesting alternative is the cosine dissimilarity. This measure will become small when the ratio between the gene expression levels is similar for the two samples considered. It differs significantly from the Euclidean distance when the data is not normalized by the $L_2$ norm.
The correlation measure evaluates if the expression levels of genes change similarly in both samples. Correlation based measures tend to group together samples whose expression levels are linearly related. The correlation differs significantly from the cosine if the means of the sample profiles are not zero. This measure is sensitive to outliers. The Spearman rank dissimilarity is less sensitive to outliers because it computes a correlation between the ranks of the gene expression

levels. An alternative measure that helps to overcome the problem of outliers is the Kendall-$\tau$.

Due to the large number of genes, the sample profiles are codified in high dimensional and noisy spaces. In this case, the dissimilarities mentioned above are affected by the 'curse of dimensionality' [1,16]. Hence, most of the dissimilarities become almost constant and the differences among them are lost. To avoid this problem, it is recommended to reduce the number of features before computing the dissimilarities.

## 3 Dissimilarity based classifiers

Classical Support Vector Machines (SVM) [21] and Diagonal Linear Discriminant Analysis (DLDA) [9] are not able to work directly from a dissimilarity matrix. In this section, the classical SVM algorithm is extended to work from a dissimilarity matrix by defining a kernel of dissimilarities. Next, we comment shortly how to adapt DLDA to work with dissimilarities.

The SVM algorithm looks for a linear hyperplane $f(\boldsymbol{x};\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{x} + b$ that maximizes the margin $\gamma = 2/\|\boldsymbol{w}\|^2$. $\gamma$ determines the generalization ability of the SVM.

The optimization problem can be solved efficiently in dual space and the discriminant function can be expressed exclusively in terms of Mercer kernels [21],

$$f(\boldsymbol{x}) = \sum_{\alpha_i>0} \alpha_i y_i K(x, x_i) + w_0 \tag{1}$$

Non-Euclidean dissimilarities can be incorporated into the SVM algorithm by defining a kernel of dissimilarities [18,19]. Next we detail the idea.
Let $d$ be a dissimilarity [7] and $R = \{p_1, \ldots, p_n\}$ a subset of representatives drawn from the training set. Define the mapping $D(z, R) : \mathcal{F} \to \mathbb{R}^n$ as:

$$D(z, R) = [d(z, p_1), d(z, p_2), \ldots, d(z, p_n)] \tag{2}$$

This mapping defines a dissimilarity space where feature $i$ is given by $d(., p_i)$. The set of representatives $R$ determines the dimensionality of the feature space. The choice of $R$ is equivalent to select a subset of features in the dissimilarity space. Due to the small number of training samples in our application, we have considered the whole sample as a representative set. It has been suggested in the literature that selecting a smaller subset of representatives does not help to improve the resulting classifier [18].

Once the patterns have been represented in the dissimilarity space, a kernel of dissimilarities can be defined as:

$$K_{ij} = \langle D(\boldsymbol{x}_i, R), D(\boldsymbol{x}_j, R)\rangle \tag{3}$$

where $\langle ., .\rangle$ denotes the scalar product in the feature space. Thus, for the linear SVM the kernel matrix is written as $K = DD^T$. This matrix is positive definite

and keeps the nice properties of the optimization problem in the original SVM algorithm.

The DLDA is a variant of the Linear Discriminant Analysis (LDA) that considers diagonal and constant covariance matrices along the classes [9]. However, in order to apply this technique, a vectorial representation of the data should be obtained. To this aim, we follow the approach of [18]. First, the dissimilarities are embedded into an Euclidean space such that the inter-pattern distances reflect approximately the original dissimilarity matrix. Next, the test points are incorporated via a linear algebra operation. Finally, the DLDA is applied considering the vectorial representation obtained.

## 4 Combination of dissimilarity based classifiers

In this section, we introduce our ensemble of classifiers to reduce the errors and comment briefly the related work.

Our method builds the diversity of classifiers considering three different kind of models such as SVM, $k$-NN and DLDA. To increase the diversity among classifiers, we have considered several dissimilarities introduced in section 2. Each dissimilarity reflects different features of the data and the resulting classifiers will produce different errors. Thus, the combination will improve the performance of classifiers based on a single dissimilarity [6,15]. Besides, the diversity of classifiers is generated considering the whole training sample. In this way, we do not reduce the size of the training set which may induce bias in the individual classifiers. Notice that the combination strategies are not able to reduce the bias of single classifiers [20].

Figure 1 shows in an intuitive way how the combination of classifiers reduces the misclassification errors. For instance bold patterns are assigned to the wrong class by one classifier but using a voting strategy the patterns will be assigned to the right class.
Hence, our combination algorithm proceeds as follows: First, the set of complementary dissimilarities introduced in section 2 are computed. As we mentioned earlier, each classifier incorporates the dissimilarities in a different way. For the SVM algorithm, the kernel of dissimilarities is computed and the optimization problem is solved in the usual way. $k$-NN is able to work directly from a dissimilarity matrix but to avoid the 'curse of dimensionality' and to increase the diversity among them it is recommended to reduce previously the number of features. For the DLDA algorithm, the dissimilarities should be embedded in an Euclidean space via a Multidimensional Scaling algorithm. The ensemble of classifiers is aggregated by a standard voting strategy [14]. The diagram 1 shows the steps of the algorithm.

A related technique to combine classifiers is the Bagging [6,3]. This method generates a diversity of classifiers considering several bootstrap samples as training

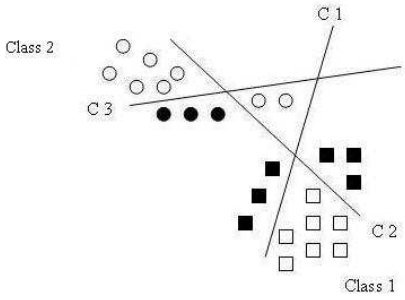


**Figure 1.** Aggregation of classifiers using a voting strategy. Bold patterns are misclassified by a single hyperplane but not by the combination.

**Algorithm 1** Aggregation of classifiers based on multiple models and dissimilarities.

1: For each measure compute the dissimilarity matrix
2: Compute the kernel of dissimilarities using equation (3) for the SVM algorithm
3: Embed each dissimilarity into an Euclidean space via MDS for DLDA algorithm
4: Train the classifiers for each dissimilarity
5: Combine the different models using a voting strategy
6: Evaluate the ensemble by ten-fold cross-validation.
7: End

sets. Next, the classifiers are aggregated using a voting strategy. Nevertheless there are three important differences between Bagging and the method proposed in this section.

First, our method generates the diversity of classifiers by considering the whole sample. Bagging trains each classifier using around 63% of the training set. In our application the size of the training set is very small and neglecting part of the patterns may increase the bias of each classifier. It has been suggested in the literature that Bagging does not help to reduce the bias [20] and so, the aggregation of classifiers will hardly reduce the misclassification error.

A second advantage of our method is that it is able to work directly from a dissimilarity matrix.

Finally, the combination of several dissimilarities avoids the problem of choosing a particular dissimilarity for the application we are dealing with. This is a difficult and time consuming task.

## 5 Experimental results

In this section, the ensemble of classifiers proposed is applied to the identification of cancerous samples using Microarray gene expression data.

Three benchmark gene expression datasets have been considered. The first one consisted of 72 bone marrow samples (47 ALL and 25 AML) obtained from acute leukemia patients at the time of diagnosis [12]. The RNA from marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays produced by Affymetrix and containing 6817 genes. The second dataset consisted of 49 samples from breast tumors [22], 25 classified as positive to estrogen receptors (ER+) and 24 negative to estrogen receptors (ER-). Those positive to estrogen receptors require a different treatment. The RNA of breast cancer cells were hybridized to high-density oligonucleotide microarrays produced by Affymetrix and containing 7129 genes. Finally the third dataset consists of 40 tumor and 22 normal colon samples, analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. The number of genes was reduced in the original dataset to 2000 [2].

Due to the large number of genes, samples are codified in a high dimensional and noisy space. Therefore, the dissimilarities are affected by the 'curse of dimensionality' and the correlation among them becomes large [16]. To avoid this problem and to increase the diversity among dissimilarities we have reduced the number of genes using the standard F-statistic [11]. The number of genes considered for SVM and DLDA are 14% while for $k$-NN the number of genes kept is 3% because this technique is more sensitive to noise.

The dissimilarities have been computed without normalizing the variables because as we have mentioned in section 2 this operation may increase the correlation among them.

The algorithm chosen to train the Support Vector Machines is C-SVM. The $C$ regularization parameter has been set up by ten fold-crossvalidation [17,5]. We have considered linear kernels in all the experiments because the small size of the training set in our application favors the overfitting of the data. Consequently error rates are smaller for linear kernels than for non linear ones.

The number of neighbors for $k$-NN algorithm is estimated by cross-validation. Before applying DLDA the dissimilarities should be embedded in an Euclidean space using a Multidimensional Scaling algorithm.

**Table 1.** Empirical results for the best single classifier for each technique.

| Technique | Datasets | Error % | False negative % |
|---|---|---|---|
| SVM (Correlation) | Golub | 6.94% | 2.77% |
| SVM (Tau) | Breast | 6.12% | 2.04% |
| SVM(Correlation) | Colon | 14.5% | 6.45% |
| $K$-NN (Tau) | Golub | 1.38% | 1.38% |
| $K$-NN(Tau) | Breast | 8.16% | 2.04% |
| $K$-NN (Cosine) | Colon | 12.9% | 4.83% |
| DLDA (Tau) | Golub | 2.77% | 1.38% |
| DLDA(Spearman) | Breast | 8.16% | 2.04% |
| DLDA (Euclidean) | Colon | 11.29% | 4.83% |

**Table 2.** Empirical results for the combination of classifiers. The Bagging technique has been taken as reference.

| Technique | Datasets | Error % | False negative % |
|---|---|---|---|
| Bagging (SVM) | Golub | 2.77% | 1.38% |
| | Breast | 6.12% | 2.04% |
| | Colon | 12.9% | 4.83% |
| Bagging ($k$-NN) | Golub | 5.55% | 5.55% |
| | Breast | 14.28% | 6.12% |
| | Colon | 14.51% | 9.67% |
| Bagging (DLDA) | Golub | 6.94% | 4.16% |
| | Breast | 14.28% | 2.04% |
| | Colon | 11.29% | 3.22% |
| **Combination** | Golub | 1.38% | 1.38% |
| | Breast | 4.08% | 2.04% |
| | Colon | 11.2% | 3.22% |

The algorithms have been evaluated considering the global errors and the false negative errors. Both have been estimated by ten-fold cross-validation which gives good experimental results for the problem at hand [17].

Table 1 shows the experimental results for the best single classifier for each technique. Table 2 compares the method proposed with Bagging, introduced in section 3. Both, Bagging and the best classifiers based on a single dissimilarity for each model have been taken as a reference.

From the analysis of tables 1 and 2, the following conclusions can be drawn:

- The dissimilarity that minimizes the error depends strongly on the classifier and on the particular dataset considered. No dissimilarity outperforms the others for a wide range of models and datasets. Hence the choice of a proper dissimilarity is not an easy task for human experts.
- The combination strategy proposed improves the misclassification errors of the best single classifiers. In particular, the ensemble of classifiers improves significantly the SVM algorithms for the three problems considered. False negative errors are particularly reduced in Golub and Colon datasets. We also report that our method improves the best $k$-NN classifier for Breast and Colon which are the most complex datasets according to the literature. Finally, DLDA is also improved.
- The ensemble of classifiers proposed compares favorably with a widely used combination algorithm such as Bagging. Our algorithm improves often the errors of the best bagging technique. Besides, the combination strategy proposed performs always as well as the best bagging technique based on a single dissimilarity. Finally, note that Bagging fails often to reduce the misclassification errors of classifiers based on a single dissimilarity.

## 6 Conclusions and future research trends

In this paper, we have proposed an ensemble of classifiers based on a diversity of models and dissimilarities. Our approach aims to reduce the misclassification error of classifiers based solely on a single measure. The algorithm has been applied to the classification of cancerous samples using gene expression data.

The experimental results suggest that the method proposed improves the misclassification error of classifiers based on a single dissimilarity. We also report that our method compares favorably with a widely used combination algorithm such as Bagging.

## References


1. C. C. Aggarwal, "Re-designing distance functions and distance-based applications for high dimensional applications," in *Proc. of the ACM International Conference on Management of Data and Symposium on Principles of Database Systems (SIGMOD-PODS)*, vol. 1, March 2001, pp. 13–18.
2. U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat'l Acad Sci USA, 96:6745–6750, 1999.
3. E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, vol. 36, pp. 105–139, 1999.
4. Blanco, A., Martín-Merino, M. and De Las Rivas, J. :Combining dissimilarity based classifiers for cancer prediction using gene expression profiles, *BMC Bioinformatics*, BioMed Central Ltd, London, UK, 2007.
5. U. Braga-Neto and E. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.
6. L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.
7. T. Cox and M. Cox, *Multidimensional Scaling*, 2nd ed. New York: Chapman & Hall/CRC Press, 2001.
8. S. Drãghici, *Data Analysis Tools for DNA Microarrays*. New York: Chapman & Hall/CRC Press, 2003.
9. S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, pp. 77–87, 2002.
10. T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.
11. R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Berlin: Springer Verlag, 2006.
12. T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 15, pp. 531–537, 1999.

13. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.
14. J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 228–239, March 1998.
15. L. I. Kuncheva. "Combining Pattern Classifiers". John Wiley, New Jersey, 2004.
16. M. Martín-Merino and A. M. noz, "A new sammon algorithm for sparse data visualization," in *International Conference on Pattern Recognition (ICPR)*, vol. 1. Cambridge (UK): IEEE Press, August 2004, pp. 477–481.
17. A. Molinaro, R. Simon, and R. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005.
18. E. Pekalska, P. Paclick, and R. Duin, "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2001.
19. B. Schölkopf and A. Smola, *Learning with Kernels*. MIT Press, USA, 2002.
20. G. Valentini and T. Dietterich, "Bias-variance analysis of support vector machines for the development of svm-based ensemble methods," *Journal of Machine Learning Research*, vol. 5, pp. 725–775, 2004.
21. V. Vapnik, *Statistical Learning Theory*. New York: John Wiley & Sons, 1998.
22. M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, and J. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," *PNAS*, vol. 98, no. 20, September 2001.