An Unbalanced Penalty SVM for Fault Identification of BOSS

Chen Zhi-feng[1, a], Peng Min-jing[2, b], Li Bo[2, c]

[1]Business Support Center, Jiangmen Branch, China Mobile Group Guangdong Co., Ltd., 529020, China

[2]School of Economics and Management, Wuyi University, Jiangmen, 529020, China

[a]email: 13709612398@139.com, [b]email:15819748999@139.com, [c]email:li.b@joybirds.cn

Keywords: Fault Identification; Support Vector Machine; BOSS; Penalty Coefficient; Unbalanced Samples

Abstract. In order to solve classification error problems of support vector machine, which was used in the telecommunication business supporting system (BOSS), caused by the unbalanced ratio of positive samples, which stand for proper states of BOSS, and negative samples, which stand for the improper states, an unbalanced penalty SVM algorithm was proposed. In the proposed algorithm, values of the penalties were inverse to the ratio of the numbers of positive and negative samples, which means that the number of samples is higher, the lower the penalty coefficient. At last, in order to prove the effectiveness of the proposed algorithm, an experiment was conducted on the classification of running data of BOSS. The result of the experiment proved that the proposed SVM algorithm greatly reduces the negative sample misclassification when the ratio of positive and negative samples was not balanced, which proved the validity of the proposed algorithm.

Introduction

As telecommunication BOSS produce amounts of data daily. Among the data, the size of negative outliner data is very small, but this small amount of data has a very important role[1]. It is the basis for identifying the faults, which may lead the system to failure. If we could not identify and deal with it, they may cause telecommunications systems working improperly, resulting in significant losses. In such a system, the size of the normal data, which is called as positive samples, is much greater than that of failure data, which is called as negative samples. How to identify failure data is the key issue to ensure the normal operation of the telecommunications system.

In the conventional SVM model, misclassified rate is included in the optimization target. However, since the telecommunication BOSS produces amounts of data and the ratio of positive sample data (normal data) to a negative sample data (outliers) is very large, typically they are in 10,000:1 level [2]. In the current SVM, the penalty rates are the same to different misclassifications. When the sample sizes of different classes were different greatly, the current split lines would be unreasonable. Therefore, in this research, an unbalanced penalty SVM algorithm was proposed. In the proposed algorithm, penalty coefficients are determined according to the samples sizes of different classes. At last, an experiment would be conducted to verify that the algorithm is reasonable for identifying the fault working data.

Support Vector Machine

Support Vector Machine (SVM) was first proposed by Vapnik in 1995 [3]. It is a major achievement in machine learning research in recent years. And a lot of research was conducted on improving and applying SVM.

Linear Model: the classification surface equation of SVM is given in equation (1):

$$\boldsymbol{W} \cdot \boldsymbol{X} + \boldsymbol{b} = \boldsymbol{0} \tag{1}$$

Normalizing the discriminant function makes all types of samples meet with $\boldsymbol{g(X) \geq 1}$. To achieve this inequality, $\boldsymbol{W}$ and $\boldsymbol{b}$ can be simply proportional adjusted. Sample points meeting with $\boldsymbol{g(X) = 1}$ have minimal distances from the dividing line or plane. And they determine the optimal

dividing line or plane. So they are called as Support Vector (SV) [4].

Obviously, the problem of seeking the optimal hyperplane could be transformed into the optimization problem in equation (2).

$$\begin{aligned} &\min\ \varphi(W)=\frac{1}{2}W^2=\frac{1}{2}(W\cdot W) \\ &\text{s.t.}\ \ y_i(WX_i+b)-1\geq 0,\ i=1,2,...,n \end{aligned} \tag{2}$$

And the optimization problem could be transformed into the dual optimization problem in equation (3).

$$\begin{aligned} &\min\ Q(\alpha)=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(X_i\cdot X_j)-\sum_{i=1}^{n}\alpha_i \\ &\text{s.t.}\ \begin{cases}\alpha_i\geq 0 & i=1,2,...,n\\ \sum_{i=1}^{n}y_i\alpha_i=0\end{cases} \end{aligned} \tag{3}$$

Introducing nuclear methods: mapping the data in a low dimensional data space into a high dimensional feature space, the classification problem is transformed into the feature space. Vector dot product operations in feature space are corresponding with kernel functions in data space. Through introducing $(x\cdot y)\rightarrow K(x,y)$, equation (3) could be transformed into equation (4).

$$\min\ Q(\alpha)=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(X_i,X_j)-\sum_{i=1}^{n}\alpha_i \tag{4}$$

The optimization problem in equation (4) could be solved by using tools in mathematical optimization libraries.

Non-linear model: For non-linear problems in the feature space, slack variables are introduced, and the optimization problem is given in equation (5) [5]:

$$\begin{aligned} &\min\ \varphi(W)=\frac{1}{2}(W\cdot W)+C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.}\ \begin{cases}y_i(W\cdot\Phi(X_i)+b)-1+\xi_i\geq 0\\ \xi_i\geq 0\end{cases}\quad i=1,2,...,n \end{aligned} \tag{5}$$

Lagrange function is defined in equation (6).

$$\begin{aligned} &L(W,b,\xi,\alpha,r)=\frac{1}{2}W^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i(y_i(W\Phi(X_i)+b)-1+\xi_i)-\sum_{i=1}^{n}r_i\xi_i \\ &\text{s.t.}\ \ \alpha_i\geq 0\quad r_i\geq 0 \end{aligned} \tag{6}$$

Get the minimization of L on W, b and ξ respectively by solving $\partial L/\partial W=0$, $\partial L/\partial b=0$ and $\partial L/\partial\xi=0$, equations (7)-(9) could be obtained.

$$\sum_{i,j=1}^{n}\alpha_i y_i=0 \tag{7}$$

$$W=\sum_{i=1}^{n}\alpha_i y_i\Phi(X_i) \tag{8}$$

$$C-\alpha_i-r_i=0 \tag{9}$$

Replacing variables in equation (6) with equations (7)-(9), we obtain the dual in equation (10).

$$\begin{aligned} &\min\ Q(\alpha)=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(X_i,X_j)-\sum_{i=1}^{n}\alpha_i \\ &\text{s.t.}\ \begin{cases}0\leq\alpha_i\leq C & i=1,2,...,n\\ \sum_{i=1}^{n}y_i\alpha_i=0\end{cases} \end{aligned} \tag{10}$$

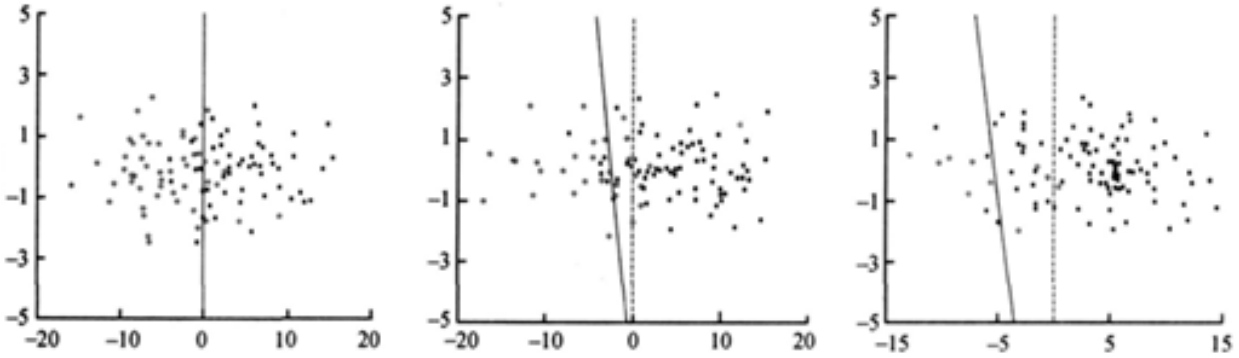
Unbalanced penalty SVM: Considering $(X\cdot x_i)$ is a kernel function, in order to facilitate the discussion, let in that function as the kernel function of SVM. Equation (11) gives a

two-dimensional normal distribution.

$$p(x,y)=(1/2\pi\sigma_x\sigma_y)\exp\left|-1/2\left|(x-u_x)^2/\sigma_x+(y-u_y)^2/\sigma_y\right|\right| \tag{11}$$

The parameters of the two types of samples are: (1) $\sigma_x=\sigma_y=3$, $u_x=3$, $u_y=0$; (2) $\sigma_x=\sigma_y=3$, $u_x=-3$, $u_y=0$.

Three cases of different proportion samples are 50:50, 70:30 and 90:10. Two types of samples are hollow and solid dots in figure 1 [6]. Three dividing lines are obtained and shown in figure 1.



(a) 50:50 (b) 70:30 (c) 90:10

Figure 1. Different proportion samples and dividing lines

It is obvious that the dividing line is biased toward the sample set with less proportion samples, because it can reduce the number of misclassified samples when the penalty coefficient are the same to the two sample sets.

To solve this problem, we set different values for two different misclassification penalties. Let the corresponding numbers of two sample sets are N_1 and N_2, respectively. N_1 is much larger than N_2. And C_1 and C_2 satisfy the formula in equation (12).

$$C_1/C_2=N_2/N_1 \tag{12}$$

And the constraints are shown in equation (13).

$$\text{s.t.}\ \begin{cases}0\le\alpha_i\le C_i & i=1,2\\ \sum_{i=1}^{2}y_i\alpha_i=0\end{cases} \tag{13}$$

Experiments

Background: Compared to the old BOSS, an advantage of Next Generation BOSS (NGBOSS) is the convergent billing. The idea is to achieve convergent billing for all users through unifying ABM (Balance Management Center), CBE (Billing Engine), AUC (Authentication Center), HSC (Information Management Center), BILL (Consolidated Accounts) database. Users order and pay for products and enjoy through CRM (Customer Relation Management). And all the businesses are processed through interface service layer and synchronized to HSC. Information in HSC is the basis of all modules of convergent billing.

The problem is that the synchronization fault can't be avoided because of disaster recovery needs, network latency and hardware stability issues.

Setup: In this experiment, the fault recognition of opening GSM 20 yuan package is taken as an example. In order to identify and locate the fault, the field "subsid" was tracked with other nine fields, which were able to affect the execution of ordering. The nine fields were "opertype" (operation type), "ordertype" (order type), "orgid" (acceptance channels), "operid" (accept operator), "recdate" (hours), "status" (order status), "accesstype" (access type), "isnotify" (SMS notification) and "verfitytype" (authentication mode).

Results : 3,336 cases of 5,336 samples were used as the training samples, and the other 2,000 cases samples were used for testing samples. In the training samples, 17 samples were negative,

3,319 cases were positive; in testing samples, 11 were of negative, and 1,989 were positive. The experiments were conducted with the SVM tool of MatLab 2011, and some modifications were made to the tool for modifying the penalty coefficients.

The standard SVM and the proposed SVM with unbalanced penalties were used to identify the fault. The results are shown in table 1.

Table 1. Classification results

SVM	Standard		Unbalanced penalty	
	Positive samples	Negative samples	Positive samples	Negative samples
The number of misclassified	6	3	8	1
The rate of misclassified	0.30	27.27	0.40	9.09

With respect to standard SVM, unbalanced penalty SVM reduced 66.67% misclassification rate for negative samples. But the positive sample m rate increased 33.33%.

Though the total error rate did not change, but for the telecommunications system, the key was the misclassification rate of negative samples. In this regard, the proposed SVM has greatly improved the performance of fault identification.

Conclusion

In the massive data produced by telecommunication systems, fault identification on negative samples is critical tasks. In this type of data, the proportion of negative samples is very low, typically is less than 5%, which is extremely unfavorable for the SVM with the same penalty coefficients to positive and negative samples.

The significance of this research work is that an unbalanced penalty SVM algorithm is proposed. And the values of penalties are determined according to sizes of both positive and negative samples. Experimental results show that the proposed SVM model for the identification of negative samples is effective.

Acknowledgement

This research is supported by Natural Science Foundation of Guangdong Province, China (10452902001004947).

References

[1] Joseph O. Sexton, Dean L. Urban, Michael J. Donohue, and Conghe Song. Long-term land cover dynamics by multi-temporal classification across the Landsat-5 record[J]. Remote Sensing of Environment, 2013, 128(3): 246-258.

[2] Zhen Zhen, Lindi J. Quackenbush, Stephen V. Stehman, and Lianjun Zhang. Impact of training and validation sample selection on classification accuracy and accuracy assessment when using reference polygons in object-based classification[J]. International Journal of Remote Sensing, 2013, 34(19): 6914-6930.

[3] Ching-Pei Lee, Ching-Pei Lee. A Study on L2-Loss (Squared Hinge-Loss) Multiclass SVM[J]. Neural Computation, 2013, 25(5): 1302-1323.

[4] V. Srinivasan, G. Rajenderan, J. Vandar Kuzhali, and M. Aruna. Fuzzy fast classification algorithm with hybrid of ID3 and SVM[J]. Journal of Intelligent and Fuzzy Systems, 2013, 24(3): 1064-1246.

[5] Abdulhamit Subasi. Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders[J]. Computers in Biology and Medicine, 2013, 43(5): 576–586.

[6] Peng Min-jing. A Classification SVM Model with Variable Punishment Coefficient Ratio[J]. Journal of Guangxi Normal University: Natural Science Edition, 2008, 26(3): 118-121.