Combining classifiers based on Gaussian Mixtures

Luis Daza
Advisor: Dr. Edgar Acuña

Departament of Mathematics, University of Puerto Rico at Mayagüez, Mayagüez, PR 00680,

Abstract

A combination of classification rules (classifiers) is known as an *Ensemble*, and in general it is more accurate than the individual classifiers used to build it. Two popular methods to construct an *Ensemble* are Bagging (Bootstrap aggregating) introduced by Breiman, (1996) and Boosting (Freund and Schapire, 1996). Both methods rely on resampling techniques to obtain different training sets for each of the classifiers. Previous work has shown that Bagging as well as Boosting are very effective for unstable classifiers. In this paper we present some results in application of Bagging and Boosting to classifiers where the class conditional density is estimated using Gaussian mixtures. The effect of feature selection on the bundling of classifiers is also considered.

1. Introduction

Many researches have investigated the technique of combining the predictions of multiple classifiers to produce a single classifier: Breiman (1996, 1998), Quinlan (1996), Freund and Schapire (1996), Maclin and Optiz (1997), Bauer and Kohavi (1999), Acuña, et al. (2002) among others. Breiman (1996) heuristically defines a classifier as unstable if an small change in the learning data L can make large changes in the classification. Unstable classifiers have low bias but high variance, meanwhile the opposite occurs for stable classifiers. CART and neural networks are not stable classifiers, linear discriminant analysis and K-nearest neighbor classifiers are stable. It is expected a reduction of the bias and variance after the classifiers are combined.

Bagging and Boosting are very effective for unstable classifiers such as decision trees: CART, C4.5 and MC4 (see Breiman (1996, 1998), Quinlan (1996), Freund and Schapire (1996), Bauer and Kohavi (1999), Dietterich (2000) and neural networks (see Maclin and Optiz (1997)). Boosting applied to decision trees and Naïve-Bayes performs generally better that Bagging, but not uniformly better, sometimes they degraded compared to the single classifier. The same conclusions were obtained for neural networks classifiers. Bagging mainly reduces the variance, whereas boosting reduces, both the bias and the variance.

2 Classifiers based on Gaussian mixtures

From a Bayesian point of view, supervised classification is equivalent to compare estimates of the probabilities of belonging to each with each other, assigning an object with measurement vector **x** to the class with the largest $\hat{f}(j/\mathbf{x}),\ j=1,2,\cdots,g$. In order to obtain such estimates, one can estimate them indirectly via the class conditional density $f(\mathbf{x}/j)$ using the Bayes' theorem. Gaussian mixtures can be used to carry out that task. For a given class j with n_j instances and a random sample $\mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_n$ of the p-dimensional random vector **x**, the Gaussian mixture estimate of the class conditional density at the point **x** is given by

$$\hat{f}(x/j) = \sum_{k=1}^{K_j} p_{jk}\, f(x, \boldsymbol{\mu}_{jk}, \mathbf{S})$$

$$j = 1,2,\cdots,g$$

where f represents the multivariate normal density con mean vector m_{jk} and covarianza matrix S, and K_j is the number of subclasses of the j-th class. In each class, It is assumed that all the subclasses have the same covariance matrix. Also, $\pi_{jk}>0$ and $\sum_{k=1}^{K_j} p_{jk} = 1$. The parameters m_{jk}, S and π_{jk} are estimated via the EM algorithm (Hastie and Tibshirani, 1994) using a random sample.

3 Experimental Methodology

We chose 11 datasets coming from the Machine Learning Database at University of California Irvine (UCI) to evaluate the effect of combining GM classifiers. A summary of the datasets appears in Table 1.

The setup for Bagging was as follows: Each dataset is randomly divided in 10 parts, the first one is taken as

the test sample and the remaining is considered as the learning sample. Next, 50 bootstrapped samples are taking from the learning sample and a Gaussian mixture classifier is constructed with each of them. Finally, each instance of the test sample is assigned to a class by voting using the 50 classifiers previously constructed. The proportion of instances incorrectly assigned will be the bagged misclassification error. We repeat the steps considering now the second part as the test set and in this way we continue until the tenth part is considered as the test set. The whole procedure is repeated 10 times and we compute the average of the bagged misclassification error.

Dataset	Instances	Features	Classes
Breastw	699	9	2
Bupa	345	6	2
Creditg	1000	20	2
Crx	690	15	2
Diabetes	768	8	2
Heartc	294	13	2
Ionosphere	351	34	2
Iris	150	4	3
Segment	2310	19	7
Sonar	208	60	2
Vehicle	846	18	4

Table 1. Datasets used in this paper

The misclassification error of a single classifier is estimated by a 10-fold crossvalidation and averaged over 50 runs. We also computed the ratio of the misclassification errors of the bagged classifier versus the single one. The results are shown in table 2

The setup for Boosting is quite similar to the one used in Bagging, the only difference is that a bootstrap sample depends on the misclassification errors on the previous one. In the first step a bootstrap sample is drawn from the original one assigning equal weight to every instance. Then a classifier is built using the bootstrap sample and its misclassification error is computed. For the second bootstrap sample a instance has more weight if was misclassified in the first step. The procedure continues until 20 bootstrap samples are drawn. Finally a weighted voting is applied to assign a object to a class.

The boosted misclassification error is averaged on 10 repetitions. We also computed the ratio of the misclassification errors of the boosted classifier versus the single one. The results appear in Table 3.

Dataset	Sub-clases	Single	Bagged	Ratio
Iris	2	2.33	2.00	0.858
sonar	3	24.24	18.90	0.780
heart-c	5	18.46	16.57	0.898
Bupa	5	32.20	30.65	0.952
Ionosfera	3	15.32	15.28	0.997
Crx	3	13.69	13.17	0.962
Breast-w	3	4.30	3.70	0.860
Diabetes	5	25.50	24.09	0.945
Vehicle	4	20.18	17.53	0.869
German	5	24.33	23.7	0.974
Segment	6	7.19	5.71	0.794
MEAN				0.899

Table 2. Comparison of misclassification error rates for single and bagged GM classifiers.

dataset	Sub-classes	single	Boosted	ratio
Iris	2	2.33	2.00	0.858
Sonar	3	24.24	20.86	0.861
heart-c	4	17.83	19.53	1.095
Bupa	2	33.07	32.93	0.996
ionosfera	3	15.32	16.07	1.049
Crx	2	13.48	13.94	1.034
Breast-w	5	4.57	4.33	0.947
Diabetes	5	25.50	24.97	0.979
Vehicle	4	20.18	20.71	1.026
German	2	24.62	26.07	1.059
Segment	6	7.19	7.25	1.008
MEAN				0.992

Table 3. Comparison of misclassification error rates for single and boosted GM classifiers.

The average of the error reduction for the 11 datasets after Boosting using the Gaussian mixture classifier was almost none, only 0.80%. When C4.5 classifier

was boosted (Quinlan, 1996) the average error reduction for the same datasets was 8.83%. Notice that the boosted classifier performed well only in Iris and Sonar.

4 Effect of Feature selection on combining Gaussian classifiers

To speed up the computation of the ensembles we perform feature selection. A forward selection procedure was used and repeated 10 times for datasets with less than 20 features and 20 times for datasets with more than 20 features.
First we select the single feature that produces the highest classification rate estimated by 10 fold cross-validation using the classifier based on kernel density estimator. Once that this is done we search for the second feature that, together with the first one yields the highest classification rate. The procedure continues until the classification rate decreases. After that we compute the average number of selected features for each dataset, rounding it if it is necessary. Finally, for each dataset, we select the features appearing more frequently in the 10 replications. Once that the predictors are selected we create one subsets of each of the datasets and then we perform bagging and boosting using GM classifiers.

Figure 1 shows a comparison of the misclassification error rates for the single and bagged classifier after feature selection. Figure 2 shows the same comparisons for the boosted classifier.

Datasets	Sub classes	Before	After	Ratio
Breast-w	3	4.30	4.19	0.974
Bupa	5	32.20	29.91	0.929
Credit-g	5	24.33	24.19	0.994
Crx	3	13.69	13.55	0.990
Diabetes	5	25.50	22.74	0.892
Heart-c	5	18.46	19.78	1.072
Ionosphere	3	15.32	12.74	0.832
Segment	6	7.19	5.48	0.762
Sonar	3	24.24	22.18	0.915
Vehicle	4	20.18	21.54	1.067
Mean				0.943

Table 4. Effect of forward feature selection on the performance of GM classifiers

Datasets	Sub classes	Single	*Bagged*	Ratio
Breast-w	3	4.19	3.56	0.851
Bupa	5	29.91	28.46	0.952
Credit-g	5	24.19	24.03	0.994
Crx	3	13.55	13.38	0.988
Diabetes	5	22.74	21.64	0.952
Heart-c	5	19.78	18.11	0.916
Ionosphere	3	12.74	11.91	0.934
Segment	6	5.48	5.04	0.919
Sonar	3	22.18	21.92	0.988
Vehicle	4	21.54	20.18	0.937
Mean				**0.943**

Table 5. Effect of forward feature selection on the Misclassification errors by Bagging.

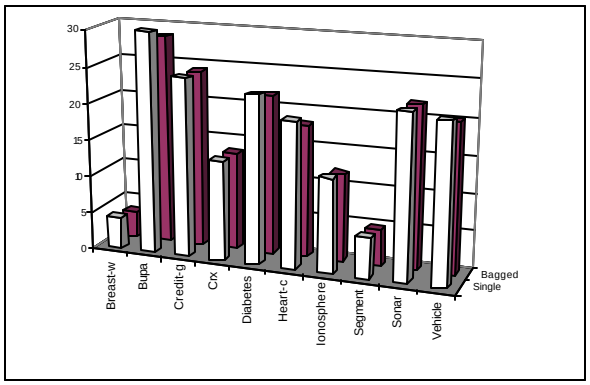

Figure 1 Comparison of the misclassification error rates for the single and bagged classifier after forward feature selection.

Dataset	Sub classes	Single	Boosted	Ratio
Breast-w	3	4.19	4.36	1.042
Bupa	5	29.91	31.48	1.052
Credit-g	5	24.19	24.40	1.009
Crx	3	13.55	13.87	1.024
Diabetes	5	22.74	23.49	1.033
Heart-c	5	19.78	21.28	1.076
Ionosphere	3	12.74	14.25	1.118
Segment	6	5.48	6.16	1.123
Sonar	3	22.18	25.87	1.166
Vehicle	4	21.54	22.67	1.053
Mean				**1.07**

Table 5. Effect of forward feature selection on the Misclassification errors by Boosting.

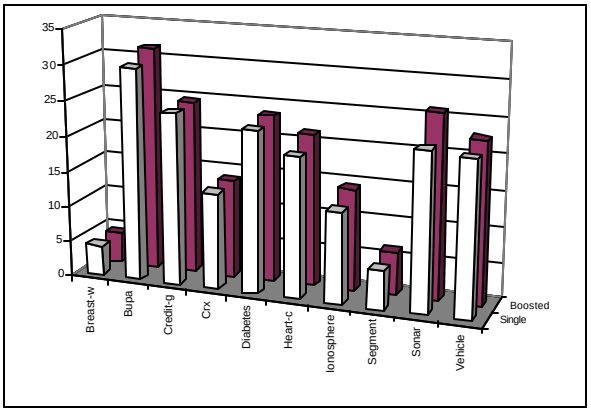

Figure 2 Comparison of the misclassification error rates for the single and boosted classifier after forward feature selection.

5 Concluding Remarks

Our experiments have lead us to the following conclusions:

a) On average, the use of bagging for gaussian mixtures classifiers uniformly reduces error misclassification rate and has a better perfomance than the Boosting algorithm when it is applied to classifiers based on Gaussian mixtures

b) Boosting classifiers is not appropriate to apply it to the classifiers GM, it does not reduce the probability of bad classification and in most of the cases it increases it.

c) The forward feature selection method applied to Gaussian mixtures classifiers yields good results. In average, a reduction in misclassification error rate is obtained, but not uniformly. In some cases, the misclassification error rate increases.

d) Ensembles of Gaussian mixtures classifiers improve after feature selection but as in the case of single classifiers this is not uniform.

6 References

ACUÑA, E. and ROJAS E. (2001). Bagging classifiers based on kernel density estimators. Proceedings of the International Conference on New Trends in Computational Statistics with Biomedical Applications. 343-350, Osaka, Japan.

BAUER, E. and KOKAVI, R. (1999): An empirical comparison of voting classification algorithms: Bagging, Boosting and variants. Machine Learning, 36, 105-139.

BLAKE, C. and MERZ, C. (1998): UCI repository of machine learning databases. Department of Computer Science and Information, University of California, Irvine, USA.

BREIMAN, L. (1996): Bagging Predictors. Machine Learning, 26,123-140.

BREIMAN, L. (1998): Arcing Classfiers. Annals of 5tatistic, 26,801-849.

DAZA, L. (2002): Combinación de Clasificadores basados en Mezclas Gaussianas. Tesis de maestria. Universidad de Puerto Rico. Recinto de Mayagüez.

DIETTERICH, T.G (2000): An Experimental comparison of three methods for constr-ucting Ensembles od decision trees: Bagging, Boosting, and randomization. Machine Learning, 26,801-849.

FREUND, Y. and SCHAPIRE, R. (1996): Experiments with a new boosting algorithm. In Machine Learning, Pi-oceedingq of the Thirteenth International Coriference, San Francisco, Morgan Kaufman, 148-156.

HASTIE, T. and TIBSHIRANI, R. (1994). Discriminant analysis by gaussian mixtures. Technical report, AT&T Bell Labs and University of Toronto.

KOHAVI, R. and JOHN, G.H. (1997): Wrappers for feature subset selection. Artificial Intelligence, 97, 273-324.

MACLIN, R. and OPTIZ, D. (1997): An empirical evaluation of Bagging and Bosting. Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI/Ml'r Press.

ORMONEIT, D. and TRESP V. (1995). Improved gaussian mixture density estimates using bayesian penalty terms an network averaging. Technical Report, University of Munich.

QUINLAN, J.R. (1996): Bagging, Boosting and C4.5. Proceedings of the Thirteenth National Conference on Artificial Intelligence, AAAI/MIT Press, 725-730.