Research on Sample Dataset Balance Method of SVM Based on GA

Xiao Han
Department of Electrical Engineering, North China Electric Power University, Baoding 071003, China

ABSTRACT: SVM was widely used in fault diagnosis, and achieved good results. However, the unbalance between normal sample datasets and fault sample datasets made it very difficult to establish a proper diagnosis model. For actual diagnosis, the normal samples are usually more than the fault ones, and it will lead to misdiagnosis. In this paper, a method based on GA to solve the imbalance problem for SVM is presented. In this method, the samples are expanded by GA so that the number of normal sample datasets and fault sample datasets keeps balance. The method of selecting parent samples is also studied. The experiments show that the method proposed in this paper improves the accuracy of diagnosis.
KEYWORD: SVM, Fault Diagnosis, Sample Balance, GA

1 INTRODUCTION

The misjudgment of the normal state will lead to the unnecessary maintenance or downtime, also the loss of benefits. The misdiagnosis of the fault state will delay the processing failures, and cause the damage of equipment. In the 1990s Vapnik put forward the SVM [1]. However, when the samples numbers of each type are unbalance, hyper plane will move toward the type which has less samples (CL). When diagnosing, the test sample which should belong to CL may be misdiagnosed as the type which more samples (CM). This is called imbalanced sample problem [2].

In the actual state diagnosis, the samples of normal operational state are often more than the fault samples. The diagnosis model, which was established by those samples, that is, the fault condition may be diagnosed as the normal state, especially the "critical fault". At present, there are two methods to solve this problem. One is to improve the algorithms [3-4], adjusting the punish coefficient C so that the support vector (SV) of CL get a greater punish coefficient to get the correct hyper plane. The other one is to reconstruct the samples set [5-6].

Based on the second principle, the method based on Genetic Algorithms (GA)[7] to solve the unbalanced sample problem of SVM was proposed in the paper. In the method, suitable parent samples in CL were selected. And by the use of crossover and mutation, the offspring samples belonging to CL were generated expand CL. Meanwhile, the paper analyzed and discussed how to select the parent sample. The practical application shows that the method could solve the unbalanced sample problem of SVM.

2 SUPPORT VECTOR MACHINE

2.1 *The basic principle*

SVM has good generalization performance on small size sample and nonlinear problems [8]. For the training set: $T=\{(x_1, y_1), (x_2, y_2),\ldots,(x_n, y_n)\}$.

By solving the optimization problem, the decision function can be obtained:

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n} y_i\alpha_i K(x,x_i)+b) \quad (1)$$

$K(x_i, x_j)$ is kernel function; $y_i \in \{1, -1\}$; α_i is the Lagrange multiplier; b is the bias terms. Optimal Lagrange coefficients α_i may be three cases[9]: (1) $\alpha_i=0$: the sample x_i corresponding to α_i is correctly classified, and these samples are not support vector (non-SV); (2) $0<\alpha_i<C$: the sample x_i corresponding to α_i is standard SV; (3) $\alpha_i=C$: the sample x_i corresponding to α_i is boundary SV. Standard SV and boundary SV collectively referred to support vector.

2.2 *Imbalanced samples problem of SVM*

In establishing the diagnosis model, the imbalance of the number of samples will make the hyper plane shifted, leading to misdiagnosis. One reason is the lack of CL information reduces the reliability of the diagnostic model; On the other hand, actually SV is misclassified samples, and the risk of misdiagnosis of each class can be calculated from a sample of each type of risk being misclassified as follows:

© 2016. The authors - Published by Atlantis Press

$$P = \frac{N_{SV}}{N} \tag{2}$$

N_{SV} is the SV number of each class, N is the sample number of each class. So, when the sample number of one class is fewer than the other, the risk of misdiagnosis will rise.

3 RESEARCH ON THE IMBALANCE SAMPLES PROBLEM

To solve the imbalanced samples problem of SVM, a sample balance method based on genetic algorithm were proposed, which can expand the number of CL.

3.1 *Crossover and mutation*

GA is an optimization algorithm which simulates the mating and mutation phenomena in the natural genetic evolution process. It departure from the initial population and generate new individuals which are more adapted to the environment [10]. The solving process of GA includes the following: (1) Coding, using digital representation of each individual, like genetic chain of chromosomes; (2) Generate the initial population; (3) Fitness evaluation, according to certain criteria, evaluate the merits of each individual; (4)Selection: Select the individual which adapt ability is better than others to the next "breeding" process; (5) Cross, the parent mate to exchange information, in accordance with the crossover probability P_c; (6) Mutation, select an individual and change its gene in some position according to the mutation probability P_m. The new individual not only save the fine genes of the parent, but also makes the population diversity.

3.2 *Analyzing and solving of the imbalance samples problem*

According to the constraints in quadratic programming, equation (3) can be got:

$$\sum_{y_i=+1} y_i \alpha_i + \sum_{y_i=-1} y_i \alpha_i = 0 \tag{3}$$

Assuming a perfect diagnostic model, which diagnostic accuracy rate is100%, can be built using the training sample, then the classification hyper plane position will be in a perfect position, and perfect SV of this diagnostic model can also be got.

Compared with the "perfect hyper plane", this position of "hyper plane" is changed when the sample number of one class is fewer than the other. Using this model, the sample belongs to CL will be misdiagnosed as CM. The root of the problem is that some samples of CL, which should not be SV, are treated as SV. Therefore, in order to avoid this situation, the CL sample need to expand, and the expansion sample should be or likely to be SV. It can be found, small sample number means higher probability of misclassification. Therefore, GA is used to expand the original sample set. First, searching for suitable sample as parent samples in CL. Using genetic manipulation to generate offspring samples. And then the number of samples could be expanded, and the samples numbers of the two categories become balanced. Finally, the new samples set are used to train and build SVM diagnosis model.

3.3 *Selecting method of the parents samples*

Before expanding the original sample set, the most important step is to select the most suitable parent samples. In the training process of the SVM diagnosis model, SV had an effect on the establishment of the model, while the other samples had no effect. So the ultimate aim is to expand the SV belonging to CL. In the GA, the offspring maintained the characteristics of the parent sample. So if the parents are SV, the offspring have a high probability to become SV of the new model. On the contrary, if the parents are not the SV, the offspring have a low probability to become SV. Therefore, when using GA to expand the sample, SV should be selected to be parent sample.

Before the establishment of new diagnostic model, its SV can not be found in advance. It can be found in the above analysis, when using the original sample to the establishment SVM diagnosis model, SV can be obtained. Although these SV are not entirely SV of "perfect super flat", but they have a high probability to become or contain SV of "perfect hyper plane." Therefore, the SV in the original sample data set is the most suitable choice as a parent sample, which offspring will most likely to be the SV of new SVM diagnosis model.

3.4 *Evaluation Method of offspring sample*

Some offspring samples will be obtained by crossover and mutation. In these new samples, not every sample can be used to expand the training set. In order to ensure fine features of the new training set, the offspring sample selection must meet two requirements:(1)The new offspring samples must belong to CL; (2)The new offspring samples should be SV. In the paper, Euclidean distance(ED) between the class center and the samples are used to measure the above conditions. First, for an n-dimensional set of samples which contain N samples $\{x_1, x_2, \ldots, x_N\}$, ED between the class center and the samples is:

$$m = \frac{1}{n} \sum_{i=1}^{N} x_i \tag{4}$$

$$d(x_i, m) = \sqrt{\sum_{j=1}^{n} (x_i(j) - m(j))^2} \tag{5}$$

Thus, the class centers m_1、m_2 of CL and CM can be obtained respectively, and also the ED between the new sample and the two class centers $d(x, m_1)$ and $d(x, m_2)$. So the above requirements are converted to:

$$d(x,m_1) < d(x,m_2) \tag{6}$$

$$\min [d(x,m_2)-d(x,m_1)] \tag{7}$$

3.5 *Realization of unbalanced sample diagnostic model*

For unbalanced sample set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, $y_i \in \{1,-1\}$, "1" represents CL and "- 1" on behalf of CM. The difference of two type samples number is l. Solving the quadratic programming problem and the optimal Lagrange coefficients $\{\alpha_1, \alpha_2,\ldots,\alpha_n\}$ corresponding to the samples in the unbalanced sample set can be obtained; Define the crossover operator $\otimes$ and mutation operator $\oplus$, calculated by equation (8-12).

$$Z_1=x_i \otimes x_j \tag{8}$$

$$Z_2=\oplus x_i \tag{9}$$

$$\alpha_i \bullet \alpha_j \neq 0 \tag{10}$$

$$y_i=1, y_j=1 \tag{11}$$

$$Z=[Z_1, Z_2] \tag{12}$$

Calculate the center of the two types m_1 and m_2 by formula (4-5), as well as the distance $d(z_i, m_1)$ and $d(z_i, m_2)$ between z_i and the two types center. Then, calculate t_i by formula (13).

$$t_i=d(z_i, m_1)- d(z_i, m_2) \tag{13}$$

Sort t_i arranged from small to large. Find out l t_i in the front of the t_i sequence, which value is positive. And select z_i correspond to those t_i to form a new offspring samples set Z^*; Put $(z_i^*, 1)$ into the original training set T, which is unbalanced, to form a new training set T^*; Training SVM by T^*, the correct diagnosis model can be established.

4 APPLICATION RESEARCH

SVM diagnosis model is established according to this method based on vibration data of the turbine rotor, and the vibration fault diagnosis was realized. Sample set includes two types of samples: normal samples and rubbing fault samples. Where normal samples for CM, rubbing fault samples for CL, as shown in Table 1. First, extracting the characteristics of those 20 training samples. In this paper, Fourier transform was used to extract the feature. 0.5*f*, 1*f*, 2*f*, 3*f*, 4*f* were selected as sample characteristics (*f* is the frequency).

Table1 samples set

	total number	Training samples	Test samples
normal samples	17	12	5
rubbing samples	13	8	5
Total	30	20	10

Then, establish the original SVM diagnosis model (model I), which model parameters C = 200; after the initial establishment, 12 SV belonging to the two types were obtained. The SV belonging to rubbing fault samples were selected as a parent sample, and were encoded. The SV belonging to rubbing fault samples were selected as a parent sample and were encoded. After getting their genetic chain, crossover and mutation operations were used to generate offspring samples; Select 4 individuals from the offspring samples, as shown in Table 2. And added the 4 individuals to the original training set to form a new training set.

Table2 offspring samples

	0.5*f*	1*f*	2*f*	3*f*	4*f*
CL21	0	0.7774	0.1225	0.04	0.0369
CL22	0.0001	0.7799	0.1625	0.017	0.0599
CL23	0.0056	0.7779	0.0811	0.0874	0.0285
CL24	0.0025	0.6311	0.2057	0.0771	0.0618

The new sample set of were used to train and establish SVM diagnosis model (model II). And 5 normal test samples and 5 rubbing fault test samples were diagnosed by model II. Meanwhile, the non-SV sample of rubbing fault samples were selected as the parent samples to generate new training samples, which were used to establish SVM model (model III). And then diagnose the same test samples by model III. Diagnosis results are shown in Table 3.

Table3 Comparison of diagnosis results

code	Type	model I		model III		model II	
		result	T/F	result	T/F	result	T/F
1	normal	normal	T	normal	T	normal	T
2	normal	normal	T	normal	T	normal	T
3	normal	normal	T	normal	T	normal	T
4	normal	normal	T	normal	T	normal	T
5	normal	normal	T	normal	T	normal	T
6	rubbing	rubbing	T	rubbing	T	rubbing	T
7	rubbing	rubbing	T	rubbing	T	rubbing	T
8	rubbing	normal	F	normal	F	rubbing	T
9	rubbing	rubbing	T	rubbing	T	rubbing	T
10	rubbing	normal	F	normal	F	rubbing	T
Accuracy		80%		80%		100%	

With respect to the model I, the accuracy of model II was raised to 100%. Unbalanced sample problem was solved reasonably. The accuracy of model III was not changed. And the misdiagnosed samples were same to mode I. This indicates that only the offspring of SV have effect on the model building, while the offspring of non-SV almost have no effect.

5 SUMMARY

This paper put forward to solve the unbalanced sample problem of SVM and the method based on GA was proposed. Selection principle of the parent sample was studied. The practical application shows that the method can establish the correct diagnosis model, greatly improving the diagnosis accuracy. The method is quite simple and intuitive. The following concludes can be obtained: (1)Using genetic algorithm to generate the offspring sample not only saved the good parent gene of generation, but also causes the population maintains diversity, such characteristics are suitable to solve the problem of unbalanced samples; (2)Sample expansion is expansion of SV, expansion of non SV samples has almost no effect on model; (3)The offspring of SV have greater probability to become the SV, while the offspring of non SV have smaller probability, so SV should be selected to be parent samples.

REFERENCES

[1] Vapnik V. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[2] Huang Xiu-li, Wang Wei. Application of SVM in imbalanced dataset[J]. Computer Technology and Development, 2009, 19(6): 190-193.

[3] Wang Juanjuan, Ren Qiushi. SVM algorithm with different error costs based on SMO[J]. Information Technology, 2006, 30(10): 45-46, 128.

[4] Liu Wan-Li, Liu San-Yang, Xue Zhen-Xia. Balance method for imbalanced support vector machines[J]. Pattern Recognition and Artificial Intelligence, 2008, 21(2): 136-141.

[5] Rehan Akbani, Stephen Kwek, Nathalie Japkowicz. Applying support vector machines to imbalanced datasets[C]. Spring- verlag Berlin Heidelberg, 2004.

[6] Yang Zhiming, PengYu. Research on classification technique for imbalanced dataset based on support vector machines[J]. Chinese Journal of Scientific Instrument, 2009, 30(5): 1094-1099.

[7] Zhao Kun, Geng Guangfei. Reactive power optimization of distribution network based on improved genetic algorithm [J]. Power System Protection and Control, 2011, 39(5): 57-62, 68.

[8] Cheng Junsheng, Yu Dejie, Yang Yu. Fault diagnosis of roller bearings based on EMD and SVM[J]. Journal of Aerospace Power, 2000, 14(4): 523-543.

[9] Fang Ruiming. Theory and application analysis of support vector machine[M]. Beijing: China Electric Power Press, 2007: 4-10.

[10] Yang Shuying. Pattern recognition and intelligent computing [M]. Beijing: Electronics Industry Press, 2008: 298-318.