ICA as a preprocessing technique for classification*

V.Sanchez-Poblador[1], E. Monte-Moreno[1], J. Solé-Casals[2]

[1]TALP Research Center
Universitat Politècnica de Catalunya (Catalonia, Spain)
enric@gps.tsc.upc.es
http://gps-tsc.upc.es/veu/personal/enric/enric.html

[2]Signal Processing Group, University of Vic (Catalonia, Spain)
jordi.sole@uvic.es
http://www.uvic.es/eps/recerca/ca/processament/inici.html

Abstract. In this paper we propose the use of the independent component analysis (ICA) [1] technique for improving the classification rate of decision trees and multilayer perceptrons [2], [3]. The use of an ICA for the preprocessing stage, makes the structure of both classifiers simpler, and therefore improves the generalization properties. The hypothesis behind the proposed preprocessing is that an ICA analysis will transform the feature space into a space where the components are independent, and aligned to the axes and therefore will be more adapted to the way that a decision tree is constructed. Also the inference of the weights of a multilayer perceptron will be much easier because the gradient search in the weight space will follow independent trajectories. The result is that classifiers are less complex and on some databases the error rate is lower. This idea is also applicable to regression

1. Introduction

The problem of classification consists on deciding a class membership of an observation vector [2]. Usually this observation vector consists of features that are related. The classification algorithm has to take a decision after the analysis of several features even though they can be mutually related in difficult ways. The dependencies between the features have an influence on the learned classifier. In the case of a decision tree, each node analyses a single feature, or a lineal combination of features, and the selection of a feature for a given level of the tree, is made in a greedy way [3]. As the decisions in the tree are made on a feature or subset of features basis, it would help if one could transform the features in such a way that instead of having the discriminative information spread through all the features, on could make each feature independent of the others and consequently simplify the decision process. This simplification of the decision process gives trees with a lower number of nodes. It is well known that there is a relationship between the complexity of a classifier and the generalization error [4]. Generally this complexity is dealt by pruning

* This work has been partially supported by the Spanish CICyT project ALIADO, by the EU integrated project CHIL and by the University of Vic under the grant R0912

the tree. We propose transforming the input so that the resulting vector has the property that each component is independent of the others. We shall do this by means of ICA In the case of classifying by means of a decision tree, as there is no statistical dependency between features the number of decisions (i.e. nodes) is lower, as it is confirmed by our experimental results. In the case of training a multilayer perceptron, the inference of the weights is made by a gradient search, which is known to be very inefficient if the features are highly correlated [2]. It is also known that the incorrelation preprocessing of the inputs of a multilayer perceptron improves the convergence of the algorithm because near a minimum the form of the error function can be approximated locally by a hyper-parabola. This explains the improvement that can be achieved by the use of algorithms such as the conjugate gradient or the Levenberg-Marquardt. Notice that the characteristics of these algorithms are adapted to the fact that the data can have correlated features. So a process of whitening the data or using these improvements of the gradient algorithms means that we are making a strong hypothesis about the data. We propose to preprocess the data in such a way that the features will be mutually independent, and therefore the gradient descent will follow a smooth surface, even if high order moments between features are present in the original pattern. In the past, ICA has been used for training decision trees [6], where the authors searched for an unsupervised method for training decision trees. This method uses the fact that by means of ICA feature space is transformed in such a way that data is aligned to the axis, therefore, they suppose that the components that go to a given node correspond somehow to a given class. As we will see in section 2, this is true depending on the distribution of the data. In our case, we use explicitly the fact that data is labeled, and the classifiers are constructed in a supervised way

2. Independent Component Analysis as preprocessing tool

Independent component analysis supposes that the observation vector was generated by a linear mixture of a set of independent random variables. This hypothesis might not be true for all the classification problems, as the generation of the observations do not always come from a linear mixture. For instance, in the classification task for the echocardiogram database, we have a set of variables, which although are related, are not generated by a linear mixture. For instance in the case of the echocardiogram database, the Left ventricular end-diastolic dimension (LVDD), and the age of the patient affected by the infarction/heart attack, are related, i.e. sick hearts are related to high values of the LVDD, and with older patients. But the age cannot be interpreted as a linear mixture of two different causes, which also generate the LVDD. As can be seen in the scatter plot of figure 1, the two classes do not follow different directions in the feature space, therefore an ICA preprocessing will not find rotations or transforms that separate classes. This effect will be seen in the results, where the preprocessing with ICA does not improve results. Nevertheless, in the case of the crabs database (see figure 2), the classes are spread along different directions, therefore, the effect of the ICA preprocessing will be the alignment of the classes with the axis, which will increase the mean distance between the instances of different classes, and at the same time sets the orientation of the class boundaries in parallel with the axes. And also, as shown in the figure, decision boundaries are simpler. In this case the whitening preprocessing yields the most simple border. In the case of vowel2 database, as can be seen in figure 3, either whitening or transforming with an ICA processing does not change significantly the relative position between the points in the feature space. In this case, the complexity of the decision boundary remains more or less the same, and therefore the recognition rate is practically the same in the three cases. In this case, there is no improvement because of the geometry of the problem. The classes are distributed in such a way that the orientation of the samples of each class is different.

Another example where the recognition rate does not change due to the preprocessing is the case of breast cancer. Figure 4, shows the scatter plot for two combinations of features, with the decision boundaries made either by a multilayer perceptron or by a decision tree. As can be seen, the distribution of samples is similar (except for a rotation), in all cases, and therefore the decision boundaries have more or less the same complexity. The error rate is more or less similar for all the cases, perhaps slightly lower for the case of a preprocessing with ICA. But in the case of the echocardiogram database, the results are different. In figure 5, we show the scatter plot for different components. It can be seen that the distribution without preprocessing or with a PCA, give distribution of points where classes are not aligned with the axes, which does not happen after the ICA processing. The result is that the class borders are smoother in the case of the multilayer perceptron, and in the case of decision trees, the borders are aligned with features. Therefore, we can expect that the classification results will improve after ICA in the cases were the feature space has a structure where the data is aligned with certain directions.

3. Experimental frameworks

In order to test the effect of an ICA preprocessing and compare with a whitening preprocessing, we did an experiment on 9 databases [5], which are summarized in table 1. Both preprocessing were done without a dimensionality reduction. The databases were divided into train/test. In the case of the multilayer perceptron this partition was done in order to use the validation sub-database for stopping the training phase. In order to compute the test results, we did 50 bootstrap samples of the data base. For each bootstrap replication, the bootstrap data was used for training (in the case of the multilayer perceptron for training and validation), and for test the samples that were not selected, therefore the relation train/test was about 65%/35 % depending on the realization. As it is well known that the CART algorithm is unstable [7], and neural nets fall into local minima, the results that we present in the tables are the mean of the 10 best results,. i.e. the mean of the best error rates and complexity of the classifiers. The pruning criterion of the decision trees and the number of hidden layer units were selected by cross validation. The ICA transformation was done by means of the Jade algorithm [8]. We used two different classifiers; decision trees and multilayer perceptrons. For the decision trees we used the CART algorithm [3], and the multilayer perceptron was trained by means of the Levenberg-Marquart algorithm. The results presented in tables (2,3,4) are somehow lower than the ones found in literature, because the objective of the paper was to assert the effect of the preprocessing, which meant dealing with the instability of the classifiers (i.e. local minima of the multilayer perceptron, different tree structure with slightly different database). In order to smooth the variability of the classifiers, the train and test databases were selected by bootstrap, which meant that the training database was poorer than the standard experiments, which assign about 90% the data, instead of a 65%.

4. Results

In Tables 2, 3 and 4 we present the classification errors and complexity of two different classifiers; decision trees and multilayer perceptrons. We compare a baseline version without data preprocessing (table 2), with the results obtained by the whitening preprocessing (table 3), and ICA preprocessing (table 4). The first conclusion that can be drawn from the results is that some databases benefit from the ICA processing, while in others the benefit is so small that is within the confidence margins, so we cannot assert that there is a real benefit or even in some cases there is a degradation. Improvements are consistent with the classification algorithms, that is, the preprocessing improves the results in both the trees and the multilayer perceptron, although the performance between classifiers can be different. The explanation of this different behavior is related to the distribution of the data on the feature space and is explained in section 2. In cases where the distribution of the classes in the feature space has regularity such that classes can be aligned in certain directions there is a clear improvement. A sign of it is that the databases that improve the error rate with ICA, also improve when we whitened the data. The improvement gotten by the whitening is in most cases lower than the one obtained by the use of ICA. The whitening process improved in 6 of the 9 databases in the case of decision trees, and 4 of the 9 in the case of the multilayer perceptron. The ICA preprocessing yielded improvements in 5 out of 9 cases for the case of the decision trees and 6 cases out of 9 for the multilayer perceptron. On the other hand, the complexity of the decision trees changes depending on the preprocessing, and is related to the improvement/degradation of the recognition rate. In the case of decision trees an improvement on the recognition rate is associated with a simpler structure (3 best improvements), while in the cases when the recognition rate does not change or degrades the number of nodes of the decision tree increases. The best results with a multilayer perceptron were not associated to smaller networks, the benefit of the ICA was due to the smoother error surface that gave a lower number of local minima.

5. Conclusions

We have shown that the use of ICA as a preprocessing tool can improve the classification results when the feature space has a certain structure. This improvement does not happen always and is related to the fact that the classes distribution in the feature spaces is such, that after the ICA preprocessing the samples of classes are better aligned with the dimensions. In these cases, ICA gives better results than PCA.

References

[1] A. Hyvärinen, J. Karhunen, E. Oja: Independent Component Analysis, Wiley, 2001

[2] Richard O. Duda, Peter E. Hart, David G. Stork: Pattern Classification (2nd Edition) Wiley Interscience, 2000

[3] Breiman, Friedman, Olshen, Stone: Classification And Regression Trees (CART) Chapman & Hall, 1984

[4] Machine Learning Tom M. Mitchell McGraw-Hill, 1997

[5] Blake, C.L. & Merz, C.J.: UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.(1998).

[6] Petteri Pajunen, Mark Girolami: Implementing decisions in binary Decision Trees using ICA, 2nd International Workshop on ICA and BSS, 483–487, Helsinki, 2000.

[7] L. Breiman: Bagging predictors. Machine Learning, 24(2): 123-140, 1996.

[8] Jean-François Cardoso and Antoine Souloumiac: Jacobi angles for simultaneous diagonalization, (SIAM) J. Mat. Anal. Appl. jan, 1996.

Table 1. Data Set Summary

Data Set	Size	Inputs	Classes
biomed	209	4	2
breast cancer	683	9	2
crabs	200	6	2
echo	62	8	2
sonar	208	60	2
titanic	2201	3	2
vowel	990	10	11
wine	178	13	3
vowel2	1520	4	10

Table 2. Test set Errors (%) without processing

Data Set	Decision Tree	Multilayer perceptron	Complexity	
			Tree/ Mean Number of nodes	MLP/ Nodes hidden layer
biomed	12.77	09.30	24	6
breast cancer	3.49	2.55	42	6
crabs	9.04	03.79	28	6
echocardio	21.96	21.41	4	8
sonar	25.53	15.97	29	10
titanic	29.13	19.97	1	6
vowel	29.80	24.13	558	20
wine	6.42	0.74	34	10
vowel2	21.17	13.44	522	20

Table 3. Test set Errors (%) whitening pre processing

Data Set	Decision Tree	Multilayer perceptron	Complexity	
			Tree/ Mean Number of nodes	MLP/ Nodes hidden layer
biomed	13.15	08.36	48	10
breast cancer	3.36	02.20	27	6
crabs	2.28	02.34	12	6
echocardiogram	20.74	22.50	2	2
sonar	19.08	17.59	39	6
titanic	28.03	20.11	4	8
vowel	37.56	21.05	648	20
wine	04.92	01.64	22	8
vowel2	25.95	14.24	577	20

Table 4. Test set Errors (%) ICA preprocessing

Data Set	Decision Tree	Multilayer perceptron	Complexity	
			Tree/ Mean Number of nodes	MLP/ Nodes hidden layer
biomed	14.58	09.17	41	10
breast cancer	04.79	02.63	50	8
crabs	06.85	02.47	25	10
echocardiogram	20.05	19.40	2	10
sonar	24.57	15.46	44	10
titanic	27.73	20.14	3	8
vowel	48.70	20.77	799	20
wine	20.86	01.47	53	10
vowel2	20.15	13.33	349	20

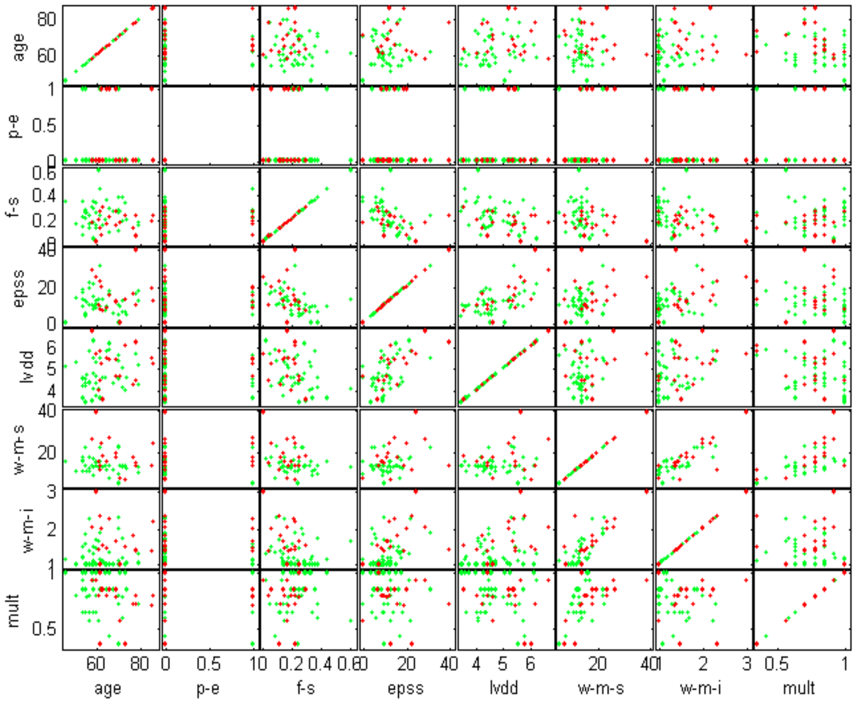

Fig. 1. Scatter plot of the echocardiogram database, before preprocessing

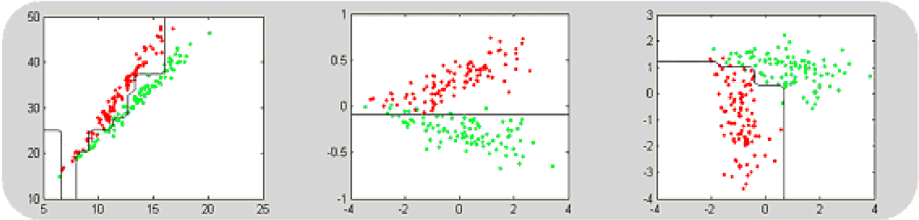

Fig. 2. Scatter plot for two features of the crabs database (length, width of carapace), the scatter plot after whitening and after ICA. The decision boundaries correspond to a decision tree.

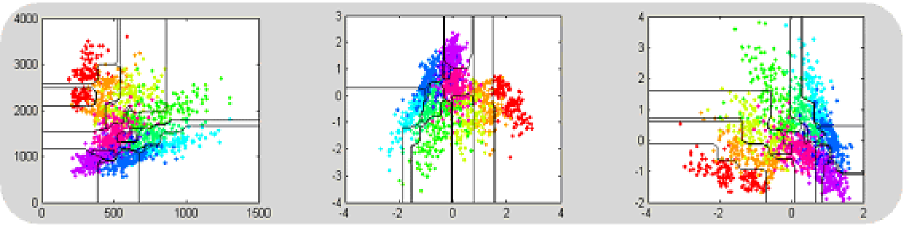

Fig. 3. Scatter plot for two features (first two formants) of the vowel2 database, after whitening and after ICA. The decision boundaries correspond to the decision tree.

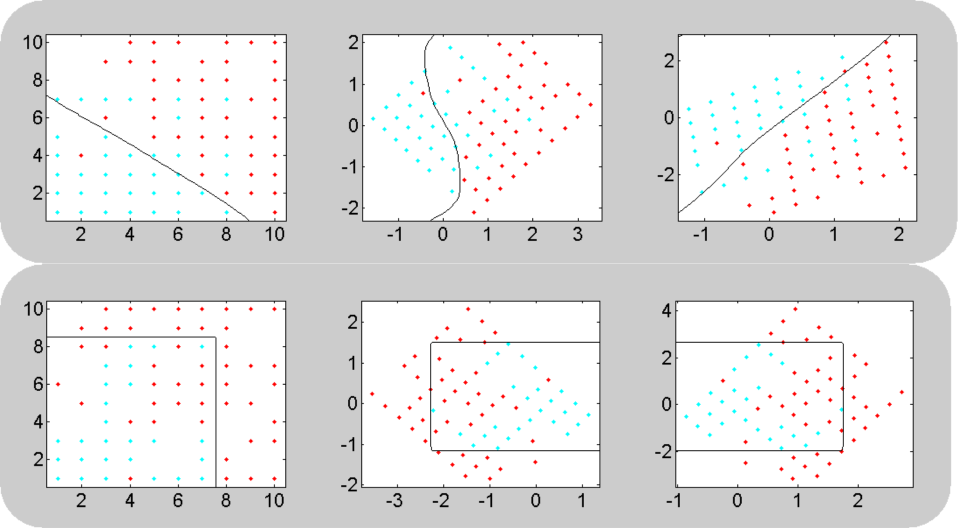

Fig. 4. Scatter plot for two features of the breast cancer database. Upper figures, features (1, 7), lower figure features (7, 8).Scatter plot without processing, after whitening and after ICA. The depicted decision boundaries correspond in the upper figures to a multilayer perceptron and in the lower to a decision tree.

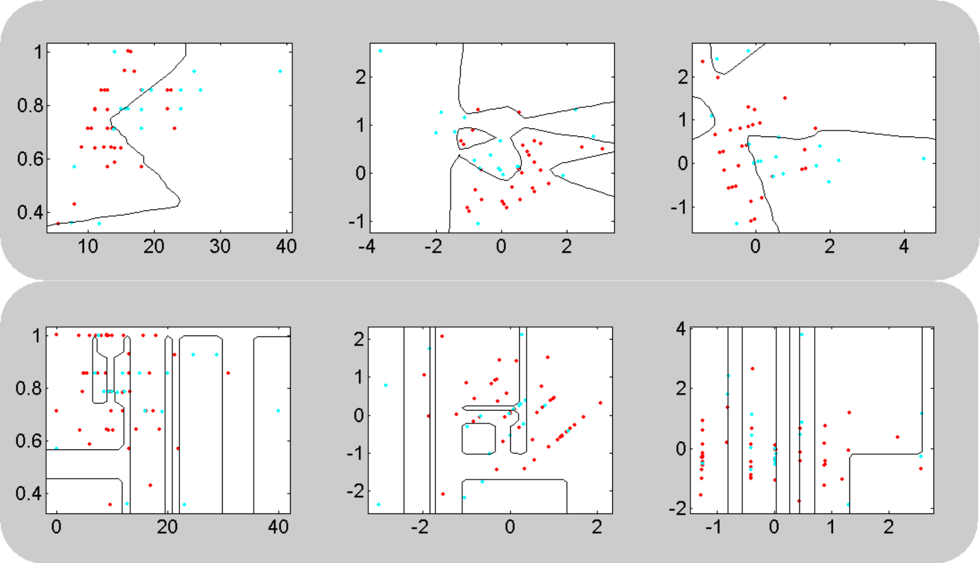

Fig. 5. Scatter plot for two features of the echocardiogram database. Upper figures, features (6, 8), lower figure features (4, 8). Scatter plot without processing, after whitening and after ICA. The depicted decision boundaries correspond in the upper figures to a multilayer perceptron and in the lower to a decision tree.