http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 8 (4): 615-618, 2009
ISSN 1812-5638

A Novel Minimax Probability Machine

[1,2]Mu Xiangyang and [1]Zhang Taiyi
[1]School of Information and Communication Engineering, Xi'an Jiaotong University, China
[2]School of Electronical Engineering, Xi'an Shiyou University, China

Abstract: This study presents an empirical study for Minimax Probability Machines (MPM) for prediction. Considering that the Euclidean distance has a natural generalization in form of the Minkovsky's distance, a novel MPM using Minkovsky's norm in Gaussian kernel function is proposed. The performance of proposed method is evaluated with the prediction for Ethernet traffic data. Result shown that the novel MPM here in using Gaussian kernels with Minkovsky's distance (α=1) and (α=5) can achieve better prediction accuracy than the Euclidean distance.

Key words: Support vector machine, minimax probability machines, gaussian kernel, Minkovsky's norm

INTRODUCTION

Minimax Probability Machine (MPM) (Lanckriet *et al*., 2002a, b) provides a lower bound on classification accuracy based on reliable estimates of means and covariance matrices of the classes from the training samples and achieves the comparative performance with the Support Vector Machine (SVM) recently (Burges, 1998). Unlike SVM, for which the samples close to the decision boundary are most important, the MPM find the margin between the means of both classes, which rather represent the typical examples of each of the classes. Furthermore, there is one constraint per class in MPM. Similarly to SVM, although the MPM is originally designed for classification task, further study extended the model to regression and prediction (Huang *et al*., 2004).

The performance of both MPM and SVM depends on the kernel functions (Perez-Cruz and Bousquet, 2004). There are many kinds of kernels such as polynomial kernel, Gaussian kernel, and tangent distance kernel, that can be used. Every kernel has its advantages and disadvantages. Unfortunately, there are currently no theories available to learn the form of the kernels. Among the possible kernels, the most used in practice is Gaussian kernel.

Due to the encouraging results with Gaussian kernel, Ribeiro (2002) investigated the SVM with a more generalized form of Gaussian kernel based on Minkovsky's distance measure. It is value for us to investigate the problem of whether a good performance could be obtained when the MPM using Gaussian kernels with Minkovsky's distance. This idea is tested on Ethernet traffic data in this study. Experiments shown that the MPM using Gaussian kernels with Minkovsky's distance can present better prediction accuracy than the Euclidean distance.

REVIEW ON MINIMAX PROBABILITY MACHINE

The linear version of MPMR: In this subsection, we outline the theoretical derivatives of a linear MPM for regression and prediction.

Given the training samples set D_N, the MPM would like to estimate the regression or prediction function f (x) by finding a model that maximizes the minimum probability of being $\pm\varepsilon$ accurate.

$$\max\left[\min P\left(\left|f(x)-t\right|\le\varepsilon\right)\right] \quad (1)$$

Assume the function f (x) has the form of

$$f(x)=w^{T}x+b_r=w^{(1)}x^{(1)}+w^{(2)}x^{(2)}+\cdots+w^{(d)}x^{(d)}+b_r \quad (2)$$

Where:

$$w=\left(w^{(1)},w^{(2)},\cdots,w^{(d)}\right)^{T}$$

MPM formulates the Eq. 1 as a binary classification problem to determine the parameters w and b_r.

We take the training samples set D_N and create two new classes $\tilde{x}$ and $\tilde{y}$ as follows:

Corresponding Author: Mu Xiangyang, School of Information and Communication Engineering, Xi'an Jiaotong University, China

$$\tilde{x}_i = (t_i + \varepsilon, x_i)^T = \left(t_i + \varepsilon, x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(d)}\right)^T, i = 1, 2, \cdots, N \quad (3)$$

$$\tilde{y}_i = (t_i - \varepsilon, x_i)^T = \left(t_i - \varepsilon, x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(d)}\right)^T, \ i = 1, 2, \cdots, N \quad (4)$$

The above-defined artificial classification problem could be solved by any binary classifier. In this study, we focus on using MPM for classification (MPMC) as the underlying classifier for the problem defined by Eq. 3 and 4. Assume the boundary obtained by the MPMC is

$$a^T u = b \quad (5)$$

Where:

$$u = (f(x), x)^T = \left(f(x), x^{(1)}, x^{(2)}, \cdots, x^{(d)}\right)^T \quad (6)$$

The parameters $a = \left(a^{(1)}, a^{(2)}, \cdots, a^{(d)}, a^{(d+1)}\right)^T$ and b in Eq. 5 can be determined by following constrained optimization problem

$$\min_a \left(\left\|R_{\tilde{x}\tilde{x}}^{1/2} a\right\|_2 + \left\|R_{\tilde{y}\tilde{y}}^{1/2} a\right\|_2\right) \quad (7)$$

s.t. $a^T(\mu_{\tilde{x}} - \mu_{\tilde{y}}) = 1$

where, $\mu_{\tilde{x}}$, $\mu_{\tilde{y}}$, $R_{\tilde{x}\tilde{x}}$ and $R_{\tilde{y}\tilde{y}}$ satisfy $\tilde{x} \sim (\mu_{\tilde{x}}, R_{\tilde{x}\tilde{x}})$ and $\tilde{y} \sim (\mu_{\tilde{y}}, R_{\tilde{y}\tilde{y}})$.

The boundary Eq. 5 obtained by the MPMC turns directly into the regression or prediction function one wants to estimate. That is to say, once the parameters a and b have been determined, we can use the classification boundary to predict the output f (x) for a new input x. When substituting expression Eq. 6 in Eq. 5, we obtain

$$a^{(1)} f(x) + a^{(2)} x^{(1)} + a^{(3)} x^{(2)} + \cdots + a^{(d+1)} x^{(d)} = b \quad (8)$$

The Eq. 8 can be reformulated as

$$f(x) = -a^{-(1)}(a^{(2)} x^{(1)} - a^{(3)} x^{(2)} - \cdots - a^{(d+1)} x^{(d)} + b) \quad (9)$$

Compared the Eq. 9 to 2, it is derived

$$w^{(i)} = -a^{(i+1)} / a^{(i)}, \quad i = 1, 2, \cdots, d \quad b_r = b / a^{(1)}$$

The kernelized version of MPMR: By mapping the training samples into a high-dimensional feature space F we can got a kerneized version of the optimization problem Eq. 7. The training samples $\{\tilde{x}_i\}_{i=1}^N$ and $\{\tilde{y}_i\}_{i=1}^N$ for the class $\tilde{x}$ and class $\tilde{y}$ are mapped as $\tilde{x}_i \to \varphi(\tilde{x}_i)$ and $\tilde{y}_i \to \varphi(\tilde{y}_i)$, where, φ is a mapping function. In the feature space, the optimization problem Eq. 7 is reformulated as:

$$\min_a \left(\left\|R_{\varphi(\tilde{x})\varphi(\tilde{x})}^{1/2} a\right\|_2 + \left\|R_{\varphi(\tilde{y})\varphi(\tilde{y})}^{1/2} a\right\|_2\right) \quad (10)$$

s.t. $a^T(\mu_{\varphi(\tilde{x})} - \mu_{\varphi(\tilde{y})}) = 1$

Through defining $a = \sum_{i=1}^N v_i \varphi(\tilde{x}_i) + \sum_{i=1}^N \beta_i \varphi(\tilde{y}_i)$ and substituting it into Eq. 10, we see that both the objective and the constraints of Eq. 10 can be written as

$$\min_\gamma \left(\left\|\frac{\tilde{K}_{\tilde{x}}}{\sqrt{N}}\gamma\right\|_2 + \left\|\frac{\tilde{K}_{\tilde{y}}}{\sqrt{N}}\gamma\right\|_2\right) \quad (11)$$

s.t. $a^T(\tilde{k}_{\tilde{x}} - \tilde{k}_{\tilde{y}}) = 1$

Where:

$$\gamma = (v_1, v_2, \cdots v_N, \beta_1, \beta_2, \cdots \beta_N)^T, \ \left[\tilde{k}_{\tilde{x}}\right]_i = \frac{1}{N}\sum_{j=1}^N K(x_j, u_i)$$

$$\left[\tilde{k}_{\tilde{y}}\right]_i = \frac{1}{N}\sum_{j=1}^N K(y_j, u_i),$$

$u_i = x_i$ for i =1,2,...,N and

$u_i = y_{i-N}$ for i =N+1,... ,2N

The matrixes $\tilde{K}_{\tilde{x}}$ and $\tilde{K}_{\tilde{y}}$ are defined as:

$$\tilde{K}_{\tilde{x}} = K_{\tilde{x}} - 1_N \tilde{k}_{\tilde{x}}^T, \ \tilde{K}_{\tilde{y}} = K_{\tilde{y}} - 1_N \tilde{k}_{\tilde{y}}^T$$

where, 1_N is a column vector with ones of dimension N. $K_{\tilde{x}}$ and $K_{\tilde{y}}$ contain respectively the first N rows and the last N rows of the kernel matrix (or Gram matrix) K. The matrix k is given by a pre-defined kernel function

$$K(u_n, u) = \varphi(u_n) \cdot \varphi(u) \quad (12)$$

THE KERNELIZED BASED FUNCTION WITH MINKOVSKY'S NORM FOR MPM

The important roles of the kernel functions for MPM: The kernel function K defined in Eq. 12 must satisfy Mercer's conditions, and the mapping function φ from the input space to high dimensional feature spaces. Therefore, the kernel's choice is closely related to the choice of the mapping function φ. The theoretical relationship between K and φ can be analyzed based on the theory of Reproducing Kernel Hilbert Space (RKHS).

The famous Moore-Aronszajn theorem states that for every reproducing kernel, there exists a unique RKHS and H vice versa (Aronszajn, 1950).

Assume the function $f \in H$ has the form

$$f(x) = \sum \mu_n \varphi_n(x) \quad (13)$$

where, $\{\varphi_n(x)\}$ is a set of linearly independent basis functions. It can be proved that the $\varphi(x)$ is the orthonormal eigenfunctions of the integral equation.

$$\int K(x,x')\varphi(x)dx = \mu\varphi(x') \tag{14}$$

The eigenvalues μ_n and the $\{\varphi_n(x)\}$ can define the kernel function.

$$K(x,x') = \sum \mu_n \varphi_n(x)\varphi_n(x') \tag{15}$$

According to the Eq. 15, using different kernels correspond to different mapping function to be φ chosen. This means that different kernels are used in MPM correspond to different representation of the samples in the feature spaces. Therefore, choosing suitable kernel is very important to the MPM.

The Gaussian kernel with Minkovsky's norm for MPM: The Gaussian kernel is one of the most frequently used kernels in practice. Its expression is

$$K(x,x') = \exp\left(-\|x - x'\|_2^2/\sigma^2\right) = \exp\left(-D_{Euc}^2(x,x')/\sigma^2\right) \tag{16}$$

where, $D_{Euc}(x,x')$ is the Euclidean distance between x and x′. Its expression is

$$D_{Euc}(x,x') = \|x - x'\|_2 = \left(\sum_{i=1}^{d}|x_i - x_i'|^2\right)^{\frac{1}{2}} \tag{17}$$

where, $x = \left(x^{(1)}, x^{(2)}, \cdots, x^{(d)}\right)^T$. The Euclidean distance has a natural generalization in form of the Minkovsky's distance

$$D_{Min}(x,x',\alpha) = \|x - x'\|_M = \left(\sum_{i=1}^{d}|x_i - x_i'|^{\alpha}\right)^{\frac{1}{\alpha}} \tag{18}$$

Euclidean and Manhattan distances are special cases of Minkovsky's distance with α = 2 and α = 1, respectively.

For the Eq. 16, the Gaussian kernels with different Minkovsky's distance can be obtained when replacing Eq. 17 with Eq. 18.

RESULTS AND DISCUSSION

To evaluate the performance of MPM using Gaussian kernels with different Minkovsky's distance, we did simulations on Ethernet network traffic prediction problems. Network traffic prediction is of significant interest in many domains, including congestion control, admission control and network bandwidth allocation (Chen *et al*., 2000; Liu *et al*., 2005). For instance, in high-speed network such as Asynchronous Transfer Mode (ATM), network traffic prediction is an essential step in building effective preventive congestion control schemes. The network traffic data can be seen as a time series s (n) varied with the time. We could predict the current traffic level by constructing a prediction model which takes into account the d observations in the past.

Before prediction, the traffic data are normalized to the interval [0, 1]. Mean Square Error (MSE) and prediction error (PE) are used as prediction accuracy measures. Their definitions are as following

$$MSE = \frac{1}{NT}\sum_{n=1}^{NT}\left[s(n) - f(x_n)\right]^2$$

$$PE(n) = s(n) - f(x_n)$$

where, N T is the number of test samples, s (n) is the actual network traffic series and $f(x_n)$ is the prediction function.

In the experiment, we use the Ethernet network traffic as real traffic series for prediction. Which is collected at Bellcore Morristown research and engineering center. And they are aggregated at different time scales of 1 and 5 sec. In this study, we only consider the time scales of 1 sec. First we construct 250 samples for training and 180 for testing. The kernel functions for MPM are Gaussian kernel with different Minkovsky's distance. The kernel width σ is determined by a simple 5-fold cross-validation technique. For convenience, the parameter ε is fixed at ε = 0.4. The comparisons for varying α in the Minkovsky's distance for kernel evaluation are taken for different values of constant d.

To analyze the effect of the parameter d at value of 3 and 5 on the predictive performance, we change the value of α from 1 to 5. The predictive performance is quantified by the MSE and is shown in Table 1.

As shown in Table 1, for each value of constant d, the worst prediction accuracy is obtained by the Gaussian kernels with Euclidean distance (α = 2). While the kernels with Minkovsky's distance given by α = 1 and α =5 present the best results with respect to MSE.

The Table 1 also show that the parameter d has a significant effect on the prediction performance. In addition, we notice that an increase in prediction accuracy is achieved for higher value d. Figure 1 showes the actual

Table 1: The prediction performance of MPM on Ethernet with different α and d

	MSE ($\times 10^{-3}$)	
Minkovsky's distance (α)	d = 3	d = 5
1	0.005	0.000
2	9.157	3.121
3	1.616	0.026
4	0.689	0.584
5	0.489	0.000

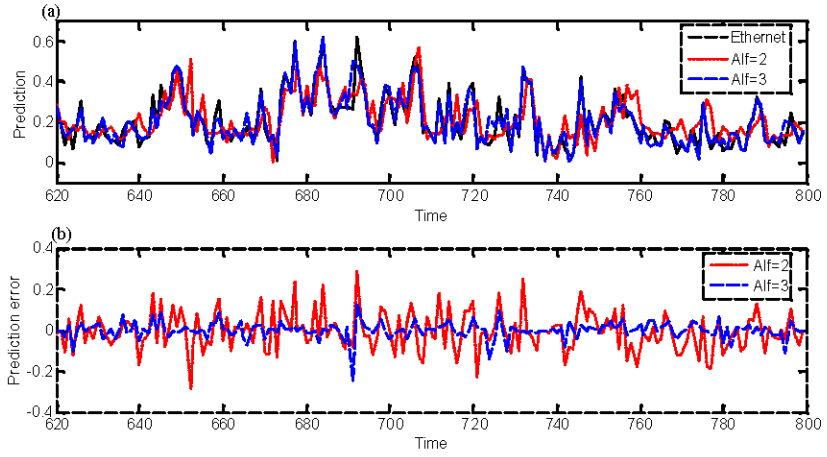

Fig. 1: The prediction performance of MPM on Ethernet traffic for $\alpha = 2$ and $\alpha = 3$. (a) Actual traffic (solid line), the prediction using Gaussian kernel with Minkovsky's distance $\alpha = 2$ (dotted line) and $\alpha = 3$ (dashed line). (b) The prediction error curves of two kernels

and predicted network traffics after MPM application using the Gaussian kernel with Minkovsky's distance $\alpha = 2$ and $\alpha = 3$.

Based on above facts, it is found that the MPM using Gaussian kernels with Minkovsky's distance can present better prediction accuracy than the Euclidean distance.

CONCLUSION

An empirical study for Gaussian kernel with Minkovsky's distance for the Ethernet traffic prediction has been investigated. The results shown that the kernels with Minkovsky's distance given by $\alpha = 1$ and $\alpha = 5$ present the best results with respect to MSE for the Ethernet traffic prediction. Therefore, the MPM using Gaussian kernels with Minkovsky's distance can achieve better prediction accuracy than the Euclidean distance.

ACKNOWLEDGMENT

This research was partially supported by the research of XSTB, China under grant No. CXY08012-1.

REFERENCES

Aronszajn, N., 1950. Theory of reproducing kernels. Trans. Am. Math. Soc., 68: 337-404.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov., 2: 121-167.

Chen, B.S., S.C. Peng and K.C. Wang, 2000. Traffic modeling, prediction and congestion contol for high-speed networks: A fuzzy AR approach. IEEE Trans. Fuzzy Syst., 8: 491-508.

Huang, K., Y. Haiqin, K. Irwin, R.L. Michael and C. Laiwan, 2004. Biased minimax probability machine for medical diagnosis. Proceedings of the 8th International Symposium on Artificial Intelligence and Mathematics, January 4-6, Fort Lauderdale, Florida, USA. pp: 1-8.

Lanckriet, G.R.G., L.E. Ghaoui, C. Bhattacharyya and M.I. Jordan, 2002a. Minimax Probablity Machines. In: Advances in Neural Information Processing Systems, Dietterich, T.G., S. Becker and Z. Ghahramani (Eds.). MIT Press, Cambridge, MA., ISBN-10:0-262-04208-8.

Lanckriet, G.R.G., L.E. Ghaoui, C. Bhattacharyya and M.I. Jordan, 2002b. A robust minimax approach to classification. J. Mach. Learn. Res., 3: 555-582.

Liu, Z.X., D.Y. Zhang and H.C. Liao, 2005. Multi-scale combination prediction model with least square support vector machine for network traffic. LNCS. Adv. Neural Networks, 3498: 385-390.

Perez-Cruz, F. and O. Bousquet, 2004. Kernel methods and their potential use in signal processing. IEEE Signal Process. Maga., 21: 57-65.

Ribeiro, B., 2002. Kernelized based functions with Minkovsky's norm for SVM regression. Proc. Int. Joint Conf. Neural Networks, 3: 2198-2203.