Improved Weighted Support Vector Machine

Li Wanling, Chen Peng, Song Xiangjun

Troops 63908 PLA

Shijiazhuang, China

759934614@qq.com

Keywords: SVM; Weighted SVM; sample; isolated points; C-SVM

Abstract. In this paper, weighted Support Vector Machine was introduced. And the weighted Support Vector Machine was improved. Some methods of determining weight were introduced, and the comprehensive method of determining weight was adopted. Simulation results indicated that the total accuracy rating by improved weighted Support Vector Machine is higher than C-SVM. It's able to improve the distribution accuracy rating with improved Supported Vector Machine effectively.

Weighted Support Vector Machine

In standard Support Vector Machine, it's equal to one in each kinds of sample weight. But some samples is bigger used in determining classified ultra-plane, and some is smaller, or another exceeds some else classified soundly. Therefore, considering conditions of some samples classified, it adjusts weight to Support Vector Machine. Each single sample-point gets different penalized coefficient, to gain much more exact classifications, and lastly it produces Weighted Support Vector Machine, among representing optimized the problems. The initial problem is:

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C\sum_{i=1}^{l} s_i \xi_i \\ s.t. \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1,2,\Lambda\ ,l \end{aligned} \tag{1}$$

Where s_i=weights of all kinds of samples, the dual problem is：

$$\begin{aligned} \min \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ s.t. \quad & 0 < \alpha_i < s_i C,\ \ i = 1,2,\mathrm{L}\ ,l \\ & y^T\alpha = 0 \end{aligned} \tag{2}$$

Against a given problem, there is a decision to the characters of Weighted Support Vector Machine with determined weight.

The methods of determining weight

It's different among the methods determining weight with Weighted Support Vector Machine, directing against different problems. It produces a way to determine the weight on Weighted Support Vector Machine, unbalanced samples, single-point, and importance of different samples.

The method of determining weight for eliminating unbalanced samples.

Let positive weight= s^+, negative weight= s^-, also, s^++s^-=1, then,

$$\frac{s^+}{s^-} = \frac{l^-}{l^+} \tag{3}$$

The method of determining weight for eliminating isolated points.

Clustering type point is a show computing the vicinity between vector x and clustering type on the condition of dense clustering type. There are some typical expresses as follows:

Where mean vector is

$$\mathbf{m}_p = \frac{1}{l}\sum \mathbf{x} \tag{4}$$

Where mean center $m_c \in S$ is defined as

$$\sum_{y\in S} d\left(m_c, y\right) \le \sum_{y\in S} d\left(z, y\right), \forall z \in S \tag{5}$$

Where median center $m_{med} \in S$ is defined as

$$med\left(d\left(m_{med}, y\right) \mid y \in S\right) \le med\left(d\left(z, y\right) \mid y \in S\right), \quad \forall z \in S \tag{6}$$

Where d is an un-similar measurement between two points. While similar measurement is referred, inequality mark is turned. It suggests a method, using weighted SVM based on clustering type point, to solve the effects on the classification by isolated point. First, it computes clustering type point of each kind of training data. Then, it expresses the computed distance between every data in all kings and the clustering type point making use of a kind of measurement, Such as Euclid Distance. Last, on the basis of the distance, the comparable weight is determined using fuzzy distribution function. It's not on the original space to solve optimal ultra-plane in SVM, but to find the solution mapping the sample with function φ(x)into characteristic space of high dimension. Therefore, it must not to calculate the clustering type point and the distance between sample vector and clustering type point in the original place. And it must be done in the characteristic space. This paper confirms the clustering type point and some distances using Euclidean distance. That the expression in the characteristic space is showed as follows:

$$d\left(x, x_i\right) = \left\|\varphi\left(x\right) - \varphi\left(x_i\right)\right\|^2 = K(x,x) - 2K(x,x_i) + K(x_i,x_i) \tag{7}$$

It employs fuzzy distribution function, according to the discussion above to compute each data's weight using fuzzy distribution function.

$$s_i = \begin{cases} \dfrac{1}{1+[a(d-c)]^b}, & d > c \\ 1, & d \le c \end{cases} \tag{8}$$

The method of determining weight based on type importance.

Not only each sample importance of every type is different, during a great many of problems, but also the importance of each type retaliated with others is .For example, if some fault is not removed timely, t must be a damage fault. So the diagnose accuracy rating has to be higher than others. There is an example with fault diagnose to explain weight determine method in view of sample importance. It's supposed that the fault diagnose divided into five grades, then, the fuzzy theory region of language value is:

X= {fatal fault, major fault, common fault, minor fault, normal condition} (9)

Quantification theory region is obtained after being quantified:

x = {1,2,3,4,5}. The subject function of fault sample type is shown in figure 1.

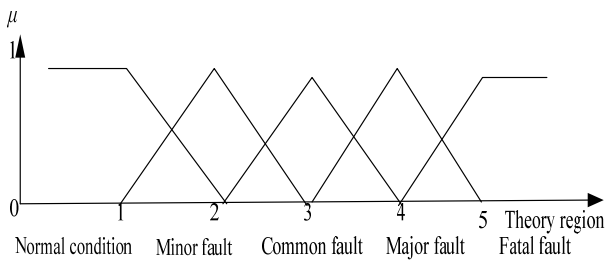

Figure 1 Subjection Function of Faults

It's able to get a subject degree vector $\mu_i, i = 1,2,\cdots,m$,being fuzzed on the basis of professional experience knowledge to fault sample type level for all the fault types, where m is a sample type number. This paper employs single to single SVM for fault diagnose. Suppose i type weight is s_i,$_j$ type weight is s_j, and s_i+s_j=1. Let

$$\frac{s_i}{s_j} = \frac{\|\mu_i^T x\|}{\|\mu_j^T x\|} \tag{10}$$

The comprehensive method of determining weight.

To solve train sample unbalance problem, single point, importance of sample type and esc .at the same time. There must to be given comprehensive consideration for three methods determining weight. Let m type fault is while training the I type and j type fault's SVM_{ij}. Let the I type weight is s_i^+, the j type weight is s_j^-,and $s_i + s_j = 1$,according to the method solving sample unbalance problem. The distance are d_{ik} and d_{jk} separately between sample with I type and j type and clustering type point, by formula (17). k=1,2,…,l, where l= sample length of all kinds. So, the means to determine sample weight as bellow:

If $s_i^+ > s_j^-$ and $s_i > s_j$, then the method of the i type to determine sample weight is ,

$$s_{ik} = \begin{cases} 1 & d_{ik} \le c \\ \dfrac{1-\max(s_i^+, s_i)}{1+[a(d_{ik}-c)]^b} + \max(s_i^+, s_i) & d_{ik} > c \end{cases} \tag{11}$$

Simulation check and result analysis

This paper presents the simulation by osu_svm3.00 SVM simulate tool cabinet, artificial data abided by Guass and breast-cancer and ionosphere data from International General Machine Learn Data Base UCI. The data is divided into training and identifying, and the training data needs to be selected less to embody the advantage of SVM small sample study. Each data length is shown in table 1.And the testing point of Gauss data is a kind of data produced stochastically, together with same distribution and independence for training sample.

Table 1 Attribute of Data

		Training data	Testing data
Guass		5	100
Breast-cancer	Benign	100	350
	Malignant	50	140
Ionosphere	Good	110	110
	Bad	60	60

If RBF kernel function is used, then formula 17 can be melted into:

$$d(x, x_i) = 2(1 - K(x, x_i)) = 2(1 - \exp\{-\gamma \|x - x_i\|^2\})$$

The choice with SVM simulation parameter (C, γ) is indicated in table 2 inside the lower brackets .And the parameters with fuzzy distribution function are a=5,b=3 and c, but c is selected by the distance distribution between two types. Meanwhile, two type data is simulated by BP network. Inside, latent layer nerve element is 20 of BP network. The simulation result is presented in table 2.

Table 2 The Simulation Results

		BP	C-SVM	Improved WSVM
Guass	Class 1	0.91	0.96	0.98
	Class 2	0.84	0.78	0.94
	Total	0.875	0.82	0.96
Breast-Cancer (100, 1)	Malignant	0.9420	0.9450	0.9886
	Benign	0.8898	0.9496	0.9812
	Total	0.9159	0.9473	0.9849
Ionosphere (0.5, 0.25)	Good	0.9582	0.9898	0.9952
	Bad	0.7368	0.9186	0.9168
	Total	0.8475	0.9542	0.9560

Every type contains a single point with Gauss data, WSVM is introduced to eliminate the effects on classification. The training result is indicated in figure 2.

There is a conclusion from the simulation result. The total accuracy rating by improved WSVM is higher than C-SVM. The accuracy rating is reduced in the much more training data, but it's

improved obviously in less. Therefore, improved WSVM is an effective improvement for C-SVM. It's able to improve the distribution accuracy rating with Supported Vector Machine effectively. The simulation result of Gauss data can make out that WSVM based on clustering type point can wake the classification effects by single points.

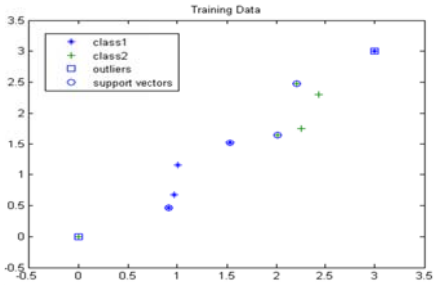

Figure 2 Simulation results of Gauss distribution data

Conclusions

Weighted Support Vector Machine was introduced in this paper. And the method of determining weight for eliminating unbalanced samples, the method of determining weight for eliminating isolated points, the method of determining weight based on type importance, the comprehensive method of determining weight were adopted to training the improved weighted Support Vector Machine. This paper presents the simulation data, and the simulation results indicated that the total accuracy rating by improved weighted Support Vector Machine is higher than C-SVM.

References

[1] Tian Yingjie, Qi Zhiquan. A New Support Vector Machines for Multi-class Classfication [C] // Proceedings of 2nd International Conference on Recent Advances in Space Technologies. 2005: 18-22.

[2] Widodo Achmad, Yang Bo-Suk. Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors [J]. Expert Systems with Applications, 2007, 33 241-250.

[3] C. W. HSU. A Comparison of Methods for Multiclass Support Vector Machines[C]. IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp. 415-425, 2002.

[4] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1999.

[5] Kyung-Soon Lee, Kyo Kageura. Virtual relevant documents in text categorization with support vector machines [J]. Information Processing and Management, 2007, 43: 902-913.