Adding a Reject Option to a Trained Classifier

RAMASUBRAMANIAN G. SUNDARARAJAN[1*] ASIM K. PAL[2]
[1] Information & Decision Technologies Lab
GE Global Research
Plot 122, EPIP Phase 2, Hoodi Village, Whitefield Road, Bangalore 560066
INDIA

[2] Professor of Information Systems & Computer Science
Indian Institute of Management Calcutta
D. H. Road, Joka, Kolkata 700104
INDIA

Abstract: - The option to reject an example in order to avoid the risk of a costly potential misclassification is well-explored in the pattern recognition literature. In this paper, we look at this issue from the perspective of statistical learning theory. Specifically, we look at ways of modeling the problem of adding a reject option to a trained classifier, in terms of minimizing an appropriately defined risk functional, and discuss the applicability thereof of some fundamental principles of learning, such as minimizing empirical risk and structural risk.

Key-Words: - Statistical learning theory, reject option, pattern recognition

1 Introduction

The primary focus of learning theory with regard to pattern recognition problems has been on algorithms that return a prediction on every example in the example space. However, in many real life situations, it may be prudent to reject an example rather than run the risk of a costly potential misclassification.

The issue of introducing a reject option in a learning machine has been dealt with extensively in the pattern recognition community. In many practical pattern recognition problems, the problem of incorporating a reject option falls into one of three categories: a) Minimize the loss of the classifier over the example space, given the costs of rejection and misclassification. b) Maximize the accuracy of the classifier, given that the rejection rate should not exceed a certain user-defined threshold. c) Minimize the rejection rate of the classifier, given that the accuracy should not go below a certain user-defined threshold.

This paper focuses on the first of these three problems. Typically, a learning algorithm first finds the optimal zero-reject hypothesis, and the optimal rejection region is then calculated on this hypothesis. We shall call this a *decoupled rejection scheme*.

The classical work in this regard, which is still used by many practitioners, is by Chow [3], which discusses the nature of the error-reject trade-off curve. This work states that, if the a posteriori class probabilities can be correctly estimated, the minimum risk rule is to choose the class with the highest posterior probability, provided that it exceeds the rejection threshold $T = \frac{b-c}{b+a}$, where

*This work was done as part of a fellow (doctoral) programme at IIM Calcutta

a denotes the gain from a correct classification, b denotes the cost of misclassification, and c denotes the cost of rejection. However, this strategy may not prove optimal in cases wherein the posterior probabilities are not accurately estimated. To deal with this drawback, methods such as class-related thresholds, and different thresholds for ambiguity and distance-based rejection, have also been explored in the literature [4, 7]. Also, for learning algorithms that do not output class probabilities, appropriate data-driven rejection methods have been applied.

An alternate approach to learning with a reject option is the case where the learning algorithm has an *embedded reject option*, i.e., it allows for rejection while training the underlying classifier, and finds the optimal rejection region and the optimal hypothesis on the predicted region simultaneously. Recently, there has been work on training a support vector machine with an embedded reject option [6], and on a modification of the perceptron algorithm with a reject option [10]. In this paper, however, we shall focus on the decoupled rejection scheme. A similar theoretical analysis for the embedded rejection scheme has been dealt with in [9].

Most of the theoretical analysis of the reject option has been based with regard to the setting wherein the classifier provides posterior probability estimates, on the basis of which the appropriate rejection strategy must be chosen. However, in many cases, the classifier used may not provide probability estimates, nor may it be easy to convert the classifier output thereof.

On the algorithmic side, methods more complicated than a simple rejection threshold on a single classifier have been explored, and their effectiveness demonstrated on various problems. Examples include ambiguity and distance related thresholds, class-related thresholds and voting schemes among classifiers [7, 4, 5].

However, to the extent of our knowledge, the issue of how this increase in complexity of the reject option trades off against a corresponding increase in performance, and how these two are related to the size of the training sample, has not been explored.

On the other hand, the analysis of data-driven methods of training a classifier without a reject option has been well studied in the statistical and computational learning theory literature. This theory deals with the study of the inductive principles that provide the motivation for algorithms that learn from a set of examples. In particular, it is concerned with the issue of how accuracy and complexity of a classifier are related to the sample size. However, it has not been extended to the case where a classifier may reject a given example.

In this paper, we try to address this gap by analyzing the problem of learning a reject option from a set of examples, from the standpoint of statistical learning theory.

Section 2 gives the mathematical formulation for the learning problem studied here, and discusses the applicability of some fundamental inductive principles such as empirical and structural risk minimization to this problem. Section 3 briefly analyzes the problem of training a classifier on one set of examples, and deciding upon the rejection region using another set. Finally, Section 4 presents some directions for further work along these lines.

2 The problem of learning a rejection hypothesis on a trained classifier

Let $S = ((x_1, y_1), \ldots (x_\ell, y_\ell))$ be a labeled i.i.d. sample drawn from an unknown but fixed joint distribution $F(x, y),\ x \in X,\ y \in Y = \{1, \ldots m\}$, where m is the number of classes. A learning algorithm L uses the sample S to arrive at a hypothesis $h \in H$, which, in case a reject option is permitted, will output $h(x) \in \{0\} \cup Y$, where 0 represents the reject option. Let $h^p \in H^p$ be the hypothesis (without a reject option) that is trained on the sample S. Let the loss function for h^p be defined as:

$$L_1(x, y, h^p) = \begin{cases} 0 & \text{if } h^p(x) = y \\ 1 & \text{if } h^p(x) \neq y \end{cases} \qquad (1)$$

The first stage of the learning problem is to find a classifier h^p_{opt} which minimizes the risk functional:

$$R_1(h^p) = \int L_1(x, y, h^p)\ dF(x, y) \qquad (2)$$

However, since the distribution $F(x, y)$ is unknown, a good inductive principle would be to choose the hypothesis h^p_ℓ that minimizes the empirical risk:

$$R_1^{emp}(S, h^p) = \frac{1}{\ell} \sum_{i=1}^{\ell} L_1(x_i, y_i, h^p) \quad (3)$$

on the basis of the sample S of size ℓ. Note, however, that finding the global minimum of the empirical risk may be non-trivial in a computational sense in many cases. Methods such as the backpropagation algorithm perform gradient descent on the hypothesis space to arrive at a local minimum of the empirical risk.

The second stage of the process is to add the reject option such that the overall cost is minimized. Let us consider a simplified scenario wherein the rationale for rejection of an example comes from the fact that the cost of rejection is lower than the cost of misclassification. We shall assume that the costs of misclassification and rejection (as well as the gain from correct classification) do not vary across classes. Without loss of generality, we can assume that the loss due to rejection of an example is $0 \leq \gamma \leq 1$.

Let $\theta \in \Theta$ be the hypothesis that decides whether or not the classifier should return a prediction on a given example. The compound hypothesis $h \in H(H^p, \Theta)$ can thus be rewritten as:

$$h = (1 - \theta(x, h^p_\ell))h^p_\ell(x) \quad (4)$$

The loss of the hypothesis h is:

$$\begin{aligned} L_2(x, y, h) &= \gamma\theta(x, y, h^p_\ell) + \\ &\quad (1 - \theta(x, y, h^p_\ell))L_1(x, y, h^p_\ell) \\ &= \begin{cases} 0 \text{ if } h(x) = y \\ \gamma \text{ if } h(x) = 0 \\ 1 \text{ if } h(x) \neq y, h(x) \neq 0 \end{cases} \end{aligned} \quad (5)$$

The objective of learning the reject problem is to find θ_{opt} that minimizes the risk functional:

$$\begin{aligned} R_2(h) &= \int L_2(x, y, h) dF(x, y) \\ &= R_1(h^p_\ell) + R_r(h) \end{aligned} \quad (6)$$

where

$$R_r(h) = \int \theta(x, h^p_\ell)(\gamma - L_1(x, y, h^p_\ell)) dF(x, y) \quad (7)$$

R_r can take values in the range $[\gamma - 1, \gamma]$, and represents the gain/loss in opportunity due to rejection. Since R_1 is constant as far as the second stage is concerned, the objective is to find θ such that the second term in equation (6) is as negative as possible. If it turns out to be positive, then a better strategy would obviously be to return a prediction on all examples. The learning problem in the second stage can now be restated as one of finding θ_{opt} that minimizes R_r.

It is illustrative to see that, by analyzing the inequality $R_r(h) \leq 0$, we arrive at the conclusion:

$$\frac{\int_{\theta(x,h^p_\ell)=1,\ L_1(x,y,h^p_\ell)=0} dF(x, y)}{\int_{\theta(x,h^p_\ell)=1} dF(x, y)} \leq 1 - \gamma$$

In other words, the probability of correct classification of an example in the rejected region should not exceed $1 - \gamma$, which is in concordance with the result given by Chow's threshold.

This method of modeling the risk functional is analogous to that given by the Local Risk Minimization principle [11], where the learning problem is to minimize

$$R_{loc}(h) = \int \theta(x, h^p) L_1(x, y, h^p)\ dF(x, y) \quad (8)$$

However, this method differs from the one described in this paper in that it works on the assumption that θ is fixed before optimizing h^p. Besides, since it does not count the cost of rejection, optimizing θ and h^p together, or optimizing θ after h^p could lead to a compound hypothesis h that operates on an extremely restricted section of the example space where no misclassifications are made.

In the next subsection, we shall discuss the applicability of the empirical risk minimization (ERM) principle for the learning problem described here.

2.1 The ERM principle for the decoupled rejection scheme

Let θ_ℓ be the rejection hypothesis that minimizes the empirical risk:

$$R_r^{emp}(h) = \frac{1}{\ell}\sum_{i=1}^{\ell} L_r(x_i, y_i, h) \tag{9}$$

where

$$L_r(x, y, h) = \theta(x, h_\ell^p)(\gamma - L_1(x, y, h_\ell^p)) \tag{10}$$

We wish to bound the risk R_2 of the hypothesis $h_\ell(h_\ell^p, \theta_\ell)$ as a function of its empirical risk and the complexity of the hypothesis classes used. The basic results used to arrive at these bounds are given in Vapnik [11].

From [11], we know that the risk of the underlying classifier (R_1) is bounded by the following inequality, with confidence $1-\eta$

$$R_1(h_\ell^p) \leq R_1^{emp}(S, h_\ell^p) + C_p \tag{11}$$

where

$$C_p = \frac{\varepsilon_p(\ell)}{2}\left(1 + \sqrt{1 + \frac{4R_1^{emp}(S, h_\ell^p)}{\varepsilon_p(\ell)}}\right) \tag{12}$$

$$\varepsilon_p(\ell) = 4\frac{d_p(\ln\frac{2\ell}{d_p} + 1) - ln\frac{\eta}{4}}{\ell} \tag{13}$$

The quantity d_p denotes the VC dimension of the loss function L_1, defined using the hypothesis class H^p.

The growth function $G^\Theta(\ell)$ of $\theta(x, h_\ell^p)$ is bounded by a function of d_r, the VC dimension of Θ, using Sauer's lemma. There are three possible values of $L_r(x, y, h)+1-\gamma$. Using the growth function of the indicator function $I((L_r(x, y, h)+1-\gamma)-\delta), 0 \leq \delta \leq 1$, we get a bound on $G^{L_r}(\ell)$ as:

$$G^{L_r}(\ell) \leq \ln\left(3\left(\frac{e\ell}{d_r}\right)^{d_r}\right) \tag{14}$$

We can apply this inequality to get the following bound for R_r, with confidence $1-\eta$

$$R_r(h) \leq R_r^{emp}(S, h) + C_r \tag{15}$$

where

$$C_r = \frac{\varepsilon_r(\ell)}{2}\left(1 + \sqrt{1 + \frac{4(R_r^{emp}(S, h) + 1 - \gamma)}{\varepsilon_r(\ell)}}\right) \tag{16}$$

$$\varepsilon_r(\ell) = 4\frac{d_r(\ln\frac{2\ell}{d_p} + 1) - ln\frac{\eta}{12}}{\ell} \tag{17}$$

By combining equations (11) and (15), the effective bound on R_2 is therefore given by the following inequality, with confidence $1-2\eta$

$$\begin{aligned} R_2(h_\ell) \leq\ & R_1^{emp}(S, h_\ell^p) + C_p \\ & + R_r^{emp}(S, h_\ell) + C_r \end{aligned} \tag{18}$$

In terms of practical applicability, we find that these bounds are reasonable only as long as the sample size ℓ is sufficiently larger than d_p and d_r. As the complexity of the classifier and the rejection hypothesis increase with respect to the number of examples, the value of the complexity terms C_p and C_r increase to the point where the upper bound for the risk exceeds 1.

2.2 VC dimension results for some rejection schemes

In order to apply these bounds in practical situations, it is necessary to know the VC dimension (or bounds thereof) for some commonly used rejection hypotheses. We present a few basic results here in this regard. Let $\Phi(i|x),\ i = 1 \ldots m$ be the output of a classifier for an m-class problem, where Φ represents the strength of the prediction for each class, given an input example x.

The VC dimension of a simple threshold on the strength of the output is 1. This is proved trivially, since no more than one example on the real line can be shattered (i.e., classified in all possible ways) by the set of simple threshold functions.

Fumera & Roli [4] proposed a system of class-related thresholds to improve classification accuracy.

$$\theta(x,\Phi) = \begin{cases} 0 & \text{if } \Phi(i_1|x) > T_{i_1} \\ 1 & \text{otherwise} \end{cases} \tag{19}$$

where $i_1 = \arg\max_{i=\{1\ldots m\}} \Phi(i|x)$. The VC dimension of the above system in an m-class problem is equal to m. By an extension of the previous result, it is trivial to prove that a configuration of m points can be shattered by a system of m thresholds. Using the pigeonhole principle, we can then show that any additional example would fall into one of the m classes, and the threshold for that class cannot shatter both points.

Le Cun et al [7] applied the following system for rejection in an application to handwritten character recognition: Let $\Phi(i|x)$ be defined as in the previous case. The rejection hypothesis $\theta(x,\Phi)$ takes a value of 0 if the following two conditions are satisfied:

$$\begin{aligned} \Phi(i_1|x) &> T_1 \\ \Phi(i_1|x) - \Phi(i_2|x) &> T_2 \end{aligned} \tag{20}$$

where i_1 is defined as in the previous case, and $i_2 = \arg\max_{i=\{1\ldots m\}\{i_1\}} \Phi(i|x)$. The VC dimension of this rejection system is 2. It can be proved that, for a set of examples x^1, x^2 such that $\Phi(i_1^1|x^1) > \Phi(i_1^2|x^2)$ and $\Phi(i_1^2|x^2) - \Phi(i_2^2|x^2) > \Phi(i_1^1|x^1) - \Phi(i_2^1|x^1)$, one can find thresholds T_1 and T_2 such that all possible labelings of x_1 and x_2 with respect to θ can be achieved. Through a simple ordering argument, one can show that there cannot exist three points that can be shattered by this system.

2.3 Minimizing structural risk

Here we briefly discuss the concept of capacity control for our learning problem. It is clear from equation (18) that the tightness of the bound on the actual risk of the hypothesis h_ℓ depends on two factors: the empirical risk, and the complexity of the hypothesis classes.

In case of a coupled rejection scheme, the entire learning happens in one stage, hence it is easier to look at the issue of capacity control in terms of the growth function of the compound hypothesis h. In a decoupled scheme, however, one would optimize the bias-variance trade-off for the underlying classifier h^p first, and then optimize the trade-off for the rejection hypothesis θ. Therefore, it may happen that the stage-wise optimization may produce a sub-optimal solution, as compared to one wherein both trade-offs are optimized together. However, in many practical situations, it may be easier to do a stage-wise optimization.

3 Use of training and validation sets

While implementing a reject option in real-life problems, many practitioners adopt a two-sample approach, wherein the classifier h^p is first trained with a sample S_1 of size ℓ_1, and the reject option θ is learned with a sample S_2 of size ℓ_2. This procedure automatically begs the question: *Given a sample of size ℓ, how do we optimally split it into subsamples of size $\kappa\ell$ and $(1-\kappa)\ell$?*

If we make the simplifying assumption that, for a given sample, both the hypothesis classes are rich enough to achieve the minimum possible value of empirical risk, then the tightness of the bound depends only on $\varepsilon_p(\ell)$ and $\varepsilon_r(\ell)$. However, there does not exist a closed-form solution for the minimum of the function $g(\kappa) = \varepsilon_p(\kappa\ell) + \varepsilon_r((1-\kappa)\ell)$.

Preliminary empirical analysis of $g(\kappa)$ suggests that, when ℓ is sufficiently larger than d_p and d_r, the relationship between $\frac{d_p}{d_r}$ and $\frac{\kappa_0}{1-\kappa_0}$ is given by:

$$\kappa_0/(1-\kappa_0) = a_1 - a_2 \exp\Big(- a_3(\sqrt{d_p/d_r} - a_4)\Big) \tag{21}$$

A reasonably good fit is also given by a linear relationship between $\kappa_0/(1-\kappa_0)$ and $\sqrt{d_p/d_r}$; however, it does not capture the non-linearity very well. Note that, while the value of κ_0 that minimizes $g(\kappa)$ gives the tightest bound under the assumptions stated above, it does not guarantee an optimal split.

From a practitioner's point of view, the optimal value of κ can be used to construct an experiment wherein the total sample of size ℓ is repeatedly divided into two subsamples and used for training the classifier

and the rejection hypothesis. This approach is akin to that of k-fold cross-validation, and helps in desensitizing the classifier to idiosyncrasies of a particular split.

4 Scope for further work

In this paper, we have approached the problem of learning with a decoupled reject option from the point of view of statistical learning theory. To this end, we have presented a two-stage formulation of the learning problem, and discussed the applicability of the ERM principle thereof. Finally, we have discussed the issue of using two samples, one for training the classifier, and the other for learning the reject option.

This analysis can be extended to other methods of applying a reject option, such as an ensemble of classifiers combined using a voting scheme. The complexity of the voting scheme (or any other combiner), as well as the complexity of the individual classifiers, will then determine the risk bound.

References

[1] M. Anthony and N. Biggs, *Computational Learning Theory - An Introduction*, ser. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 1992, no. 30.

[2] Y. Baram, Partial classification: The benefit of deferred decision, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 769–776, 1998.

[3] C. K. Chow, On optimum recognition error and reject trade-off, *IEEE Transactions on Information Theory*, vol. 16, pp. 41–46, 1970.

[4] G. Fumera and F. Roli, Multiple reject thresholds for improving classification reliability, Univ. of Calgary, Tech. Rep., 1999.

[5] G. Fumera and F. Roli, Analysis of error-reject trade-off in linearly combined classifiers, in *Proceedings of the International Conference on Pattern Recognition (ICPR 2002)*, vol. 2, 2002.

[6] G. Fumera and F. Roli, Support vector machines with embedded reject option, in *Pattern Recognition with Support Vector Machines - First International Workshop, Proceedings*, S.-W. Lee and A. Verri, Eds. Springer, 2002.

[7] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Handwritten digit recognition with a backpropagation network, in *Advances in Neural Information Processing Systems*, vol. 2. Morgan Kaufmann, 1990, pp. 396–404.

[8] J. M. R. Parrondo and C. Van den Broeck, Error vs. rejection curve for the perceptron, *Europhysics Letters*, vol. 22, pp. 319–324, 1993.

[9] R. Sundararajan and A. K. Pal, Learning with an embedded reject option, Indian Institute of Management Calcutta, Tech. Rep., 2004.

[10] R. Sundararajan and A. K. Pal, A conservative approach to perceptron learning, 2004, accepted at 5th WSEAS International Conference on Neural Networks & Applications (NNA'04).

[11] V. N. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.