

Supervised learning rule selection for multiclass decision with performance constraints

Nisrine Jrad, Edith Grall-Maës and Pierre Beuseroy
Institut Charles Delaunay, Université de technologie de Troyes, FRE CNRS 2848
12 rue Marie Curie, BP 2060, 10010 Troyes cedex - France
{nisrine.jrad, edith.grall, pierre.beuseroy}@utt.fr

Abstract

A procedure to select a supervised rule for multiclass problem from a labeled dataset is proposed. The rule allows class-selective rejection and performance constraints. The unknown probabilities are estimated with a Parzen estimator. A set of rules are built by varying the Parzen's smoothness parameter of the marginal probabilities estimates and plugging them into the statistical hypothesis rules. A criterion that assesses the quality of these rules is estimated and used to select a rule. Resampling and aggregation methods are used to show the efficiency of the estimated criterion.

1 Introduction

Classification of an unknown pattern into one of a finite number of known classes is a common problem in many fields. The performance of a classification system is characterized by a loss function, for example the error rate. For certain applications like currency verification or fraud detection, it can be more costly to make a wrong decision than to withhold making a decision. A reject option provides a means to control the loss. The class-selective rejection scheme allows to select only the best classes and to reject those that are most unlikely to issue the pattern. The required performance of a classification system can be fixed using constraints. The classification problem with class-selective rejection options and performance constraints has been tackled in [2]. A general formulation for this problem was proposed and the optimal decision rule within the framework of statistical theory was expounded. However, the conditional probabilities are usually not known because classification problems are generally described by a labeled database. In this case, one way pointed out in [3] is to determine a rule from an estimation of the unknown probabilities, using for example the Parzen estimator [6]. To assess the quality of a supervised learn-

ing rule a performance criterion was proposed in [3]. It takes into account the constraints and the loss function to be minimized. However the computation of the exact value of this criterion requires the exact density probabilities. Since the latest are unknown, an estimation of this criterion is necessary.

In this paper, we study the selection of a learning rule and the efficiency of the estimated criterion using only a labeled dataset. The aim is to propose a learning procedure and a quality assessment method. It is necessary to take maximum advantage of a given size dataset in order to select the "best" rule that guarantee a good generalization ability. The supervised rule selection involves two linked difficulties: the estimation of the rule and the estimation of its performance criterion. Observations are divided into a training set and a validation set. The training set serves to construct a set of rules depending on the smoothness parameter of the Parzen window. The validation set is used to estimate the criterion and consequently to choose the "best" rule. With a focus on model selection, three resampling methods are considered in this study: Holdout, Bootstrap and v-Cross Validation [4], [5]. However, if the learning algorithm is unstable for the resampled training set, Bagging procedure [1] is used to improve its stability.

This paper is organized as follows. Section 2 describes the multiclass decision problems with performance constraints. Section 3 presents the supervised learning rules and their quality assessment. In section 4 are exposed the resampling and the aggregation techniques. The efficiency of the estimated criterion is shown through a simulated example using the different resampling methods in section 5. Our work is concluded in section 6.

2 Multiclass decision problem

Assuming that the problem is characterized by N classes $w_1 \dots w_N$ and that any observation $x \in \mathbb{R}^d$ belongs to one class, a decision rule consists in a partition

Z of \mathfrak{R}^d in I sets Z_i composed of elements x assigned to the decision option ψ_i . In the class-selective rejection scheme, options are defined by admissible or a subset of classes (ie. $x \in \psi_i = \{1; 3\}$ means that x is assigned to both classes w_1 and w_3 with ambiguity). The probability that elements of w_j are assigned to ψ_i is:

$$P(D_i/w_j) = \int_{Z_i} P(x/w_j) dx$$

The problem consists in finding the decision rule Z^* that minimizes a given loss function $\bar{c}(Z)$ and respects K given constraints respectively defined by:

$$\bar{c}(Z) = \sum_{i=1}^I \sum_{j=1}^N c_{i,j} P_j P(D_i/w_j)$$

$$e^{(k)} = \sum_{i=1}^I \sum_{j=1}^N \alpha_{i,j}^{(k)} P_j P(D_i/w_j) \leq \gamma^{(k)}$$

γ^k are the threshold, $c_{i,j}$ and $\alpha_{i,j}^{(k)} \in \mathfrak{R}$ are the cost of deciding that an element x belongs to the set ψ_i when it belongs to the class w_j , in the expressions of the loss function and the constraints. P_j are the a priori probabilities. Z^* is the solution of the following problem:

$$\min_Z \bar{c}(Z) \quad \text{subject to } e^{(k)}(Z) \leq \gamma^{(k)} \quad \forall k = 1..K$$

The determination of the solution in the statistical decision theory was expounded in [2]. It consists in finding the saddle point (Z^*, μ^*) of the Lagrangian $L(Z, \mu)$ associated to the problem where μ is the vector of the lagrange multipliers $\mu_k \geq 0, k = 1..K$ associated with the constraints. It can be obtained by solving the optimization problem defined by $\max_{\mu \in \mathfrak{R}^{K+}} \{\min_Z L(Z, \mu)\}$. $\{\min_Z L(Z, \mu)\}$ is easy to obtain since it can be analytically determined.

3 Supervised learning and quality assessment of a decision rule

In the supervised learning framework, P_j and $P(D_i/w_j)$ are unknown. A solution for the decision problem is to use the decision rule obtained in statistical learning theory with an estimate of these probabilities. In this paper, we study the repercussions due to the estimation of $P(x/w_j)$ and we consider that P_j are known. Parzen window method [6] is used as an estimator of the unknown probabilities. The probabilities estimates depend on the training set and on the Parzen's smoothness parameter σ varied between $[0, 1]$ in order to balance between the bias and the variance of the estimation. Given a labeled dataset, two steps are essential to select the optimal decision rule. First, a set of rules Z_σ^* is built using the estimated conditional probabilities

$\hat{P}_\sigma(x/w_j)$ with different values of σ . μ_σ^* is the vector of the lagrange multipliers associated to Z_σ^* . Second, the performance of these rules are evaluated and the best rule $Z_{\sigma_{opt}}^*$ is chosen. To measure the performance of a rule, a criterion κ that takes into account the loss function and the constraints was proposed in [3]:

$$\kappa = \sum_{i=1}^I \sum_{j=1}^N (c_{i,j} + \mu^{*T} \alpha_{i,j}) P_j P(D_i/w_j) - \mu^{*T} \gamma \quad (1)$$

with $\alpha_{i,j} = [\alpha_{i,j}^{(k)}]_{k=1..K}$, $\gamma = [\gamma^{(k)}]_{k=1..K}$, μ^* is the vector associated to the theoretical rule and $P(D_i/w_j)$ are the real probabilities. This criterion, evaluated with the theoretical probabilities P_j and $P(D_i/w_j)$ and the lagrangian multipliers associated to the theoretical rule, is relevant [3]. It reaches its minimum when the constraints are satisfied and the loss function is minimum. However, even if this criterion shows good performance using theoretical knowledge, it is not evident to confirm that, without any theoretical information on the datasets, it keeps its performances. In this paper, the relevance of κ will be shown without using theoretical knowledge, which is more realistic. κ is estimated by:

$$\hat{\kappa}_{\mu, \hat{P}} = \sum_{i=1}^I \sum_{j=1}^N (c_{i,j} + \mu^T \alpha_{i,j}) P_j \hat{P}(D_i/w_j) - \mu^T \gamma \quad (2)$$

with \hat{P} an estimate of the probabilities evaluated by using the labeled data set and μ is chosen from the μ_σ^* .

If the generalization criterion is assessed on the same samples used to construct the decision, the selected rule may have poor ability to correctly predict new observations. To solve this problem, resampling techniques are used. They divide data into a training set and a validation set. The estimated rule is established using the training set. The criterion will be evaluated using the validation set to estimate the probabilities and the lagrangian multipliers obtained by the estimated rule.

It is important to compare different rules Z_σ^* using a unique estimator of $P(D_i/w_j)$ and the same value of μ . In our studies the empirical estimator is used. The choice of μ is a crucial point. Different strategies are developed and analyzed to select one.

4 Resampling and aggregation methods

In this section three resampling methods [4], [5] and the aggregation technique Bagging [1] are exposed. Resampling methods divide the original set into a training set and a validation set in order to select the best decision rule. McLachlan introduced the Holdout method [4], [5] that splits data into two mutually exclusive subsets. It is known for its ease of computation but it is a pessimistic biased method. First, since only a portion of the data is used to predict the rule, the learned model over-estimate the criterion. Second,

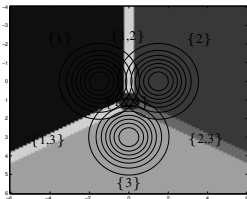


Figure 1. Theoretical density functions and partition

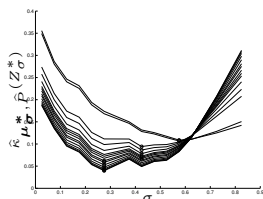


Figure 2. $\hat{\kappa}_{\mu^*, \hat{P}}(Z_\sigma^*)$ computed with all the values of μ_σ^*

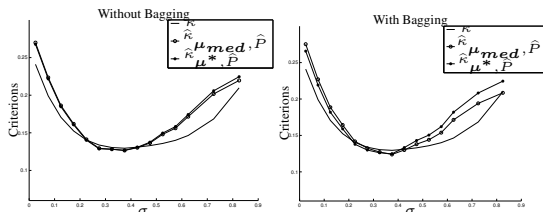


Figure 3. $\kappa(Z_\sigma^*)$ (solid), $\hat{\kappa}_{\mu_{med}, \hat{P}}(Z_\sigma^*)$ (circles) and $\hat{\kappa}_{\mu^*, \hat{P}}(Z_\sigma^*)$ (stars) in respect to σ , computed using 10-Bootstrap without Bagging (left) and with Bagging (right)

there is a bias caused by the independence of the learning and the validation sets. Also, since the rule is developed only once, the prediction has a large variance. To overcome these inconveniences, data can be randomly split, using v -Cross-Validation [4], [5], into v mutually exclusive subsets of approximately equal size. Subsequently, the learning set contains $(v - 1)$ of the partitions. The left partition serves as the validation set. The criterion is evaluated for each of the v sets and then averaged over v . A particular case is the Stratified Cross-Validation which divides the observations into v folds containing approximately the same proportion of labels as in the original set. Bias is known to decrease when v increases. Another method, Bootstrap [4], [5], creates a learning set by sampling, with replacement, n patterns uniformly from the dataset. The patterns left out $(0.368n)$ serve as the validation set. The learning set contains $0.632n$ unique observations which leads to an over-estimation of the criterion: a decrease on the learning set leads to an increase in the bias. Bootstrap samples are generated B times. The criterion is assessed B times and averaged over B .

To improve the predictive power of a learning rule, aggregation techniques are recommended. In this paper, Bagging [1] is used. It generates T rules $Z^{*t}(x)$ ($t = 1 \dots T$) for each pattern x of the training set built by one of the above resampling methods. The rule $Z^{*t}(x)$ is constructed on the t trial. The number of trials is assumed fixed. The multiple versions are obtained by

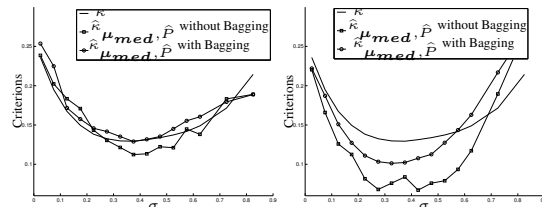


Figure 4. Two 1/2-Holdout experiments

making bootstrap replicates of the learning set and using them as the new learning sets. The bagged rule $Z^*(x)$ gives the final class of x . It is formed by aggregating the T rules by voting. Bagging gives a stable rule with unstable datasets at the cost of additional computation.

5 Simulation results and discussions

The considered problem is composed of 600 patterns $x \in \mathcal{R}^2$ drawn from three equiprobable Gaussian distribution presented on figure 1. The classification problem is described by: the seven possible decision options shown in figure 1, two constraints $P_E \leq 0.05$ and $P_I \leq 0.08$ where P_E is the probability of error and P_I the probability of indistinctness and the loss function defined by $\bar{c} = P_E + 0.5P_I + P_R$ with P_R is the probability of total rejection. The partition associated with the theoretical decision rule obtained by using the theoretical densities is represented in figure 1. It is obtained with a criterion κ_{theo} equal to 0.1157.

First, a set of decision rules Z_σ^* was developed using the whole original data and the criterion $\kappa(Z_\sigma^*)$ was assessed using μ^* and the theoretical probabilities by varying σ between 0.025 and 0.825. It reaches its minimum for $\sigma_{opt} = 0.375$. Second, data was divided into a training set and a validation set in order to determine $\hat{\kappa}$. It is then necessary to use a common value for μ in equation (2). A pragmatic choice is the median value μ_{med} of all the μ_σ^* . To study the performance of this choice, an experiment was carried out using 1/2-Holdout method. For each rule $\hat{\kappa}_{\mu_\sigma^*, \hat{P}}(Z_\sigma^*)$ was evaluated with each of the μ_σ^* . The results are reported in figure 2. The minimum values are shown by circles. Figure 2 shows that using extreme values of μ_σ^* to assess $\hat{\kappa}$ can lead to a wrong model selection.

Another experiment was carried out using the resampling methods to study the performance of the approximate criterion. The aim of this experiment is not to compare the efficiency of these methods but to show that the estimated criterion is as relevant as the theoretical one, despite all the imprecisions due to the estimation of the probabilities and the choice of the lagrangian multipliers. Datasets are split into a learning set and a validation set according to Holdout (the proportion

Table 1. Without Bagging: values of the real and estimated criterions and their appropriate optimal parameter σ_{opt} . Values of \hat{c} , \hat{P}_E and \hat{P}_I of the selected decision rule

	σ_{opt}	$\hat{\kappa}_{\mu_{med}, \hat{P}}(Z_{\sigma_{opt}}^*)$	$\hat{\kappa}_{\mu^*, \hat{P}}(Z_{\sigma_{opt}}^*)$	\hat{c}	\hat{P}_E	\hat{P}_I
5-CV	0.375	0.1029	0.1079	0.125	0.04	0.08
10-CV	0.325	0.0941	0.1029	0.1333	0.0333	0.0733
HO-1/3	0.325	0.1034	0.1048	0.11443	0.044776	0.079602
HO-1/2	0.425	0.0672	0.0901	0.13667	0.026667	0.066667
HO-1/2	0.375	0.1124	0.1125	0.11333	0.05	0.066667
Bootstrap (5)	0.375	0.1065	0.1065	0.11773	0.04375	0.078036
Bootstrap (10)	0.375	0.1266	0.1267	0.12251	0.052264	0.081167
$\kappa(Z_{\sigma_{opt}}^*) = 0.12954$	0.375			0.13114	0.050308	0.072512

of left data is varied), Stratified Cross-Validation (the number of folds is modified) and Bootstrap (the number of Bootstrap samples is changed). For each method, a set of rules was built using the training set with $\sigma \in [0.025, 0.825]$. The criterion was assessed again using μ^* , and noted $\hat{\kappa}_{\mu^*, \hat{P}}$, to show the impact of choosing μ_{med} when \hat{P} is used. Values of real and estimated criterions are given in table 1. Similar simulations were done with Bagging and results are as following:

- Table 1 and figure 3 show that $\hat{\kappa}_{\mu_{med}, \hat{P}}$ and $\hat{\kappa}_{\mu^*, \hat{P}}$ are similar and their minimum is reached for the same σ_{opt} leading to the same model selection. σ_{opt} is either equal or close to its theoretical value which proves that estimating the criterion using \hat{P} and μ_{med} on an appropriate validation set leads to chose the optimal rule.
- Table 1 and figure 4 show that, for two different 1/2-Holdout experiments, different models are selected. This poor performance is caused by the large bias.
- In the case under study, the performance of v-Cross-Validation is independent of the number v of folds when using moderate v .
- Table 1 shows that the results given by Bootstrap depend on B . For $B = 10$ (figure 3), Bootstrap performs the best and the estimated criterion follows approximately the real one. It is a relevant method due to its reliability and its moderate computational cost.
- Bagging is strongly recommended if disturbing the learning set causes significant changes in the selected rule. This is the case of rules built with Holdout. Figure 4 shows κ and $\hat{\kappa}_{\mu_{med}, \hat{P}}$ calculated with two 1/2-HO experiments, with and without Bagging in each experiment. To make the estimated criterion close to the real one, Bagging is required with $T = 150$ in the first experiment and $T = 10$ in the second one. However, Bagging seems to induce an extra cost computation with Bootstrap and Cross-Validation without subsequent performance improvement. Figure 3 shows that using Bagging does not lead to an improvement on the criterions determined by 10-Bootstrap. Similar results to those of table 1 obtained by using Bagging prove these claims.

6 Conclusion

The selection of a multiclass decision rule with class-selective options and performance constraints from a labeled set has been discussed. The proposed solution consists in using densities estimated by Parzen to solve the optimization problem. A set of rules is built using different values of the Parzen's smoothness parameter. To select a rule, an estimator of a performance criterion that assesses the quality of each rule is proposed. It uses the empirical conditional probabilities and the median value of the lagrange multipliers associated to the estimated decision rules. To obtain an accurate estimated of the criterion, resampling and aggregation techniques have been considered. Simulations on a problem with two constraints show that, for all resampling methods, the optimal criterion value is very close to the theoretical one. Furthermore, the estimated and the real criterions select the same rule. Thus the estimated criterion performs well and leads to find the optimal rule. Future work may tackle the problem with real data sets in a higher dimension space. Since estimating the probabilities and choosing convenient lagrange multipliers lead to very complex computations on a high dimension grid, other approaches, like kernel methods, should be used.

References

- [1] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [2] E. Grall, P. Beausery, and A. Bounsiar. Multilabel classification rule with performance constraints. In *Proceedings of IEEE conference ICASSP'06*, France, May 2006.
- [3] E. Grall, P. Beausery, and A. Bounsiar. Quality assessment of a supervised multilabel classification rule with performance constraints. In *EUSIPCO'06*, Italy, 2006.
- [4] R. Kohavi. *Wrappers for Performance Enhancement and Oblivious Decision Graphs*. PhD thesis.
- [5] A. Molinaro, R. Simon, and R. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.
- [6] F. van der Heijden, R. Duin, D. de Ridder, and D. Tax, editors. *Classification parameter estimation and state estimation*. John Wiley, England, 2004.