International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.3 (2016), pp.273-282
http://dx.doi.org/10.14257/ijmue.2016.11.3.26

Prediction of Damage to Insulation Joints Based on SVM with Unbalanced Data Sets

Dong Yu[1] and Xiao Zi-Qiang[1,2]

School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
ziqiangx@yahoo.com

Abstract

As a key part of track circuit, the state of insulation joints is related to safe, normal and efficient operation of railway. In order to accurately obtain different degrees of insulation joints, a prediction model based on support vector machines has been proposed to study damage to insulation joints. For unbalanced data sets in the research process, a KNN under-sampling is presented to remove redundant and noise samples. By means of BSMOTE over-sampling method to further take full advantage of the data, KNN-BSMOTE-SVM algorithm of hybrid sampling is given to achieve balanced data sets. The theoretical analysis and simulation results show that the proposed algorithm increases classification performance of SVM classifier. Compared with KNN classifier, the classification results of SVM are better, support vector machines used in insulation damaged joints prediction is feasible and effective.

Keywords*: unbalanced data sets; support vector machines (SVM); insulation joints; track circuit; prediction*

1. Introduction

Insulation joints are an integral part of railway network infrastructure, which guarantee the normal operation of railway line and ensure traffic safety. In the electrical service system, damaged insulation joints are the main cause of track circuit failure. Once rail insulation joints are damaged by squeezing, trains may be obstructed, which directly affects the normal order of rail transport, causing unnecessary economic losses. The traditional methods for determining the damaged insulation joints (microcomputer monitoring system and Megger) are slower, poor real-time, hysteretic nature and strong subjective. Meanwhile, the judge results only have two states, "normal" and "abnormal", not a good description of the extents for damaged insulation joints. At this stage, the study of insulation joints rather focus on polar transposition and damaged reasons [1] than prediction of damage to insulation joints.

SVM (support vector machines) is a statistical theory of machine learning method, which is based on structural risk minimization criterion and VC dimension concept, has been very successfully applied to prediction field [2-3], fault diagnosis [4], image recognition and intrusion detection. The basic ideas [5]: To find a hyperplane to correctly classify two types of sample points, and achieve the maximum class interval. In recent years, insulation joints prediction has not been reported in domestic and overseas.

SVM is a supervised classification algorithm, training balanced data set samples could achieve good generalization ability. Therefore, the study on the training data set of SVM, often assume that all kinds of the number of samples is basically equal, is balanced. But in most cases, lots of data samples are unbalanced, the SVM is used to directly deal with the issue of unbalanced data sets, generalization ability is not high. In practical application, the number of positive sample is often difficult to obtain, but it is the essence of these data to better reflect the problem very well. All along, many scholars are concerned about how

ISSN: 1975-0080 IJMUE

to improve the classification performance of SVM algorithm with unbalanced data, making it better. The [6] puts forward a cost sensitive SVM algorithm based on hybrid sampling. And [7] proposes a new fault detection algorithm based on border synthetic minority over-sampling technique combined with cost sensitive SVM. Although to some extent, under-sampling cost sensitive SVM algorithm has been improved the classification performance of SVM, it is limited to a subset of the negative sample, not up to the ideal especially when serious unbalanced data.

Therefore, in order to improve the ability of anti-jamming of noise sample, K nearest neighbor under-sampling method is put forward to construct SVM algorithm. Meanwhile, to further strengthen the boundary samples, a KNN-BSMOTE-SVM algorithm, combine with BSMOTE algorithm, is proposed to achieve balanced data. In the course of the experiment, the algorithm is applied to predict damage to insulation joints, the results show that the algorithm has better classification performance with unbalanced data, compared with other algorithms.

2. Theoretical Background

In this paper, multi-class SVM is devised to predict damage to insulation joints. To solve this problem, we suppose that there are training data ($\boldsymbol{x}_i$, $\boldsymbol{y}_i$), (i=1,2,…,N), $\boldsymbol{x}_i \in \mathrm{R}^d$, $\boldsymbol{y}_i \in \{+1, -1\}$, where each data is a d-dimensional vector, $\boldsymbol{y}$ is the category label, $\boldsymbol{w}_1$ class marks +1, and -1 belongs to $\boldsymbol{w}_2$ class. Samples given are linearly separable, that is, the hyperplane: $g(\boldsymbol{x})=(\boldsymbol{w}\cdot\boldsymbol{x})+ b=0$, which presents to separate all N samples with no error, where $\boldsymbol{w}\in \mathrm{R}^d$ means weight of linear discriminant function, b is the constant term, $(\boldsymbol{w}\cdot\boldsymbol{x})$ represents the inner product of vector $\boldsymbol{w}$ and $\boldsymbol{x}$. To N samples could be correctly classified and hyperplane has classification interval, requirements are proposed as $\boldsymbol{y}_i[(\boldsymbol{w}\cdot\boldsymbol{x})+b]\geq 1$, ($i$=1,2,…,$N$), and the class interval is maximum, that is M=2/$\|\boldsymbol{w}\|$, where $\|\boldsymbol{w}\|$ is die of the weight vector, that is, $\|\boldsymbol{w}\|=(\boldsymbol{w}\cdot\boldsymbol{w})^{1/2}$.

Therefore, to solve the optimal hyperplane is to minimize $\|\boldsymbol{w}\|^2$. The problem could be solved by Lagrange method. A Lagrange coefficient $\alpha_i \geq 0$, (i=1,2,…,N) is introduced to each sample, obtaining the optimization problem of equivalent conversion.

$$L(\boldsymbol{w},b,\alpha)=\frac{1}{2}(w\cdot w)-\sum_{i=1}^{N}\alpha_i\{[(w\cdot x_i)+b]-1\} \quad (1)$$

Where $L(\boldsymbol{w},b,\alpha)$ is the Lagrange function, the solution is to find the smallest $\boldsymbol{w}$ and b, but the largest α, the best solution has to be got at the saddle point of $L(\boldsymbol{w},b,\alpha)$.

Then, to seek the partial differentiation of $\boldsymbol{w}$ and b, then to zero them, adjust Eq.(1), getting the dual problem of the original problem.

$$\frac{\partial L(w,b,\alpha)}{\partial w}=0 \Rightarrow w=\sum_{i=1}^{N}\alpha_i y_i x_i \quad (2)$$

$$\frac{\partial L(w,b,\alpha)}{\partial b}=0 \Rightarrow \sum_{i-1}^{N}\alpha_i y_i=0 \quad (3)$$

To finish Eq.(1), Eq.(2) and Eq.(3), Eq.(4) could be obtained.

$$L=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \quad (4)$$

According to Kuhn–Tucker conditions of optimization theory, Lagrange function has to satisfy at the saddle point in Eq.(4).

$$\alpha_i\{y_i[(w\cdot x_i)+b]-1\}=0, (i=1,2,\ldots,N) \quad (5)$$

Conditions could be seen from the above Eq.(5), only when α_i is zero or α_i is non-zero but with condition, $\boldsymbol{y}_i[(\boldsymbol{w}\cdot\boldsymbol{x})+b]=1$, Eq.(5) is set up, which corresponds to samples are support vectors.

For non-linear support vector machine, as long as $K(\boldsymbol{x}_i \cdot \boldsymbol{x}_i)$ meets Mercer condition, according to the theory of functional space, function $K(\boldsymbol{x}_i \cdot \boldsymbol{x}_i)$ is the inner product of swap space. Therefore, the problem to solve non-linear support vector machine by designing nonlinear transformation could be realized by directly designing the kernel function. At the same time, taking the situation into account, some samples could not be classified correctly by the best classified surface, so slack variables ε_i is introduced to optimal hyperplane, that is, $\varepsilon_i \geq 0$, (i=1,2,…,N), then the constraint of optimal hyperplane is as follows:

$$y_i\left[w^T x + b\right] \geq 1 - \varepsilon_i, i = 1, \cdots, N \tag{6}$$

Then, we have got dual form of optimization problem under non-linear classified.

$$w(a) = -\frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{N} \alpha_i \tag{7}$$

With constraints,

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C, i = 1, \cdots, N \tag{8}$$

Finally, SVM decision function is got in the new feature space, namely (N is the number of support vectors):

$$f(x) = sign\{\sum_{i=1}^{N} \alpha_i y_i [\varphi(x_i) \cdot \varphi(x) + b]\} \tag{9}$$

3. Insulation Joints Description

Insulation joints are vital for the safety and normal transportation of railway, with insulation roles, to divide interval separation, and ensure current transported directly along the rail in track circuit. Above all, as an important part of the track circuit, the weakest link of line, the main part of a combination of electrical engineering, sensitive areas of restricting traffic safety, if insulation joints are broken or damaged, forming a red belt, causing a short circuit fault, it has affected the normal railway driving, badly interfered with transportation. Failure rate of poor insulation joints is high and long delay. Damaged insulation joints show the track circuit fault.

To analysis electrical characteristics at both insulation joints ends and influencing factors, we have found that there are differences on the value of voltage current and resistance. Research shows that different sample value of insulation joints presents different status. Therefore, the characteristic components (electrical characteristics and influencing factors) are selected to identify the conditions of insulation joints. Data provided has two parts, some are track circuit test data, others are glued insulation test data. Characteristic components are as shown in Table 1.

Table 1. Characteristic Components

Test data	Characteristic components	Standard value	Consider or not
Track circuit	Weather and ballast		No
	Sending voltage/V	[0.5,1]	Yes
	Limited resistor voltage /V	1.6	Yes
	Track circuit voltage/V	[15,24]	Yes
Glued insulation	Rail voltage/V	Under 1	Yes
	Broken trough		Yes
	Rail joints [8]/mm	[6,10]	Yes
	Measured resistance/Ω	More than 20Ω	Yes
	Inside voltage/V	0	Yes

	Outside voltage/V	0	Yes
	Leakage current/A	0	No

The effects of weather, ballast and leakage current are not big in the short time, no consider temporarily. To analysis Table 1, nine characteristic components are selected. Measured sample data is within the normal range when insulation joints are good. If not, are broken. As a consequence, to determine what are the statues of insulation joints is to judge sample characteristic components are normal or not. A wide range of decision rules are provided to define four states of insulation joints in Table 2.

Table 2. Rules

States of insulation joints	Description
Normal	All characteristic components value were normal.
Slight damage	One of characteristic components was abnormal.
Medium damage	Two were abnormal.
Serious damage	At least three were abnormal.

4. KNN-BSMOTE-SVM Algorithm

For SVM research, the number of samples of various types of training sample data sets is assumed to roughly equal, that is, the prior probability distribution is balanced or misclassification cost is equal of each type. But there is lots of normal data, abnormal data is very limited during the study, the experimental data is an unbalanced data set. For research of damage to insulation joints with unbalanced data sets, the damaged data better reflects states. Although information asymmetry has higher overall classification accuracy, increasing error rate of the positive sample information, the classifier generalization capability is not high. Therefore, to research insulation joints with unbalanced data sets is crucial. Currently, unbalanced research mainly focuses on the level of algorithms and data, a SVM classification algorithm based on combination of KNN and BSMOTE is proposed to remove redundant and noise samples to improve classification accuracy.

4.1 KNN Under-Sampling

KNN, K-Nearest Neighbor, is promotion to nearest neighbor method, with rules, to observe k nearest known samples of unknown sample, $\boldsymbol{x}$ belongs to category of a greater number among k nearest neighbors. The nearest neighbor is calculated according to Euclidean distance. In KNN algorithm the value of the parameter k is determined by experience. Algorithm idea can be described as follows.

Step1: To calculate the distance $d(\boldsymbol{x}_i, \boldsymbol{x}_p)$ between each unspecified sample $\boldsymbol{x}_p$ and $\boldsymbol{x}_i$.

Step2: To Select k smallest distances sample with, named $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_k$.

Step3: The categories of k samples reference numeral category label of $\boldsymbol{x}_p$.

KNN under-sampling method is divided into four subdivisions, that is, NearMiss-1, NearMiss-2, NearMiss-3 and “most remote”. NearMiss-1 selects the negative sample with average minimum distance to the nearest three positive sample. While NearMiss-2 elects the negative sample with average minimum distance to the farthest three positive sample. NearMiss-3 gives each positive sample for a given number recent the negative to ensure the positive is surrounded by negative. The “most remote” method selects the negative sample with the maximum average distance to the nearest three positive sample. Experimental results have shown NearMiss-2 method have a better unbalanced optimal performance [9].

4.2 BSMOTE Over-Sampling

In addition to the negative by reducing the data to achieve a balanced, over-sampling method could also be used to increase the positive to complete balanced. Over-sampling method has two kinds of random replication and synthetic. Random copying method achieves a balanced by copying positive sample randomly, but is easy to over-fitting. To improve this deficiency, synthetic oversampling method is proposed which generated with a new algorithm, currently, more using this method. SMOTE is a powerful method in over-sampling, to use the positive sample to generate artificial ones, but with a certain blindness. Therefore an improved algorithm border SMOTE (Border Synthetic Minority Over-sampling Technique) is proposed on the basis of SMOTE, the algorithm is more attention to the distribution of the sample.

Support vectors are the main concepts of SVM algorithm, which determine the classification boundary, the algorithm could be used to copy only the positive sample near the classification boundaries to achieve equalization of training samples.

BSMOTE oversampling algorithm specifically describes as follows (T is training set, the positive sample $F=\{f_1,f_2,\ldots,f_n\}$).

Step1: According to k nearest neighbors of each sample in F of the positive in the training sample set T to classify F sample. If k nearest neighbors are the negative sample, set as the noise sample, then put into set N. If they are the positive ones away from classification boundaries, put into B. If mixing, put into S.

Step2: Take out boundaries sample set $B=\{\boldsymbol{f}_1',\boldsymbol{f}_2',\ldots,\boldsymbol{f}_b'\}$, to calculate k nearest neighbors $\boldsymbol{f}_{ij}$ of each boundaries sample in the positive F, and randomly select s nearest neighbors, where $s \in (1,b)$. Then to count difference of properties, $d_{ij}=\boldsymbol{f}_i'-\boldsymbol{f}_{ij}$, $j=1,2,\ldots,s$. And to multiply by a random number r_{ij}, which is taken to between 0 and 1. Final artificially generates positive sample $\boldsymbol{h}_{ij}$, $\boldsymbol{h}_{ij}=\boldsymbol{f}_i'+r_{ij}\times d_{ij}$, $j=1,2,\ldots,s$. In addition, when the k nearest neighbors sample belong to the collection of N or S, $r_{ij} \in (0,0.5)$.

Step3: To repeat Step2 until the positive achieves a balanced sample set, then to finish algorithm.

4.3 KNN-BSMOTE-SVM Algorithm

Combining above sampling methods, a mixed algorithm, which combines KNN under-sampling and BSMOTE over-sampling, the KNN-BSMOTE-SVM is proposed. The algorithm could remove noise and duplicate information of the negative sample, meanwhile, increase boundary sample information of the positive in the case of retaining useful information of the negative, improving the utilization of data and achieving a balanced sample. The algorithm flow is as shown in Figure 1.

5. Experimental Analysis

According to track circuit and rubber insulation test data, KNN-BSMOTE-SVM algorithm, named as A-algorithm, is used to forecast the state of insulation joints with unbalanced data set. In order to verify the feasibility of SVM, SVM and KNN classifiers are selected to classify experimental data, then compare results.

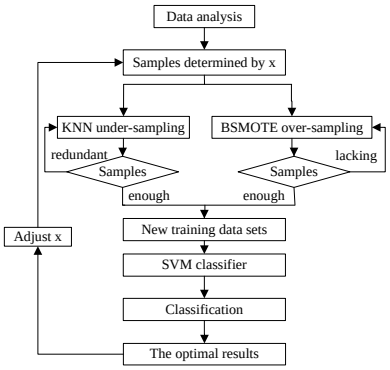

Figure 1. KNN-BSMOTE-SVM Algorithm Flow

5.1 Experimental Data

Data is track circuit and rubber insulation test data that comes from a certain electricity department. We have to convert the raw data into digital vector that SVM could recognize. For example, "Broken trough" is a Chinese character not a digital vector, we make "0" represent normal, "1" is abnormal. Characteristic parameters of the sample data show in Table 1, and Table 3 shows experimental feature information. Parameters of SVM could be got through training, damage to insulation joints is a multi-classification problem. "one-against-one" MSVMs is chose to train and test sample in this paper, *A*, *B*, *C* and *D* represent one category of normal, slight damage, medium damage and serious damage. Meanwhile, considering fail-safe, to prevent the failure to upgrade, the highest degree of damage is determined to be the smallest tag number. So, a SVM prediction model is established.

Table 3. Raw Data Sets

Data sets	Properties	Number of negative sample	Number of positive sample	Number of categories
Track circuit	3	39	22/17/3	4
Rubber insulation	6	112	85/4/8	4
Mixed data	9	39	22/17/8	4

5.2 Evaluation Standard

To classify unbalanced data, we tend to be more concerned about the positive sample, but the proportion of it is small, easy to classify wrongly. The traditional sorter usually takes the accuracy as the evaluation criteria to pursue high rate of accuracy. This kind of evaluation standard in unbalanced study of the question is obviously incorrect. Therefore, this paper uses accuracy, precision, recall and F-measure to evaluate and compare the precision model, these performances come from hybrid matrix. The hybrid matrix is as shown in Table 4.

Table 4. Hybrid Matrix

	Forecast number of negative sample	Forecast number of positive sample
Actual number of negative sample	TN	FP
Actual number of positive sample	FN	TP

N represents the negative sample, *P* is the positive, *T* and *F* are shorthand of TRUE and FALSE. TN is the number of forecasting the negative sample correctly. FN is the number of forecasting the negative sample wrongly. FP is the number of forecasting the positive sample wrongly. TP is the number of forecasting the positive sample correctly. It is follows:

Accuracy, the proportion of forecast results and actual.

$$Accuracy=(TP+TN)/(TN+FP+FN+TP) \quad (10)$$

Precision, the actual positive samples in prediction.

$$Precision=TP/(FP+TP) \quad (11)$$

Recall, the proportion of the positive.

$$Recall=TP/(FN+TP) \quad (12)$$

F-measure, harmonic mean of Precision and Recall.

$$F=2/((1/precision)+(1/Recall)) \quad (13)$$

5.3 Simulation Results

Table 5 shows the new data dealt with KNN-BSMOTE-SVM algorithm. Figure 2 and 3 give performance and training status curve of mixed data in equilibrium process.

Table 5. New Data Sets

Data set	Track circuit	Rubber insulation	Mixed data
New samples	19	79	14
Test samples	60	160	80

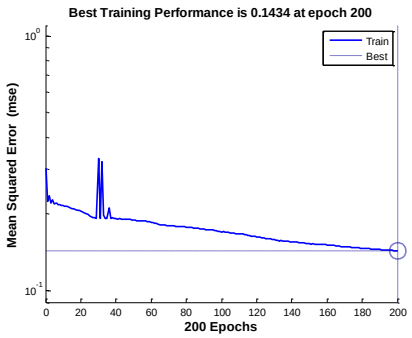

Figure 2. Performance Curve

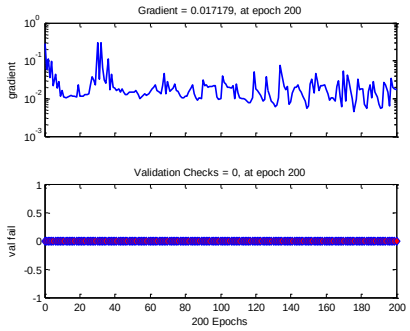

Figure 3. Training Status Curve

SVM parameters are selected by PSO algorithm, Gaussian function selects as kernel function. PSO settings: accelerating factor C1=1.5, C2=1.7, population is 20, 200

iterations. The nearest neighbor parameter k of BSMOTE and KNN algorithm makes 5, with a 10-fold cross-validation. The results are as shown in Table 6.

Table 6. Comparison of Experimental Results (%)

Data	Algorithm	Accuracy	Precision	Recall	F
Track circuit	KNN	53.75	85	40	54.4
	SVM	75	83.73	73.29	74.8
	A-algorithm	73.33	74.34	73.33	71.13
Rubber insulation	KNN	89.07	90	89.09	89.54
	SVM	92.75	90.56	90.56	90.45
	A-algorithm	93.13	93.57	93.13	93.1
Mixed data	KNN	53.99	87.08	35	49.93
	SVM	72.41	62.95	68.54	62.13
	A-algorithm	88.75	89.33	88.75	88.2

The feature is not absolute independence in the prediction of insulation joints. Analysis Table 6, concluding: SVM is more suitable for prediction of damage to insulation joints, compared with KNN classification algorithm. SVM algorithm after equalization has improved the classification performance of SVM. Meanwhile, the results of three types data show that sole rubber insulation data could better reflect the state of insulation joints, we could use these data to predict damage to insulation joints directly. Figure 4 and 5 give Recall-Precision curve of normal and serious damage insulation joints based on rubber data. In summary, it is feasible to apply SVM to predict damage to insulation joints.

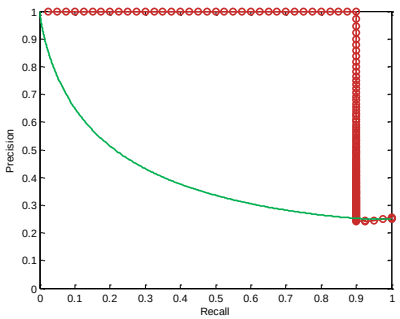

Figure 4. Normal Curve

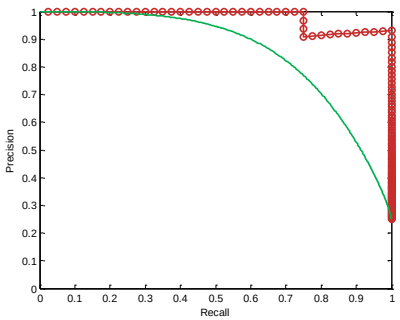

Figure 5. Serious Damage Curve

6. Conclusion

The more detailed insulation joints state can better reflect the damaged degrees of insulation joints, it is convenient for on-site staffs to take a more reasonable and timely response measures according to different extents of damaged. Accurate prediction can ensure the safety of railway transportation and normal economic benefits. SVM algorithm is used to handle small sample data, commendably solve nonlinear problems, this paper applies SVM to predict the extents of damage to insulation joints in the railway. The experimental data is unbalanced in the prediction process. Therefore, This paper uses the KNN-BSMOTE-SVM algorithm in terms of data to achieve balanced data. That's because, the algorithm combines under-sampling and over-sampling method, making up the defects of single sampling method, removing redundant and noise samples and improving utilization rate of the effective data. The experimental results show that SVM algorithm is more suitable for prediction of damage to insulation joints than KNN algorithm, and balanced data can improve the generalization ability of SVM classifier from the evaluation results. What's more, from the experimental data, forecasting results of the individual rubber insulation data are better, it can reach more than 90%. It should be noted that attribute reduction to original data is the next research direction.

Acknowledgement

The author would like to thank all colleagues who contributed to this study. This work is partially funded by National Natural Science Foundation of Science Fund Projects under the Project No.61164010.

References

[1] Y. Shiwu, J. Xiyi and W. Xinghui, "Study and Simulation Tests of Burning Damage to Insulation Joints and Rails in High-speed Railway Stations", Journal of the China Railway Society, vol. 35, no. 10, **(2013)**, pp. 82-88.
[2] T. Rao, N. Rajasekhar and Dr T. Rajinikanth, "An Efficient Approach for Weather Forecasting Using Support Vector Machines", International Conference on Computer Technology and Science, **(2012)**, pp. 208-213.
[3] N. LaoutiSami and Othman, "Combination of Model Based Observer and Support Vector Machines for Fault Detection of Wind Turbines", International Journal of Automation and Computing, vol. 11, no. 3, **(2014)**, pp. 274-287.
[4] Y. He, C. Y. Du and C. B. Li, "Sensor Fault Diagnosis of Superconducting Fault Current Limiter with Saturated Iron Core Based on SVM", IEEE Transactions on Applied Superconductivity, vol. 24, no. 5, **(2014)**.
[5] Z. Hua and Z. Jie, "Wind Speed Forecasting Model Study Based on Support Vector Machine", Acta Energiae Solaris Sinica, vol. 31, no. 7, **(2010)**, pp. 928-932.
[6] L. Gang, "The Research of Cost Sensitive SVM Supervised Learning", Nan Jing, Inrolle of Education Science, Nanjing Normal University, **(2007)**, pp.26-27.
[7] T. Xinmin, L. Furong, T. Zhijing and Y. Libiao, "Novel Fault Detection Method Based on SVM with Unbalanced Data Sets", Journal of Bration and Shock, vol. 29, no. 12, **(2010)**, pp. 8-12.
[8] Ministry of Railways, Technical Standards for Railway Signal Maintenance Rules (I), Beijing, China Railway Publishing House, **(2008)**.
[9] J. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distributions, A Case Study Involving Information Extraction", Proceeding Int'l Conference Machine Learning (ICML'2003), Workshop Learning From Imbalanced Data Sets, **(2003)**.

Authors

Dong Yu, (1962), male, professor, director of master, engaged in the study on Railway Signal. His research interests include Computer Interlocking, Railway Signal Applied Technology *etc.*

Xiao Ziqiang, (1989), female, a student of school of Automation and Electrical Engineering in Lanzhou Jiaotong University, master, major in Traffic Information Engineering and Control. Received a B.S. degree in Automatic Control from Lanzhou Jiaotong University 2013. E-mail: ziqiangx@yahoo.com.