Combining Classifiers Based on Kernel Density Estimators

Edgar Acuna
University of Puerto Rico at Mayaguez, Department of Mathematics
Mayaguez, PR 00680
edgar@math.uprm.edu

Alex Rojas
University of Puerto Rico at Mayaguez, Department of Mathematics
Mayaguez, PR 00680
alexr@math.uprm.edu

1. Introduction

A lot of research is being conducted on combining classification rules (classifiers) to produce a single one, known as an *ensemble*, which in general is more accurate than the individual classifiers making up the *ensemble*. Two popular methods for creating *ensembles* are *Bagging* introduced by Breiman, (1996), Ada*Boosting* by Freund and Schapire (1996). These methods rely on resampling techniques to obtain different training sets for each of the classifiers. Previous work has demonstrated that combining techniques are very effective for unstable classifiers, such as decision trees, neural networks and naive Bayes. In this paper we present some results in application of combining techniques to classifiers where the class conditional density is estimated using Kernel density estimators.

2. Classifiers based on Kernel Density estimators

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

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1h_2....h_p}\sum_{i=1}^{n}\prod_{v=1}^{p} K(\frac{x_v - \mathbf{X}_{iv}}{h_v})$$

where the kernel K will be usually a radially symmetric unimodal density function, for instance the multivariate normal density, and h_v represents the bandwidth of the v-th predictor. There are several approaches to select the optimal bandwidth. In the Statlog Project (Michie, et. al. 1994), where 23 classifiers are compared in 22 datasets, classifiers based on kernel density estimators (ALLOC80) performed better than CART (a 13-8 victory) and tied with C4.5 (11-11). However, ALLOC80 appeared as the top 5 classifier for 11 datasets whereas C4.5 and CART appeared only for 5 and 3 datasets respectively. Classifiers based on kernel density estimation are unstable due to singularities presented in the log-likelihood function, and the selection of the bandwidth is affected by the presence of outliers. For categorical predictors we have followed the Titterington's proposal.

3. Experimental Methodology

We chose 8 datasets coming from the Machine Learning Database at UCI to evaluate the effect of combining KDE classifiers considering mixed type of predictors: continuous and

categorical and fixed as well as adaptive bandwidths. For each dataset we have performed the following procedure: The 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 KDE 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 procedure is repeated 10 times. The misclassification error of a single classifier is estimated by a 10-fold crossvalidation and averaged over 10 runs. The bagged misclassification error is averaged on 100 repetitions. We also computed the relative error reduction (Improv.). A S-plus function has been developed to carry out all our tasks. The results are shown in the table below.

Dataset	Cases	Classes	Features	Classical Kernel			Adaptive Kernel		
				Single	Bagged	Improv	Single	Bagged	Improv
Iris	150	3	4	4.00	3.33	16.75	4.67	4.00	14.34
Glass	214	6	9	44.97	40.52	9.90	35.20	33.25	5.54
Heart-C	297	2	13	22.09	20.05	9.23	23.60	19.80	16.10
Breast-W	699	2	8	4.34	4.10	5.53	4.88	4.53	7.17
Diabetes	768	2	8	25.43	25.05	1.49	26.46	25.92	2.04
Vehicle	846	4	18	35.62	32.54	8.65	37.17	33.58	9.66
Credit-G	1000	2	24	37.00	35.21	4.84	36.17	34.21	5.42
Segment	2310	7	16	15.76	14.25	9.58	13.33	12.54	5.92

The average of the error reduction for the 8 datasets after Bagging using the classical Kernel was 8.25% whereas for the Adaptive Kernel was 8. 27%. When C4.5 classifer was bagged (Quinlan. 1997) the average error reduction for the same datasets was 9.07%, quite similar to our result. Presently we are carrying out feature selection to deal with the curse of dimensionality and so far we are getting good results.

REFERENCES

Breiman, L. (1996). Bagging Predictors. *Machine Learning,* 26,123-140.

Freund, Y and Schapire, R. (1996). Experiments with a new boosting algorithm. *In Machine Learning, Proceedings of the Thirteenth International Conference*, 148-156., San Francisco, Morgan Kaufman..

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

Michie, D., Spigelhalter, D.J. and Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Classification.* London: Ellis Horwood.

Titterington, D.M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics,* 22, 259-268.

RÉSUMÉ

L'objectif de ce travail est d'appliquer les techniques récentes de Bagging et Boosting aux classificateurs où la densité conditionnelle de classe est évaluée par les estimaeurs de densité employant la méthode de kernel.