Neurocomputing 151 (2015) 296–303

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Steel plates fault diagnosis on the basis of support vector machines

Yang Tian, Mengyu Fu, Fang Wu *

State Key Laboratory of Robotics and Systems (HIT), Harbin Institute of Technology, Heilongjiang 150001, China

ARTICLE INFO

Article history:
Received 30 June 2014
Received in revised form
15 August 2014
Accepted 19 September 2014
Communicated by Hongli Dong
Available online 7 October 2014

Keywords:
Fault diagnosis
Support vector machine
Machine learning
Steel plate faults dataset
Parameter optimizing

ABSTRACT

Fault diagnosis is always a big concern in industry production. As industrial technology has developed a lot, new fault diagnosis methods are needed to distinguish faults with only fine distinctions. The higher quality a production is required to have, the better fault diagnosis method the factories should apply. A fault diagnosis method based on modified Support Vector Machines (SVMs) is shown in this paper. With this method, dimension of samples is effectively reduced by recursive feature elimination (RFE) algorithm, and computing time is saved at the same time. Besides, classification accuracy is improved by parameter optimizing and sample size balancing strategy. A faults dataset of steel plates is taken as a practical case. And SVMs that are modified by different algorithms are utilized to complete fault diagnosis. This combined measure shows its superiority in sorting common faults of steel plates over original SVMs. Some essential procedures in model developing, such as normalization and cross validation, are also referred to.

1. Introduction

Fault diagnosis is aimed at discovering the time, location, and size of certain faults in modern industry process [1]. It is usually based on available data on the spot and systematic fault classification records. An effective method that can determine fault types and causes will not only lower maintenance cost and unexpected waste, but also improve production efficiency and quality level of products [2]. Besides, further treatments such as recycling are also based on accurate fault diagnosis [3]. Traditionally, even experts with fault diagnosis manuals have to analyze operational environments carefully to infer potential causes of a particular fault. However, more intelligent means, derived from study of machine learning, have developed a lot to address this problem quickly and correctly [4]. Typically, they include artificial neural networks (ANNs) [5], logistic regression (LR) [6], decision tree (DT) [7], principal component analysis (PCA) [8], correspondence analysis (CA) [9], canonical variate analysis (CVA) [10], kernel independent component analysis (KICA) [11] and support vector machines (SVMs) [12].

Among all these algorithms, ANNs have a relatively long history and have been widely used in machine learning and related fields [13]. ANNs are computational models inspired by an animal's central nervous systems (in particular the brain). Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs [14]. However, ANNs have obvious shortcomings. For example, structure of the network has to be specified in advance or fixed by heuristic algorithm during training process, which is particularly difficult for multi-layer network; network's weight coefficients have very limited adjusting methods; solution found by ANNs may be local optimal; quality of the model depends too much on quantity and quality of training samples; and no theory has been found to analyze the rate of convergence of training quantitatively [15]. The deadliest problem is: ANNs concern only empirical risk, neglecting whether empirical risk converges to real risk or not and the corresponding convergence condition [16]. All these shortcomings limit the use of ANNs and their variants, especially when the designers do not have enough experience and priori knowledge. In contrast with traditional machine learning algorithms with typical shortcomings, SVM has been provided with incredibly complete theoretical backgrounds [17]. It has shown several major advantages in the field of machine learning [18]. First, SVM excels in addressing high-dimension problem and solving the problem of small sample with the help of kernel function [19]. Second, SVM has better generalization ability than traditional machine learning algorithms thanks to consideration of structural risk minimization theory. It overcomes the appearance of over-fitting and under-fitting in ANNs. Third, the final solution of SVM is unique and globally optimal. Moreover, special dimension mapping algorithm makes computing time short when solving high-dimension problems, which cannot be achieved by ANNs, LR, DT, etc. In fault diagnosis problems, a huge number of features are taken into consideration and many practical occasions require high efficiency of working. Thus, SVM is regarded as a proper choice. For example, in a similar case of fault diagnosis, worn cutting knives are

* Corresponding author.
E-mail address: wufangzailushang@gmail.com (F. Wu).

http://dx.doi.org/10.1016/j.neucom.2014.09.036
0925-2312/

required to be distinguished from unworn ones. In this case, ANNs have sorting accuracy of 79.2% and SVM has that of 91.7% [20].

Recursive Feature Elimination (RFE) is an attribute extracting algorithm [21]. The relevant degree of every attribute to the known label is pondered according to its sorting coefficient, which is obtained based on weight factor ω of SVM model in each iteration. Eventually, all features are ranked in a descending order, and an appropriate number of top features can retain enough information for classification. In this way, dimension of data is reduced [22].

In the first part of this paper, brief introduction of basic ideas of SVM and SVM-RFE is presented, including their mathematical backgrounds and introduction of practical applications. Afterwards, parameter optimizing algorithms and sample size balancing criterion are concerned as assistance for SVM. Then common faults recorded in steel plate faults dataset are selected as an object of fault diagnosis and is used to test classification performances of concerning algorithms. Three main parts compose fault diagnosis of steel plates, namely initial treatment, full dimensional classification and reduced dimensional classification. At last, precisions of different methods are compared and corresponding conclusions are made. This example shows the importance of proper parameter setting and feature set choosing in SVM classification.

2. Related words

2.1. Support vector machine

Vapnik was one of the founders of statistics learning theory and also a major inventor of SVM. SVM algorithm was first proposed by a team lead by Vapnik in AT&T BELL laboratories in 1995. It was initially used in pattern recognition. In this initial test, SVM shows better accuracy than ANNs, reaching 80%. In the latest 20 years, SVM developed a lot and many variants appeared. Usually in SVM training process, the problem to be solved is divided into subproblems, thus the final result can be obtained by solving all these subproblems. According to the division of subproblems and iteration strategy, there are two types of SVM: chunking algorithm and fixed training set algorithm. Fixed training set algorithm has four fast algorithms: SVM-light, SMO, BSVM and LIBSVM. Meanwhile, by adding function items, variables or factors, deciding function of SVM can be changed to cater needs in different circumstances. Among these variants, υ-SVM uses parameter υ to control the number of support vectors, making it easier to select proper support vectors; double υ-SVM can assign different error rates for all varieties flexibly. One-class SVM comes up with the idea of hypershpere, solving problems that cannot be solved by hyperplane. Weighting SVM compensates the negative effect of category differences to improve classification accuracy. Fuzzy SVM (FSVM) provides fuzzy memberships for every sample to promote accurate rate. Direct SVM sets up identification methods and rules for unknown samples according to known samples, which is superior to traditional inductive learning.

SVM claims advantages over other machine learning algorithms on small, nonlinear or high dimensional datasets. The principles on which SVM is built are VC dimension theory and structural risk minimization theory [23]. VC dimension is introduced to weigh complexity of a function. Structural risk concerns two major parts, namely empirical risk and confidence. The former is regarded as main objective by traditional learning machines, leading to a satire that the training model can classify train dataset a hundred percent correctly but totally messes up test dataset. So structural risk has to be taken into consideration to improve reliability when predicting samples that are not in train dataset.

To find solution to nonlinear problems, kernel functions are introduced to map the low dimensional space onto a high dimensional one, which can realize inner product in the low dimensional space and eschew strenuous computations in high dimension with nonlinear functions. Mathematical background of basic SVM is introduced in detail as follows.

Consider sample sets:

$$T=(x_i,m_i)_{i=1}^{N} \tag{1}$$

where x_i denotes a sample in input data whose size is N and m_i denotes label of a sample, which is either $m_i=1$ or $m_i=-1$. Assume that a hyperplane that classifies the samples labeled as 1 from those labeled as -1 exists. Its mathematical expression is

$$w^T\cdot x+b=0 \tag{2}$$

where w is the normal vector of the hyperplane, which is also known as weight vector, and b is the constant bias.

The greater the distance between sample points and the hyperplane is, the more possible the sample can be classified correctly. Thus, overall object is to find the best plane, or classifier, which can maximize the minimum geometric margin from points in sample set. The simplified optimization problem is

$$\begin{aligned}&\min\quad \frac{1}{2}\|\omega\|^2\\ &\text{s.t.}\quad y_i(\omega^T\cdot x_i+b)\geq 1,\quad i=1,2,\ldots,N\end{aligned} \tag{3}$$

Noticing it is a standard quadratic programming (QP) optimization, its dual problem can be achieved by Lagrange duality, which puts constraint condition into the objective function:

$$\min_{\omega,b}\max_{\alpha_i\geq 0}\left[\frac{1}{2}\|\omega\|-\sum_{i=1}^{N}\alpha_i(m_i(\omega^T\cdot x_i+b)-1)\right]=p^* \tag{4}$$

where α_i is the Lagrange multiplier. Because the problem satisfies Karush–Kuhn–Tucker (KKT) condition, the positions of min and max can be exchanged without changing optimization solution:

$$\max_{\alpha_i\geq 0}\min_{\omega,b}\left[\frac{1}{2}\|\omega\|-\sum_{i=1}^{N}\alpha_i(m_i(\omega^T\cdot x_i+b)-1)\right]=d^* \tag{5}$$

Solution to Eq. (5) is easy to be attained. Firstly, α_i is fixed and extreme point of ω and b can be computed:

$$\frac{\partial\Gamma(\omega,b,\alpha)}{\partial\omega}=0\Rightarrow\omega=\sum_{i=1}^{N}\alpha_i m_i x_i \tag{6}$$

$$\frac{\partial\Gamma(\omega,b,\alpha)}{\partial b}=0\Rightarrow\sum_{i=1}^{N}\alpha_i m_i=0 \tag{7}$$

Substitute Eqs. (6) and (7) back into Eq. (5), and Eq. (8) can be obtained:

$$\Gamma(\omega,b,\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j m_i m_j x_i^T x_j \tag{8}$$

Solution of α can be acquired by Sequential Minimal Optimization (SMO) and the ultimate results are

$$\omega=\sum_{i=1}^{N}\alpha_i m_i x_i \tag{9}$$

$$b=-\frac{1}{2}\left(\max_{m_i=-1}\omega^T\cdot x_i+\min_{m_i=1}\omega^T\cdot x_i\right) \tag{10}$$

The decision function for test samples is developed to be

$$m=\operatorname{sgn}\left(\sum_{i=1}^{N}\alpha_i m_i\langle x_i,x\rangle+b\right) \tag{11}$$

Kernel function is utilized to simplify the projection from a low dimensional space onto a high dimensional one. Originally, for a

nonlinear problem, nonlinear function $\psi(n)=(\psi_1(x),\ldots,\psi_i(x))$ replaces linear part of Eq. (10), so Eq. (11) becomes

$$m=\operatorname{sgn}\left(\sum_{i=1}^{N}\alpha_i m_i\langle\psi^T(x_i),\psi(x)\rangle+b\right) \tag{12}$$

However, kernel function makes it possible to realize inner product in a low dimension and classification in a high dimension:

$$m=\operatorname{sgn}\left(\sum_{i=1}^{N}\alpha_i m_i K\langle x_i,x\rangle+b\right) \tag{13}$$

If the case is linearly non-separable, slack variables and penalty factors should be introduced to tolerate outliers that have an adverse impact on determination of hyperplane. As a consequence, the optimization problem becomes

$$\begin{aligned}&\min\quad \frac{1}{2}\|\omega\|^2+E\sum_{i=1}^{N}\xi_i\\&\text{s.t.}\quad m_i(\omega^T\cdot x_i+b)\geq 1-\xi_i\\&\qquad \xi_i\geq 0,\quad i=1,2,\ldots,N\end{aligned} \tag{14}$$

where ξ_i is the slack variable and E is the penalty factor. Its dual problem is

$$\begin{aligned}&\max\quad \sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j m_i m_j\langle x_i^T x_j\rangle\\&\text{s.t.}\quad 0\leq\alpha_i\leq E\\&\qquad \sum_{i=1}^{N}\alpha_i m_i=0,\quad i=1,2,\ldots,N\end{aligned} \tag{15}$$

The result can be obtained via precedent solution.

Since kernel function plays an important role in setting up SVM model, choosing an appropriate kernel function is of significance. Four kernel functions, including linear function, polynomial function, radial basic function and sigmoid function, are most widely used. In this paper, all SVM models are built based on radial basic function, it has the following expression:

$$K(x,y)=\exp\left(-\gamma\times\|u-v\|^2\right) \tag{16}$$

As is shown above, SVM is designed for two-class classification. Nevertheless, it can be generalized to solve multi-class classification problems. Commonly used approaches are one-versus-rest SVMs (1-v-r SVMs), one-versus-one SVMs (1-v-1 SVMs) and hierarchical SVMs (H-SVMs) [24]. Assume that N is the total number of categories. In the first method, there will be N SVM models, judging whether a sample belongs to certain class or not. Then every sample is filtered through the N models, thus its label is settled by the model which decides the sample with maximal classification value. While in the second method, the number of models will be increased to $N(N-1)/2$. Then a test sample will get its category by being voted according to the results of these models. As for H-SVMs, a binary decision tree is constructed. In every round, a test sample will be categorized into either of the two subsets and this process continues until there is only one category in the subset. Lately, a new method proposed by Weston is aimed at solving multi-class classification problems directly by SVM. However, complexity of QP problem impedes extensive use of the algorithm. In this paper, both 1-v-1 SVMs and 1-v-r SVMs are utilized.

2.2. Recursive feature elimination algorithm

RFE algorithm aims at getting a ranking list of sample's attributes. In the training process of SVM, weight vector w can be attained and sorting coefficients can be computed subsequently. The minimum coefficient hints that the corresponding feature is relatively unimportant in contrast with other features. So such feature is eliminated and only the rest are considered in next round of model development. Similarly, the second relatively unimportant feature is figured out. When the iteration finishes, a complete importance descending list is acquired. First few features of the list are selected to compose a feature subset. When accuracy of classification is considered, the optimal feature subset can be found out by comparing classification results using different subsets [25].

Due to the fact that the higher dimension one problem has, the harder its solving process becomes, dimension reduction is a significant method in fault diagnosis [26]. RFE algorithm is a relatively outstanding data processing method that can reduce the dimension of one problem [27]. As dimension is reduced, computing time can be shortened as well. Besides, not all features are closely related to the category of a sample. Treating all features equally may cause dominating factors to be buried. So it is usually better to ignore some unimportant features and focus on features that are more decisive [28].

However, conflict between accuracy and computing time should be balanced [29]. In this paper, curve of accuracy is plotted to provide reference for determining feature subset.

3. Algorithm

3.1. Optimization

In this paper, two variables need to be optimized—penalty factor and γ in radial basic kernel function. The aim of optimization is to improve accurate rate. Three different methods are applied and two of them are introduced in detail as follows.

3.1.1. Genetic algorithm

Biological evolution is a marvelous optimization process. By selective elimination, sudden mutation and gene crossover, excellent species that can adapt to the environment best are found. Genetic algorithm (GA) is a heuristic globally optimizing algorithm inspired by the evolutionary mechanism. It has been applied broadly in pattern recognition, artificial neural networks, picture processing, machine learning, industrial optimal control, adaptive control and bioscience [30]. In GA, potential solutions are regarded as chromosomes, namely some strings of binary code. Then these initial individuals are pondered by their fitness values and appropriate parents are selected to produce next generation by mutation and crossover. After several iterations, fitness values of some individuals meet stop criterion and the final optimization result is spotted. GA is good at locating globally optimal solution because it takes all individuals into consideration and processes them at the same time [31]. However, parameters' setting is of great significance in the optimization and can lead to premature convergence. Besides, the number of initial population can have vital impact on the performance. If initial population contains too many individuals, individuals with poor fitness values will take up resources and operation time is prolonged as well; but if it contains too few individuals, optimal results are more likely to be neglected [32].

3.1.2. Particle Swarm Optimization

Particle Swarm Optimization (PSO), proposed by Kennedy, Eberhart and Shi, is another artificial optimizing algorithm. It is based on observation and simulation of social foraging behavior of animals. It has been utilized in fitting function, training artificial neural networks, adjusting parameters and referred to accomplish combinatorial optimization. The optimal solution is treated as a kind of food for particles. Special tricks are employed by them to approach food. They change their locations and velocities

according to the distance between their present positions and current known best position (where food is most likely to be stored). At first, these particles' distribution is dispersive and random. By continuously adjusting their directions and speeds, they eventually get together at the position of globally optimal position. Obviously, PSO is distinctive from GA, since there is no mutation and crossover in PSO and globally optimization is obtained by following the current optimal particle. However, irrational parameter setting can slow down convergence speed seriously [33].

3.2. Penalty factor and sample size balancing

Assume that there are two categories (A and B) to be classified. And category A has relatively small sample size. In this condition, SVM model tends to ignore samples in category A and predict that most of the testing samples belong to category B. Although overall accuracy seems to be high for train dataset, but model's generalization ability is poor. This high accuracy is specious. In other words, this kind of classification has no practical value.

One way to solve the problem is to set greater penalty factor for category A, showing that samples in this class are valued and the model should not give them up. This method greatly improves classification accuracy for category A, but that of category B may become lower [34]. For steel plate faults diagnosis, penalty factor is determined by optimizing methods.

In multi-classification, it is impractical to set proper penalty factors for all classes. Thus, sample size balancing is significant in optimizing process of parameters. Because qualities of different parameter settings are determined by results of cross validation, parameters with specious accuracy may be determined as final choice. Conclusively, sample sizes should be balanced before looking for proper parameters. Thus, generalization ability of SVM model can be optimized. However, in realistic cases, it is not always easy to get samples that distribute in every category equally.Since major aim is to make sample size of different classes equal to each other, two simple ways can be adopted. The first one is to give up some of the samples, making every category has the same sample size as the class that has the least samples. The second method is to use some of the samples repeatedly, enlarging sample sizes to keep balance among categories. If the first method is applied, some information must be sacrificed. There is a possibility that support vectors are different because of information loss, which will greatly influence the accuracy of SVM predicting. So there is hardly any choice that the second method should be implemented. To avoid preferring some of the samples, all samples in one category are repeatedly utilized for the same time. A method of using samples repeatedly is applied in this paper. Assume that there are k categories to be classified, and their sample size can be noted as k_i. Among these categories, class m has the largest sample size k_m which is taken as standard size. All samples in the other $k-1$ categories are used repeatedly and repeating time c_i can be calculated as follows:

$$c_i = \left\lfloor \frac{k_m}{k_i} + \frac{1}{2} \right\rfloor \tag{17}$$

In this way, all samples in one category are treated equally, and the sample size differences among categories are successfully compensated.

4. Steel plate faults dataset

Steel plate is an important raw material in hundreds of industry manufactures. Mature as its producing technology has been, there are still products of inferior quality that need to be picked out and sorted for further treatment. Steel plate faults dataset is provided by Semeion, Research of Sciences of Communication, Via Sersale 117, 00128, Rome, Italy. In this dataset, the faults of steel plates are classified into 7 types. Since it has been donated on October 26, 2010, this dataset has been widely used in machine learning for automatic pattern recognition. Types of fault and corresponding numbers of sample are shown in Table 1.

As is shown in Table 1, the numbers of sample vary a lot from one category to another. Meanwhile, fault 7 is a special class because it contains all other faults except the first six kinds of fault. In other words, samples in class 7 may have no obvious common characteristics.

For every sample, 27 features are recorded, providing evidences for its fault class. All attributes are expressed by integers or real numbers. Detailed information about these 27 independent variables is listed out in Table 2.

Different features have different value ranges, and some have extremely small ranges in contrast with other features. To balance effect of different attributes in SVM model building, all attributes should be normalized to a conventional range. In steel plate faults diagnosis, the following equation is used:

$$\hat{f} = \hat{f}_{\min} + \frac{f - f_{\min}}{f_{\max} - f_{\min}} \times (\hat{f}_{\max} - \hat{f}_{\min}) \tag{18}$$

where $\hat{f}_{\max}$ and $\hat{f}_{\min}$ are desired ranges, $f_{\max}$ and $f_{\min}$ are original range of certain attribute. In the following computation, $\hat{f}_{\max} = 1$, $\hat{f}_{\min} = 0$.

5. Fault diagnosis

5.1. Initial treatment for fault 7

Steel plate faults data is recorded in the form of matrix, in which every column from 1 to 27 represents an independent attribute and every row represents a single sample. Every column

Table 1
Types of faults and sample sizes.

Fault type	Type of faults	Number of samples
1	Pastry	158
2	Z Scratch	190
3	K Scratch	391
4	Stains	72
5	Dirtiness	55
6	Bumps	402
7	Other Faults	673
Overall number		1941

Table 2
Independent attributes of steel plates.

Number	Attribute	Number	Attribute
1	X Minimum	15	Edges Index
2	X Maximum	16	Empty Index
3	Y Minimum	17	Square Index
4	Y Maximum	18	Outside X Index
5	Pixels Areas	19	Edges X Index
6	X Perimeter	20	Edges Y Index
7	Y Perimeter	21	Outside Global index
8	Sum of Luminosity	22	Log of Areas
9	Minimum of Luminosity	23	Log X Index
10	Maximum of Luminosity	24	Log Y Index
11	Length of Conveyer	25	Orientation Index
12	Type of Steel_A300	26	Luminosity Index
13	Type of Steel_A400	27	Sigmoid of Areas
14	Steel Plate Thickness		

from 28 to 34 initially represents the corresponding categories. For example, only samples belong to category 1 are marked by 1 in column 28, only samples in class 2 are marked by 1 in column 29, and so on. For convenience of SVM training and testing, Column 28 is redefined as label column, in which number 1, 2, 3, 4, 5, 6, 7 represents the class of a sample, and columns from 29 to 34 are deleted. For every category, approximately half of all samples are chosen as train dataset while the other half are taken as test dataset (shown in Table 3).

In model testing, true positive samples are defined as samples that are correctly classified into class X, and true negative samples are samples that are sorted from class X aright. Applying these basic definition, sorting accuracy for class X is defined as

$$\alpha = \frac{TP+TN}{N} \tag{19}$$

where TP is the number of true positive samples, TN is the number of true negative samples, and N is the total number of all samples. Overall accuracy in multi-classification can also be defined as

$$\alpha = \frac{TP}{N} \tag{20}$$

where TP and N are also the number of true positive samples and the total number of all samples.

According to the fact that fault 7 is not a specific kind of fault but a combination of several faults that are different from fault 1 to 6, special treatment is needed for fault 7. Thus, to avoid being confused and disturbed by fault 7 in multi-classification, the first step of classification is to sort fault 7 out from all faults. This process can be approached by viewing fault 7 as category A and all other faults as category B. It is necessary to guarantee the accuracy of this step because further classification is based on this initial treatment.

It is quite difficult to pick samples of fault 7 out from other faults because samples in this class do not share certain specific attributes, meanwhile it is also hard to find dominating features for training samples in the other 6 classes. Moreover, some of the samples in class 7 may have similar features with samples that belong to other classes, which leads to a result that SVM tends to classify them to combinational faults rather than fault 7.

It is better to adopt 1-v-1 SVMs to meet the need of high accuracy. In other words, method of setting up 21 different models to judge whether a sample belongs to class A or class B should be applied. There are 6 models judging whether a sample belongs to class X or class 7, where X can be 1, 2, 3, 4, 5, 6. Their classification accuracies are shown in Table 4.

In Table 4, original accuracy is obtained without variable optimizing and promoted accuracy is obtained with grid search optimizing method. In 1-v-1 SVMs, the category that gets most votes is determined to be predicted class for a sample. Applying this deciding method, result in Fig. 1 is obtained. It shows that 92.5697% of samples of fault 7 are picked out from all samples.

Table 3
Train dataset and testing dataset.

Fault number	Train dataset		Test dataset	
1	80	50.6%	78	49.4%
2	100	52.6%	90	47.4%
3	200	51.2%	191	48.8%
4	40	55.6%	32	44.4%
5	30	54.5%	25	45.5%
6	200	49.8%	202	50.2%
7	350	52.0%	323	48.0%

Table 4
Classification accuracies for fault X and fault 7.

Model	Original accuracy (%)	Promoted accuracy (%)
Fault 1 vs fault 7	80.5486	83.2918
Fault 2 vs fault 7	90.3148	91.2833
Fault 3 vs fault 7	91.8288	92.4125
Fault 4 vs fault 7	90.9859	98.5915
Fault 5 vs fault 7	92.8161	92.2414
Fault 6 vs fault 7	73.1429	81.8095

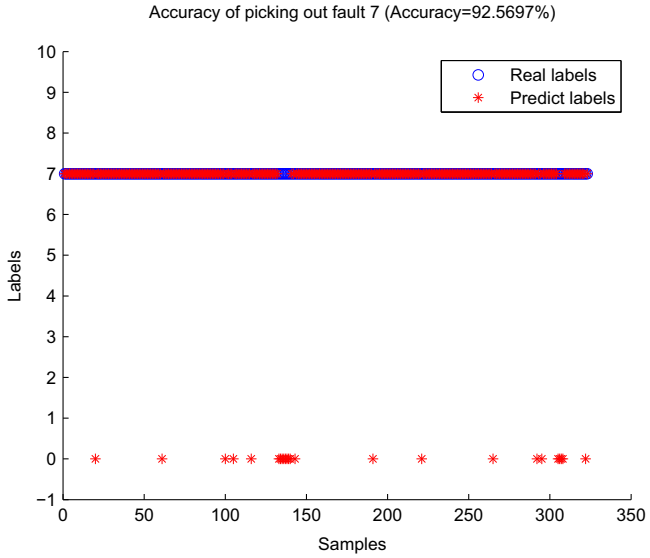

Fig. 1. Result of overall initial treatment.

5.2. Full dimensional multi-classification

In full dimensional multi-classification, SVMs are promoted by parameter optimizing, sample size balancing or both. Accurate rates of three different SVMs are listed out. It is shown that SVMs with balanced sample and parameters optimized by PSO has the highest accuracy.

Although samples of fault 7 are not perfectly removed from the dataset, following classifications ignore this blemish and is based on an assumption that there are only 6 types of faults in the dataset. 650 samples with fault 1 to 6 are chosen to be train dataset while the rest 618 samples are treated as test dataset.

Firstly, original SVMs are applied to classify all 618 samples. In original SVMs, penalty factor is 1 for all classes. Without assistance of variable optimizing methods, accuracies obtained from this model are only 81.2308% for train data and 68.7702% for test data, which is far from satisfactory. Corresponding result is shown in Fig. 2.

It can be told from Fig. 2 that categories with small number of samples, such as fault 5, are mostly classified to a wrong class which has a larger number of samples. When sample sizes are different between categories, model that developed in training process tends to discriminate some of the classes. This unbalance can be compensated by carefully arranged penalty factors. Nevertheless, it is difficult to determine value of penalty factors. Three different optimizing methods are utilized to find the best penalty factor respectively. They are Grid Search (GS), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Classification accuracies promoted by GS are 94.6154% for train dataset and 77.6699%

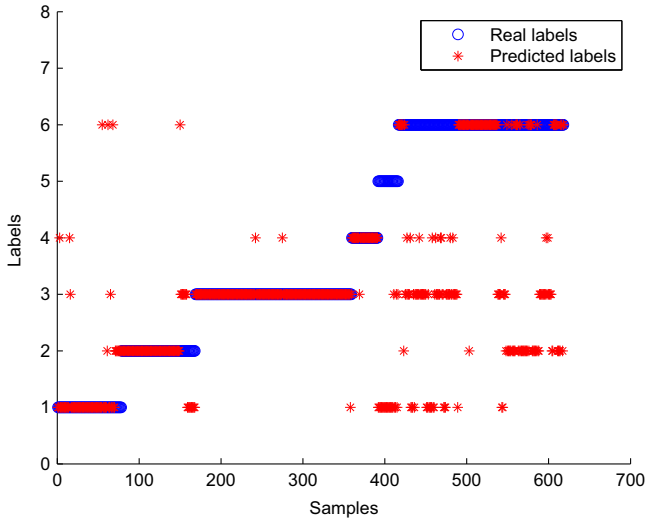

Fig. 2. Result of original SVMs without variable optimizing.

for test dataset. While GA improves training accuracy to 95.2308% and testing accuracy to 77.1845%. As for PSO, training precision is only 88% but testing precision is higher than previous two methods, reaching 77.9935%. Conclusion can be made that GA works the best in improving accuracy of train data classifying, but it also has the worst accuracy in sorting test data. And result optimized by PSO has the highest accuracy for test data and the lowest accuracy for train data. If testing accuracy is focused on, it can be found that there is no obvious difference between these three optimizing methods. But they all promoted testing accuracy by nearly 10% compared to the original SVMs.

Parameter optimizing methods are only capable in finding optimal value when the number of parameters is small. Under the condition of multi-fault diagnosis, it is preferred to set different penalty factors for every category in order to compensate sample size unbalance. As is mentioned above, setting penalty factors for all categories is difficult. So sample size balancing strategy is applied.

Though no additional information is put into train data, applying sample-size balanced SVMs (balanced SVMs) can force optimizing method to treat categories equally in the training process, successfully preventing specious accuracy caused by sample-size unbalance. After sample-size balancing measure and variable optimizing method are both utilized, training precisions optimized by GS, GA and PSO all reach 100%. Meanwhile, testing accuracies are improved by around 1.5%, reaching nearly 79%, which are 78.8026% for GS, 78.6480% for GA and 78.8026% for PSO.

Although result of PSO has exactly the same accuracy with that of GS, it takes much longer time to find the final solution. And computing time for GA is a little longer than GS. Setting proper initial parameters for GA and PSO is important for cutting down computing time. Because all parameter optimizing methods are aimed at finding globally optimal solution, the only reason why they show different testing accuracies is that parameters of possible values show equal ability in sorting train data, but generalization abilities of these models are not the same. Thus, solutions found by different methods may be completely different. Judging which of them works better in lowering error rate and saving computing time is of importance. All these discussions help to make final choice of SVM promoting strategy.

All accuracies for single categories in full dimensional multi-classification are shown in Table 5. Unbalanced SVMs and balanced SVMs both improve sorting precision for single category.

Table 5
Accuracies for single category.

Class	Original SVM (%)	Unbalanced SVM (%)			Balanced SVM (%)		
		GS	GA	PSO	GS	GA	PSO
Fault 1	89.2	93.5	93.7	93.7	93.5	93.9	93.9
Fault 2	89.2	90.9	90.9	91.6	90.3	89.6	93.7
Fault 3	87.5	93.2	92.9	93.4	95.2	94.8	93.7
Fault 4	97.3	98.4	98.2	98.5	98.1	98.1	98.9
Fault 5	96.0	96.1	95.6	96.4	96.9	96.8	99.0
Fault 6	78.5	83.2	83.0	84.3	83.7	83.2	79.0
Average	89.6	92.6	92.4	93.0	92.9	92.7	93.0

Table 6
Ranking list of attributes.

Rank	No.	Attribute	Rank	No.	Attribute
1	20	Edges Y Index	15	16	Empty Index
2	21	Outside Global index	16	10	Maximum of Luminosity
3	25	Orientation Index	17	22	Log of Areas
4	19	Edges X Index	18	24	Log Y Index
5	12	Type of Steel_A300	19	23	Log X Index
6	26	Luminosity Index	20	14	Steel Plate Thickness
7	17	Square Index	21	4	Y Maximum
8	13	Type of Steel_A400	22	3	Y Minimum
9	11	Length of Conveyer	23	18	Outside X Index
10	9	Minimum of Luminosity	24	8	Sum of Luminosity
11	2	X Maximum	25	6	X Perimeter
12	1	X Minimum	26	5	Pixels Areas
13	27	Sigmoid of Areas	27	7	Y Perimeter
14	15	Edges Index			

5.3. *Reduced dimensional multi-classification*

Based on full dimensional multi-classification, RFE algorithm is applied to confirm accuracy improvement and computing time saving when different numbers of attributes are extracted. The result shows that reduced dimensional classification works better than full dimensional classification.

Because some of the samples are repeatedly used, overall sample size is enlarged, leading to a result that computing time is also prolonged. One way to cut down computing time is to reduce dimension of the sample. As sample's dimension is reduced, computing efficiency can be obviously improved.

As is already mentioned above, SVM-RFE is a feature extracting method. By choosing a reasonable number of attributes, not only computing time is shortened, but also classification accuracy is guaranteed.

The ranking list of all 27 attributes in Table 6 is computed by SVM-RFE. The higher rank one attribute has, the more important it is in determining categories.

In reduced dimensional multi-classifications, both unbalanced and balanced samples are tested. Generally speaking, balanced SVMs get higher accuracies by 1% than unbalanced SVMs as is shown in Table 7. Focusing on achieving higher accuracy, results of balanced samples are discussed. After feature sets with different numbers of attributes are used, accuracy curves for test data is plotted.

Fig. 3 shows that classification accuracies remain the same when 1–13 attributes are removed from feature set. So it is believed that the last 13 features in the ranking list provide little help in classification. Moreover, when 5–7 least important features are abandoned, sorting precision is even improved. Conclusively, if computing time and precision are both focused on, the number of attributes should be around 13. If the only concern is overall accuracy, reserving about 19 attributes in the feature set may be

Table 7
Contrast of sorting accuracies.

Optimizing method	Full dimension				Reduced dimension (test)	
	Unbalanced (%)		Balanced (%)		Unbalanced (%)	Balanced (%)
	Train	Test	Train	Test		
GS	94.6	77.7	100	78.8	77.8	79.8
GA	95.2	77.2	100	78.6	78.0	80.3
PSO	88	78.0	100	78.8	79.6	80.7

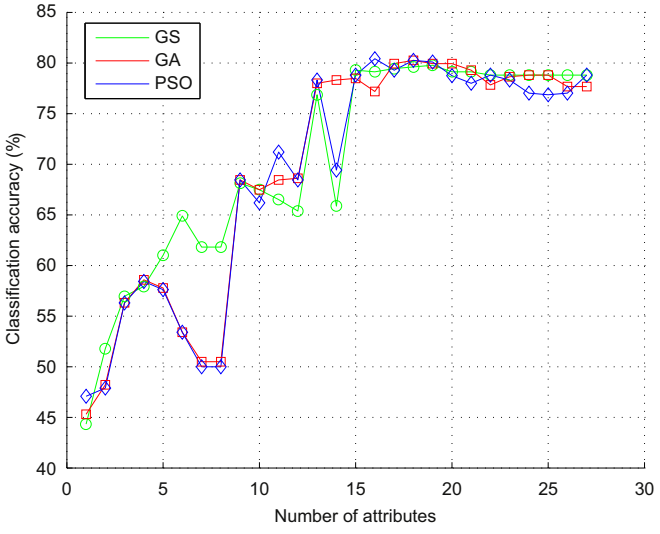

Fig. 3. Accuracy curves of RFE-SVM.

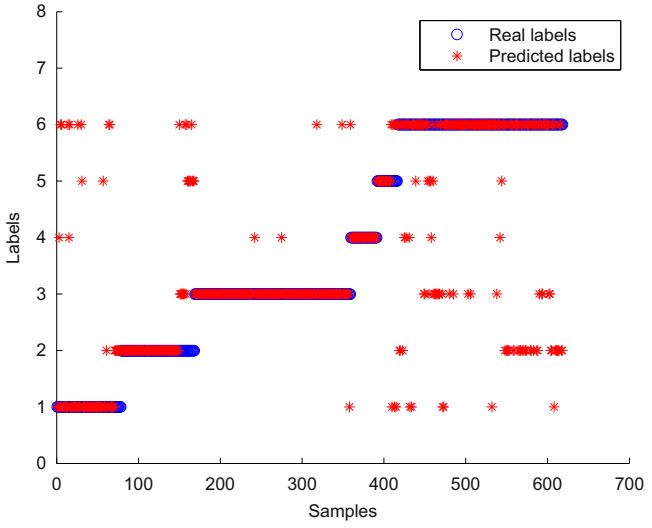

Fig. 4. Balanced SVM-RFE optimized by PSO.

a better choice. If SVMs are optimized by GS, overall accuracy reaches 79.7735% (77.8317% for unbalanced samples) when 19 attributes are reserved in the features set. Comparing to GS, GA works better in improving sorting precision, which reaches 80.2589% (77.9935% for unbalanced samples) if 18 features are reserved. Sorting accuracy promoted by PSO reaches 80.7443% (79.6117% for unbalanced samples) when subset of 20 features are chosen.

Finally, overall accuracies comparison of full dimensional SVMs and reduced dimensional SVMs is listed out in Table 7. According to all above, sorting accuracy finally reaches 80.7443% which is shown in Fig. 4.

6. Conclusion

Fault diagnosis of steel plates is divided into two parts. In the first part, samples of fault 7 which is a combination of various different faults are picked out from all samples. In the second part, multi-classification of fault 1–6 is concerned. In the second part, both full dimensional and reduced dimensional classification are adopted. In full dimensional sorting, SVMs with balanced sample sizes show its advantage over unbalanced SVMs, especially when sorting train dataset. Among all three different optimizing algorithms, PSO works the best to get higher accuracy but GS saves the most time of computing. Besides, SVMs with reduced dimension which combines sample size balancing, parameter optimizing and RFE algorithm works better in sorting test dataset. Among all 27 independent attributes, 7 least important features offer little help in classifying test dataset. Sorting accuracy is improved after abandoning them. Meanwhile, 14 of the 27 attributes are of significance in fault diagnosis. Using only these 14 attributes guarantees high accuracy and short computing time at the same time. In reduced dimensional classification, PSO has the best performance.

In conclusion, parameter optimizing, sample size balancing and dimension reducing method all have positive effect on SVMs' performance. With the help of all these three methods, precision of test data finally reaches 80.7443%, which is 11.9741% higher than original SVMs. However, there are still problems to be concerned. Because sample sizes are balanced according to a simple criterion, sorting precision of train data can reach 100% easily, causing a result that optimal parameter values with best generalization ability are not found successfully. If sample sizes of different classes are similar in train dataset and test dataset, an assumption can be made that probabilities of different faults' occurrence is certain in the producing process. Thus, other sample balancing criterion may bring better result than the one used in this paper.

Acknowledgment

This work is supported by State Key Laboratory of Robotics and System (HIT) with No. SKLRS-2014-MS-01 and Special Financial Grant from China Postdoctoral Science Foundation No. 2014T70339.

References

[1] R. Isermann, Model-based fault-detection and diagnosis—status and applications, Annu. Rev. Control 29 (1) (2005) 71–85.
[2] H. Dong, Z. Wang, H. Gao, Fault detection for markovian jump systems with sensor saturations and randomly varying nonlinearities, Circuits and Systems I: Regular Papers, IEEE Transactions on 59 (10) (2012) 2354–2362.
[3] S. Yin, S.X. Ding, A. Haghani, H. Hao, P. Zhang, A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process, J. Process Control 22 (9) (2012) 1567–1581.
[4] A. Widodo, B.-S. Yang, Support vector machine in machine condition monitoring and fault diagnosis, Mech. Syst. Signal Process. 21 (6) (2007) 2560–2574.
[5] I. Basheer, M. Hajmeer, Artificial neural networks: fundamentals, computing, design, and application, J. Microbiol. Methods 43 (1) (2000) 3–31.
[6] S. Yin, S. Ding, X. Xie, H. Luo, A review on basic data-driven approaches for industrial process monitoring, IEEE Trans. Ind. Electron. 61 (11) (2014) 6418–6428.
[7] W. Du, Z. Zhan, Building decision tree classifier on private data, in: Proceedings of the IEEE International Conference on Privacy, Security and Data Mining, vol. 14, Australian Computer Society, Inc., 2002, pp. 1–8.
[8] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graphical Stat. 15 (2) (2006) 265–286.

[9] J. Braga, Y. Heuze, O. Chabadel, N. Sonan, A. Gueramy, Non-adult dental age assessment: correspondence analysis and linear regression versus Bayesian predictions, Int. J. Legal Med. 119 (5) (2005) 260–274.
[10] E.L. Russell, L.H. Chiang, R.D. Braatz, Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, Chemom. Intell. Lab. Syst. 51 (1) (2000) 81–93.
[11] F.R. Bach, M.I. Jordan, Kernel independent component analysis, J. Mach. Learn. Res. 3 (2003) 1–48.
[12] S. Yin, X. Zhu, O. Kaynak, Improved pls focused on key performance indictor related fault diagnosis, IEEE Trans. Ind. Electron. (2014).
[13] W.-H. Chen, S.-H. Hsu, H.-P. Shen, Application of svm and ann for intrusion detection, Comput. Oper. Res. 32 (10) (2005) 2617–2634.
[14] S. Yin, G. Wang, H.R. Karimi, Data-driven design of robust fault detection system for wind turbines, Mechatronics 24 (4) (2014) 298–306.
[15] S. Ding, S. Yin, K. Peng, H. Hao, B. Shen, A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill 9(4) (2012) 2239–2247.
[16] L.M. Elshenawy, S. Yin, A.S. Naik, S.X. Ding, Efficient recursive principal component analysis algorithms for process monitoring, Ind. Eng. Chem. Res. 49 (1) (2009) 252–259.
[17] S.X. Ding, P. Zhang, S. Yin, E.L. Ding, An integrated design framework of fault-tolerant wireless networked control systems for industrial automatic control applications, IEEE Trans. Ind. Inform. 9 (1) (2013) 462–471.
[18] S. Yin, X. Yang, H.R. Karimi, Data-driven adaptive observer for fault diagnosis, Math. Probl. Eng. 2012 (2012).
[19] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, Improvements to Platt's smo algorithm for svm classifier design, Neural Comput. 13 (3) (2001) 637–649.
[20] V. Koltchinskii, Rademacher penalties and structural risk minimization, IEEE Trans. Inf. Theory 47 (5) (2001) 1902–1914.
[21] S. Yin, G. Wang, X. Yang, Robust pls approach for kpi-related prediction and diagnosis against outliers and missing data, Int. J. Syst. Sci., no. (2014) 1–8, ahead-of-print.
[22] J. Bedo, C. Sanderson, A. Kowalczyk, An efficient alternative to svm based recursive feature elimination with applications in natural language processing and bioinformatics, in: AI 2006: Advances in Artificial Intelligence, Springer, 2006, pp. 170–180.
[23] S. Yin, X. Li, H. Gao, O. Kaynak, Data-based techniques focused on modern industry: an overview, 2014.
[24] S. Yin, X. Zhu, Fault detection based on a robust one class support vector machines, Neurocomputing 145 (5) (2014) 263–268.
[25] V. Cherkassky, Y. Ma, Practical selection of svm parameters and noise estimation for svm regression, Neural Netw. 17 (1) (2004) 113–126.
[26] S. Yin, X. Zhu, H.R. Karimi, Quality evaluation based on multivariate statistical methods, Math. Probl. Eng. 2013 (2013).
[27] Y. Ding, D. Wilkins, Improving the performance of svm-rfe to select genes in microarray data, BMC Bioinform. 7 (Suppl 2) (2006) S12.
[28] A.S. Naik, S. Yin, S.X. Ding, P. Zhang, Recursive identification algorithms to design fault detection systems, J. Process Control 20 (8) (2010) 957–965.
[29] H. Dong, Z. Wang, J. Lam, H. Gao, Fuzzy-model-based robust fault detection with stochastic mixed time delays and successive packet dropouts, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42 (2) (2012) 365–376.
[30] F. Busetti, Genetic Algorithms Overview, Retrieved on December, vol. 1, 2007.
[31] S. Yin, X. Gao, H.R. Karimi, X. Zhu, Study on support vector machine-based fault detection in tennessee eastman process, in: Abstract and Applied Analysis, vol. 2014, Hindawi Publishing Corporation, 2014.
[32] R. Leardi, Application of genetic algorithm—pls for feature selection in spectral data sets, J. Chemom. 14 (5–6) (2000) 643–655.
[33] J. Kennedy, Particle swarm optimization, in: Encyclopedia of Machine Learning, Springer, 2010, pp. 760–766.
[34] S. Yin, S. Ding, X. Xie, H. Luo, A review on basic data-driven approaches for industrial process monitoring, IEEE Trans. Ind. Electron. 61 (11) (2014) 6418–6428.

Yang Tian was born in Shanxi, China, in 1993. He is currently an undergraduate student with State Key Laboratory of Robotics and Systems (HIT) in Harbin Institute of Technology. His research interests include the application of learning machines such as neural networks and support vector machines, focusing on improvements by means of data preprocessing algorithms.

Mengyu Fu, born in 1992, now is studying as a bachelor candidate in Honors School, Harbin Institute of Technology. She is currently doing research with State Key Laboratory of Robotics and Systems (HIT). Her researches mainly include datamining, machine learning and the corresponding variants or combinations of basic algorithms.

Fang Wu received his B.E. degree in automation from Harbin Engineering University, Harbin, China. He is currently working toward the master degree in control science and engineering with the Key Laboratory of Robotics and Systems in Harbin Institute of Technology. His research interests include fault detection and diagnosis, and machine learning.