Communications of the Korean Statistical Society Vol. 15, No. 4, 2008, pp. 543–549

Empirical Choice of the Shape Parameter for Robust Support Vector Machines

Ro Jin Pak[1)]

Abstract

Inspired by using a robust loss function in the support vector machine regression to control training error and the idea of robust template matching with M-estimator, Chen (2004) applies M-estimator techniques to gaussian radial basis functions and form a new class of robust kernels for the support vector machines. We are specially interested in the shape of the Huber's M-estimator in this context and propose a way to find the shape parameter of the Huber's M-estimating function. For simplicity, only the two-class classification problem is considered.

Keywords: Huber's M-estimator; support vector machine; two-class classification.

1. Introduction

Combining robust estimation with support vector machine(SVM) turns out very effective in improving the performance of SVM, severely affected by gross error (Schölkopf and Smola, 2002; Mangasarian and Musicant, 2000). The Huber's M-estimating function as objective function or cost function is most often used. Although the idea of a robust M-estimator based loss function has been applied in support vector regression to control training error, the resultant effect on margin maximizing is obscure (Schölkopf and Smola, 2002). By contrast, by applying robust M-estimator to kernels, both margin and training errors are explicitly forced to be robust (Chen, 2004).

Recent works concerning robustness of SVM witnessed the usefulness of M-estimating functions, specially the Huber's function, however, the main concern of these works was not on the shape of the Huber's function. In this article we are concerned about the shaping parameter of the Huber's function. It is reported that the numbers for the shaping parameter, suggested by Huber (1981), are quite useful in robust SVM (Schölkopf and Smola, 2002). However, these numbers are based on the assumption that data are from the normal probability distribution and that assumption is not often guaranteed in practice. The numbers are often determined by a trial and error through computer simulations.

We are interested in finding a proper(empirical) value of the shaping parameter for a given data set. In this article, we propose a method to find a proper value, defined as the value maximizing expected accuracy for a two-class classification.

The remainder of this article is organized as follows. In Section 2, the (approximate) expected accuracy in a two-class classification is derived in terms of a shaping parameter.

1) Professor, Department of Information & Statistics, Dankook University, Yongin, Kyunggido, 448-701, Korea. E-mail: rjpak@dankook.ac.kr

The method and an example of finding an empirical value of a shaping parameter are present in Section 3.

2. Expected Accuracy

A model H, with a k-dimensional parameter vector $\mathbf{w}$, consists of its functional form f, the distribution $P(D|\mathbf{w}, H)$ that the model makes about the data D and a prior parameter distribution $p(\mathbf{w}|\lambda, H)$ with a regularization parameter λ. The first level of inference infers the posterior distribution of $\mathbf{W}$ for a given value of λ by using the Bayes' rule:

$$p(\mathbf{w} \mid D,\ \lambda,\ H) \propto p(\mathbf{w} \mid \lambda, H)p(D \mid \mathbf{w}, H).$$

Consider the following probability model:

- The prior over $\mathbf{w}$ is the Gaussian prior

$$p(\mathbf{w} \mid \lambda, H) \propto \exp\left(-\frac{\lambda}{2} \parallel \mathbf{w} \parallel^2\right).$$

- The probability distribution $p(y_i \mid \mathbf{x}_i, \mathbf{w}, H)$ for $y_i = \pm 1$ is given by

$$\begin{aligned} p(y_i \mid \mathbf{x}_i, \mathbf{w}, H) &= \frac{\exp\left(-[1 - y_i a_i]_+\right)}{\exp\left(-[1 - a_i]_+\right) + \exp\left(-[1 + a_i]\right)} \\ &\simeq \exp\left(-[1 - y_i a_i]_+\right) \\ &= \exp\left(-\xi_i\right), \end{aligned} \tag{2.1}$$

where $a_i = a(\mathbf{x}_i, \mathbf{w})$ and $[u]_+ = uI_{u>0}$. For details about the above results, refer Kwok (2000).

Let Ω=$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n) \mid \mathbf{x}_i \in R^d,\ y_i \in \{-1, 1\},\ i = 1, 2, \ldots, n\}$ be a set of n input-output training data pairs, where R^d space is referred to as the input vectors onto some feature space by a non-linear function ϕ and then finds a linear separating hyperplane in the feature space H, that is, $\mathbf{w}'\mathbf{x} + b = 0$ where $\mathbf{w} \in H$ and $b \in R$. SVM classifiers are obtained by solving the object function

$$\min_{\mathbf{w}, b, \epsilon_i} \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{n} \epsilon_i \tag{2.2}$$

subject to

$$y_i(\mathbf{w}'\boldsymbol{\phi}(\mathbf{x}_i)) + b \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \ \text{ for all } \ i = 1, \ldots, n,$$

where the constant $C > 0$ determines the tradeoff between margin maximization and training error minimization.

After solving the above optimization by using Lagrangian techniques, we can obtain the classification rule SVM

$$\operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right),$$

where $K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}_i)$ and α_is are Lagrange multipliers. The commonly used Mercer's kernels include Gaussian radial based function(RBF),

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right), \tag{2.3}$$

where $\sigma \in R$ is a kernel parameter. Let a d-dimensional pattern (object) $\mathbf{x}$ have n coordinates, $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, where $x_i \in R$ for $i = 1, 2, \ldots, d$. Note that $||\mathbf{x} - \mathbf{x}'||^2$ is typically defined as

$$||\mathbf{x} - \mathbf{x}'||^2 = \left(\sum_{i=1}^{d} |x_i - x_i'|^2\right)$$

and is called as the sum of square differences(SSD) in the signal processing society. Similar to the definition of SSD and the idea in (Chen, 2004), we can define the sum of robust differences(SRD) between x and x' as follows:

$$\text{SRD}_{\rho,\gamma}(\mathbf{x}, \mathbf{x}_i) = \sum_{j=1}^{d} \rho(|x_i - x_i'|, \gamma), \tag{2.4}$$

where $\rho(,)$ is a robust error measure(or a robust loss function) and γ is a parameter controlling the shape of $\rho(,)$. For example, for ρ in (2.4) we may use the Huber M-estimating function

$$\rho(z, \gamma) = \begin{cases} \dfrac{z^2}{2}, & \text{if } z \leq \gamma, \\ \gamma\left(z - \dfrac{\gamma}{2}\right), & \text{otherwise.} \end{cases}$$

If we replace ρ in (2.4) by the ρ of Huber M-estimating function, we can get a Huber M-estimator based SRD as a similarity measure. By adding SRD to Gaussian RBF kernels (*i.e.*, by combining (2.3) and (2.4)), a M-estimator based robust Gaussian RBF kernels (Chen, 2004) is defined as

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{\frac{\text{SRD}_{\rho,\gamma}(\mathbf{x}, \mathbf{x}_i)}{\sigma^2}\right\}. \tag{2.5}$$

In order to utilize the Chen's proposal we need proper values for γ and σ, respectively. Chen (2004) did a full search in some regions specified in advance to get the values, but of course it is time consuming and there is no guarantee that we are searching on right regions. In this article, we propose to choose values for the parameters in the sense of optimizing certain criterion. We consider the two-class classification problem and let expected accuracy as a criterion.

Denote y_i^{pos} and y_i^{pri} are the posterior class label and the prior class label, respectively. Let $N = \{i \mid y_i^{pri} = -1 \text{ with probability } p\}$ and $n_{(-1,-1)}$, $n_{(1,1)}$ be the numbers of cases correctly classified, respectively and n be the total number of observations then the

expected frequency

$$E\left[n_{(-1,-1)}+n_{(1,1)}\right] = E\left[\sum_{i=1}^{n} I(y_i^{pos}=-1,\, y_i^{pri}=-1)+I(y_i^{pos}=1, y_i^{pri}=1)\right]$$
$$= \sum_{i=1}^{n} P\left[y_i^{pos}=-1,\, y_i^{pri}=-1\right] + P\left[y_i^{pos}=1,\, y_i^{pri}=1\right]$$
$$= \sum_{i=1}^{n} P\left[y_i^{pos}=-1 \,|\, y_i^{pri}=-1\right] p + P\left[y_i^{pos}=1 \,|\, y_i^{pri}=1\right](1-p).$$

The approximate posterior probabilities can be expressed by using (2.1) as the function of γ as

$$P[y_i^{pos}=-1|\, y_i^{pri}=-1] \;\propto\; \exp\left(-[1+a_i^-(\gamma)]_+\right)$$
$$= \begin{cases} \exp(-1-a_i^-(\gamma)), & \text{if } a_i^-(\gamma) > -1, \\ 1, & \text{otherwise} \end{cases}$$

and

$$P[y_i^{pos}=1|\, y_i^{pri}=1] \;\propto\; \exp\left(-[1-a_i^+(\gamma)]_+\right)$$
$$= \begin{cases} \exp(-1+a_i^+(\gamma)), & \text{if } a_i^+(\gamma) < 1, \\ 1, & \text{otherwise,} \end{cases}$$

where $a_i^-(\gamma) = -\sum_{i\in N}\alpha_i K_\gamma(\mathbf{x},\mathbf{x}_i)+b$ and $a_i^+(\gamma)=\sum_{i\in N^c}\alpha_i K_\gamma(\mathbf{x},\mathbf{x}_i)+b$. Then the expected accuracy,

$$E\left[\frac{n_{(-1,-1)}+n_{(1,1)}}{n}\right] \;\propto\; p\sum_i \exp\left(-\left[1-\sum_{j\in N}\alpha_j K_\gamma(\mathbf{x}_i,\mathbf{x}_j)+b\right]_+\right)$$
$$+(1-p)\sum_i \exp\left(-\left[1-\sum_{j\in N^c}\alpha_j K_\gamma(\mathbf{x}_i,\mathbf{x}_j)-b\right]_+\right),$$

which is maximized when

$$\left[1-\sum_{j\in N}\alpha_j K_\gamma(\mathbf{x}_i,\mathbf{x}_j)+b\right]_+ \quad \text{and} \quad \left[1-\sum_{j\in N^c}\alpha_j K_\gamma(\mathbf{x}_i,\mathbf{x}_j)-b\right]_+$$

are as small as possible for each i.

3. Proper Choice of the Shape Parameter

We propose that the proper γ should be the value which makes both

$$\left[1-\sum_{j\in N}\alpha_j K_\gamma(\mathbf{x}_i,\mathbf{x}_j)+b\right]_+ \quad \text{and} \quad \left[1-\sum_{j\in N^c}\alpha_j K_\gamma(\mathbf{x}_i,\mathbf{x}_j)-b\right]_+ \tag{3.1}$$

to be 0 as close as possible in order for (2.6) to be maximized.

Example 3.1 The Exclusive-OR Problem

The exclusive-or problem is as follows: Find an optimal separating hyperplane that classifies the following data set without error (*i.e.* $C = \infty$ and $\epsilon_i = 0$ in (2) for all i):

Index i	x	y
1	(1, 1)	1
2	(1, −1)	−1
3	(−1, −1)	1
4	(−1, 1)	−1

Here, $N = \{2, 4\}$ and $N^c = \{1, 3\}$. Suppose the kernel is the Gaussian RBF, then we have eight equations from (3.1) but only two distinct equations are relevant among them;

$$1 - (1.3375)\left[\exp\left\{-\rho(|(0,0)|, \gamma)\right\} + \exp\left\{-\rho(|(2,2)|, \gamma)\right\}\right],$$

and

$$1 - (1.3375)\left[\exp\left\{-\rho(|(0,-2)|, \gamma)\right\} + \exp\left\{-\rho(|(-2,0)|, \gamma)\right\}\right].$$

According to the ranges of γ, we have

$$\gamma \le 2: \quad 1 - (1.3375)\left[\exp(0) + \exp\left\{\frac{-\gamma\left(\sqrt{8} - \frac{\gamma}{2}\right)}{\sigma^2}\right\}\right] \quad \text{and} \tag{3.2}$$

$$1 - (1.3375)\left[2\exp\left\{\frac{-\gamma\left(2 - \frac{\gamma}{2}\right)}{\sigma^2}\right\}\right]. \tag{3.3}$$

$$2 < \gamma \le \sqrt{8}: \quad 1 - (1.3375)\left[\exp(0) + \exp\left\{\frac{-\gamma\left(\sqrt{8} - \frac{\gamma}{2}\right)}{\sigma^2}\right\}\right] \quad \text{and} \tag{3.4}$$

$$1 - (1.3375)\left[2\exp\left(-\frac{2}{\sigma^2}\right)\right]. \tag{3.5}$$

$$\gamma > \sqrt{8}: \quad 1 - (1.3375)\left[\exp(0) + \exp\left(-\frac{4}{\sigma^2}\right)\right] \quad \text{and} \tag{3.6}$$

$$1 - (1.3375)\left[2\exp\left(-\frac{4}{\sigma^2}\right)\right]. \tag{3.7}$$

The expressions in (3.2), (3.4) and (3.6) are always negative and the expression in (3.3), (3.5) and (3.7) is negative on the dark area and positive on the bright area of Figure 3.1. The expected accuracy is maximized at every combination of (σ, γ) on the dark areas. We have three areas to consider,

- $\{(\sigma, \gamma) \in \mathbb{R}^2 | 0 < \sigma \le 1.4255,\ 0 < \gamma \le 2,\ 1 - (1.3375)[2\exp\{-\gamma(2 - \gamma/2)/\sigma^2\}] \le 0\}$ [Figure 3.1(a)].
- $\{(\sigma, \gamma) \in \mathbb{R}^2 | 1.4255 < \sigma \le 2.0162,\ 2 < \gamma \le \sqrt{8}\}$ [Figure 3.1(b)].
- $\{(\sigma, \gamma) \in \mathbb{R}^2 | \sigma > 2.0162,\ \gamma > \sqrt{8}\}$ [Figure 3.1(c)].

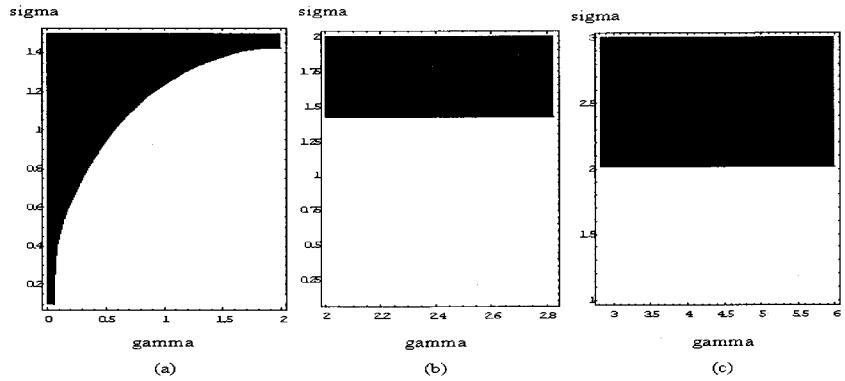

Figure 3.1: Contour plots for the expression in (3.3), (3.5) and (3.7), respectively, from (a) to (c) over γ and σ. The expressions are negative on dark area and positive on bright area

From the robust statistical point of view, we better choose the shape parameter γ as large as possible as long as robustness is maintained (Huber, 1981). In practice, suppose an estimate $\hat{\sigma}$ for σ in (2.5) is available, and

- if $0 < \hat{\sigma} \leq 1.4255$, our choice of the γ should be the corresponding value of the γ on the border line of the contour plot [Figure 3.1(a)],
- if $1.4255 < \hat{\sigma} \leq 2.0162$, our choice of the γ would be $\sqrt{8}$ [Figure 3.1(b)],
- if $\hat{\sigma} > 2.0162$, our choice of the γ would be ∞ [Figure 3.1(c)].

4. Conclusions

In this article, we propose the method of finding a shaping parameter of the Huber's function for the RBF kernels in a two-class classification problem. The shaping parameter γ is very much related with a scale parameter σ and choice of γ is under control of a scale parameter. In this sense, it requires a more through study on estimating a scale parameter in the sequel. Though the scope of this article is too narrow to be little difficult to generalize the discoveries, we raise the possibility that the Guassian RBF itself is quite enough to handle robustness regardless of M-estimating functions.

References

Chen, J. (2004). M-estimation based robust kernels for support vector machines, In *Proceeding of the* 17^{th} *International Conference on Pattern Recognition*, 168–171.

Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Kwok, J. T. Y. (2000). The evidence framework applied to support vector machines, *IEEE Transactions on Neural Networks*, **11**, 1162–1173.

Mangasarian, O. L. and Musicant, D. R. (2000). Robust linear and support vector regression. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **22**, 950–955.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels-Support Vector Machines, Regularization, Optimization and Beyond*, The MIT Press, Cambridge.

[Received May 2008, Accepted June 2008]