Measurement 47 (2014) 715–723

Contents lists available at ScienceDirect

Measurement

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

A new approach for rule extraction of expert system based on SVM

Ai Li, Guo Chen *

College of Civil Aviation, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, PR China

ARTICLE INFO

Article history:
Received 18 October 2012
Received in revised form 9 August 2013
Accepted 19 August 2013
Available online 14 September 2013

Keywords:
Support Vector Clustering
Support Vector Machine
Rule extraction
Knowledge acquisition
Expert system
Genetic Algorithm
Feature selection

ABSTRACT

Based on the SVM's excellent generalization performance, a new approach is proposed to extract knowledge rules from Support Vector Clustering (SVC). In this method, the first step is to choose the features of the sample data by using Genetic Algorithm for improving the comprehensibility of the knowledge rules. Then the SVC algorithm is adopted to obtain the Clustering Distribution Matrix of the sample data whose features have been chosen. Finally, hyper-rectangle rules are constructed using the Clustering Distribution Matrix. To make the rules more concise, and easier to explain, hyper-rectangle rules are simplified further by using rules combinations, dimension reduction and interval extension. In addition, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm is adopted to resample fault samples in order to solve the serious imbalance problem of samples. The UCI datasets are used to validate the new method proposed in this paper, the results compared with other rules extraction methods show that the new approach is more effective. The new method is used to extract knowledge rules for aero-engine oil monitoring expert system, and the results show that the new method can effectively extract knowledge rules for expert system, and break through the bottleneck in expert system knowledge dynamic acquisition.

1. Introduction

At present, knowledge acquisition through data mining [1,2] occurs mainly through machine learning or statistics. Correlation analyses [3], artificial neural networks [4], rough sets [5], and decision trees [6] are extensively employed for data mining. If data mining is applied to an expert system and if the knowledge rules are extracted automatically from real data, then the intelligence level and knowledge acquisition ability of the expert system will be greatly improved.

In recent years, the Support Vector Machine (SVM) [7] has become an emerging classification technology in data mining. The SVM can approximate any continuous bounded nonlinear function because of the perfect generalization theory and strong nonlinear mapping ability. The SVM has several advantages over the neural network, such as better generalization ability, no local minimum problem, the ability to automatically construct the learning machine, no dimension curse, and the ability to deal with small samples. These advantages have caused data mining technology based on SVM to receive the attention of researchers worldwide. Furthermore, a number of promising SVM rule extraction algorithms published to date [8–14] are not only simple but also broadly applicable. Nunez et al. [9] introduced a rule extraction approach based on the SVM, in which K-means clustering is used to obtain clustering centers, which are then combined with support vectors (SVs) to define ellipsoid rules. Finally, the "if-then" rules can be obtained when the ellipsoid rules are mapped to the input space. However, the generated ellipsoid rules seriously overlap. In addition, the solution quality of K-means strongly depends on the initial values for the centers, and it is difficult to control the quantity and quality

* Corresponding author. Tel./fax: +86 025 84891850.
E-mail address: cgzyx@263.net (G. Chen).

0263-2241/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.measurement.2013.08.028

of the obtained rules. In a similar study, Zhang et al. [10] introduced the hyper-rectangle rule extraction (HRE) algorithm to extract rules from the trained SVM. The authors used the Support Vector Clustering (SVC) algorithm to find prototype vectors for each class, and then used those vectors with the SVs to generate hyper-rectangles. A nested generalized exemplar algorithm is utilized to first construct small hyper-rectangles around the prototypes, which are then grown incrementally until the stopping criteria based on a user-defined minimum confidence threshold (MCT) or minimum support threshold (MST) are met. If-then rules are then generated by projecting these hyper-rectangles onto coordinate axes. The published results for this method show that the rules provide good accuracy. However, all the features are present as antecedents of these rules. This limits their explanation capability, since no indication is given about the most important features for the classification.

Based on the aforementioned limitations, here, a new method is proposed to extract knowledge rules from SVC. The first step in this method is to choose the features of the sample dataset using a Genetic Algorithm (GA) for improving the comprehensibility of the knowledge rules. The next step is to map the chosen features of the training samples into a high-dimensional feature space to get optimal separating hyper-planes and SVs. Finally, the hyper-rectangles are constructed using the Clustering Distribution Matrix of the data obtained by the SVC, and the if-then rules are generated by projecting these hyper-rectangles onto coordinate axes. In order to make the rules more concise and easier to explain, hyper-rectangle rules are further simplified using a combination of rules, dimension reduction, and interval extension. In addition, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm is adopted to resample fault samples in order to solve the serious imbalance problem of samples. Experimental results show that it is easy to control the number and the support degree of the generated rules; feature selection and simplification of rules can greatly improve their explanation capability.

Spectral oil diagnosis expert system is the advanced stage of aero-engine wear fault diagnosis. At present, some oil monitoring expert systems have been developed, such as, the advanced rapid analysis system PFALink developed by the United States Mobil oil company, lubricating oil analysis expert system Lube Analyst and Atlas developed by the United States and Canada. But these software only provides a framework and management system, and the users need to develop its core knowledge base and provide the monitored wear element threshold value. In the intelligent diagnosis expert system, these problems, such as weak knowledge acquisition, hard knowledge updating and poor knowledge adaptability, still did not get effective to be overcome. The expert system knowledge acquisition is basically by means of the mechanical learning methods based on the experiences. The knowledge is hard to update and the rules exist serious problems such as inconsistent, redundancy, and combination explosion. Therefore, in this paper, the new method is applied to the knowledge acquisition of aero-engine spectral oil diagnosis expert system. Experimental results to real dataset show the effectiveness and the correctness of the new method.

2. Knowledge rules extracting method based on GA_SVC

The rule extraction process includes data preprocessing, SVC, hyper-rectangle rule extraction and rule simplification. The entire rule extraction procedure is shown in Fig. 1.

2.1. Data preprocessing

2.1.1. Dalancing to unbalance data

In data mining experiments, the datasets are usually assumed to balance distribution, which is the number of various types of samples is almost the same, while it is almost non-existent in the real. In many real datasets, the number of class with different label is unequal. These datasets are called unbalanced datasets. Usually, the minority class samples will be taken out as noise so that no rules about the minority class can be extracted. Therefore, in order to extract rules of various types of samples completely, and improve the recognition rate of the rules, the first step is to preprocess the unbalance data into balance data before rules extraction.

In this paper, we resample fault samples by using Synthetic Minority Over-sampling Technique (SMOTE) which is the typical sampling algorithm. SMOTE [15] algorithm is an over-sampling method put forward by Chawla. In order to make the dataset be equilibrium, the main concept of the method is to use k neighbor method and linear interpolation method to insert new samples according to certain rules between the two closer samples of minority class. In Fig. 2, a two dimensional example $\{\mathbf{X} = (x_1, x_2)\}$ is enlarged by using SMOTE over-sampling method. It can be seen from Fig. 2 that the new re-sampling samples focus

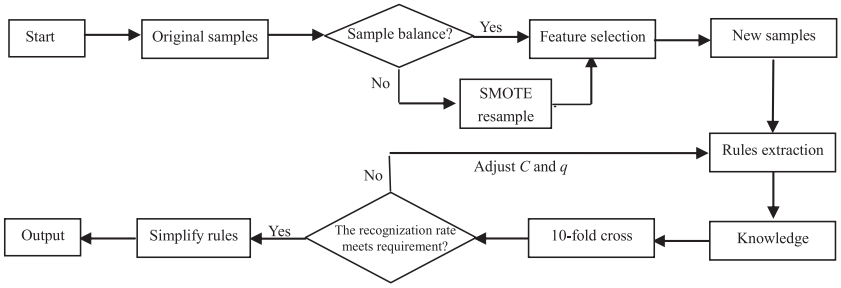

Fig. 1. Rules extraction procedure.

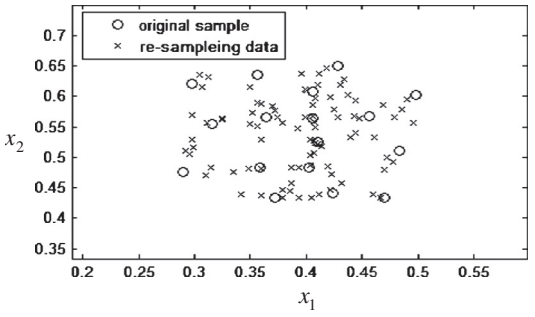

Fig. 2. SMOTE re-samples results.

around the original sample, and reflect the sample potential distribution very well.

2.1.2. Feature selection based on Genetic Algorithm

In this paper, of the fealy using SMOTE over-sampling methodIn order to reduce the sample feature dimensions, improve rules extraction efficiency, and enhance the rules comprehension, the feature selection is a key step. In this paper, a feature selection method based on the Genetic Algorithm (GA) is proposed because of its implicit parallelism and global search ability. The basic principle of feature selection is to use GA to search an optimal binary code, each of the code corresponding to a feature selection result. If the *i*th bit is 1, the corresponding feature will be selected, and this feature will appear in the classifier, but if the *i*th bit is 0, the corresponding feature will not be selected, and this feature will not appear in the classifier.

With the problem of the feature selection, the fitness function is very important, and it is constructed mainly based on class separability criterion and the feature classification ability. The effectiveness of the fitness function will directly determine the search direction and evolution results of GA. In this paper, the fitness function based on *k*-nearest neighbor classification is constructed. The neighbor method is a kind of nonparametric pattern recognition method, which belongs to supervised learning, and its classification ability can be used as the characteristic evaluation function.

2.2. Support Vector Clustering (SVC)

SVC is a novel clustering method proposed by Ben-Hur et al. [16], in which data points are mapped to a high dimensional feature space by means of a Gaussian kernel, where we search for the minimal enclosing sphere. When this sphere is mapped back to the data space, it can be separated into several components, each enclosing a separate cluster of points.

2.2.1. Mathematical description of the clustering

Let $\{x_i\} \subseteq X$ be a dataset that includes n points, with $X \subseteq R^d$ as the data space. Using a nonlinear transformation φ from x_j to a high dimensional feature-space, we look for the smallest enclosing sphere of radius R. This is described by the constraints:

$$\|\Phi(x_j) - a\|^2 \leqslant R^2 \quad \forall j \tag{1}$$

where $\|\bullet\|$ is the Euclidean distance and a is the center of the sphere. In order to allow some points outside the sphere, soft constraints are used by adding slack variables ζ_j, namely:

$$\|\Phi(x_j) - a\|^2 \leqslant R^2 + \zeta_j \tag{2}$$

The mathematical descriptions for the above problems are:

$$\min R^2 + C\sum_j \zeta_j \tag{3}$$

$$\text{s.t.} \|\varphi(x_j) - a\|^2 \leqslant R^2 + \zeta_j \tag{4}$$

$$\zeta_j \geqslant 0 \tag{5}$$

where C is a penalty term. The bigger the C value, the less is allowed for the emergence of the noise. To solve this problem, transform into the Wolfe dual form:

$$\max W = \sum_j \Phi(x_j)^2 \beta_j - \sum_{i,j} \beta_i \beta_j \Phi(x_i) \cdot \Phi(x_j) \tag{6}$$

with the constraints:

$$\sum_j \beta_j = 1 \tag{7}$$

$$0 \leqslant \beta_j \leqslant C \quad j = 1, 2, 3 \cdots N \tag{8}$$

We adopted the SV method and represented the dot products $\Phi(x_i)\Phi(x_j)$ by an appropriate Mercer kernel $K(x_i, x_j)$. We used the following Gaussian kernel:

$$K(x_i, x_j) = \exp\left(\|x_i - x_j\|^2 / q^2\right) \tag{9}$$

The Lagrangian W can then be written as:

$$W = \sum_i K(x_j, x_j)\beta_j - \sum_{i,j} \beta_i \beta_j K(x_i, x_j) \tag{10}$$

According to Eqs. (7), (8), and (10), Eq. (6) can be written as:

$$\max W = 1 - \sum_{i,j} \beta_i \beta_j K(x_i, x_j) \tag{11}$$

$$\text{Namely,} \ \min W' = \sum_{i,j} \beta_i \beta_j K(x_i x_j) \tag{12}$$

$$\text{St}: \ 0 \leqslant \beta_j \leqslant C \quad j = 1, 2, 3 \cdots N \tag{13}$$

Finally, the solution of β_j can be obtained.

The distance between a data sample and the center of the feature space hyper-sphere is computed as:

$$D(x_i) = \sqrt{\sum_{i,j=1}^{N} \beta_i \beta_j K(x_i, x_j) + K(x_i, x_i) - 2\sum_{j=1}^{N} K(x_j, x_i)\beta_j} \tag{14}$$

The radius of the smallest enclosing hyper-sphere in the feature space is determined by $R = D(x_i) | \forall 0 < \beta_i < C$.

Further, the contours that enclose the points in the input space are defined as the set $\Omega = \{x|D(x) = R\}$.

2.2.2. Cluster assignment

The cluster description algorithm does not differentiate between points that belong to different clusters. To do this, we used a geometric approach involving $D(x)$, based on the following observation: given a pair of data points that belong to different components (clusters), any path that connects them must exit from the sphere in feature space. Therefore, such a path contains a segment of points y such that $D(y) > R$. This leads to the definition of the adjacency matrix A_{ij} between the pairs of points x_i and x_j whose images lie in or on the sphere in the feature space:

$$A_{ij} = \begin{cases} 1 & \text{if, for all } y \text{ on the line segment connecting } x_i \text{ and } x_j D(y) \leqslant R \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Clusters are now defined as the connected components of the graph made by A. The line segment is checked by sampling a number of points (20 points were used in our numerical experiments).

Because SVs lie on cluster boundaries that determine the shape and size of the hyper-sphere in the feature space, so in this paper, only the SVs are assigned to the cluster and others are left.

2.2.3. Analysis of SVC parameters

There are two important parameters in the SVC algorithm, one is the scale parameter q of the Gaussian kernel, and the other is the penalty factor C. The parameter q determines the number of the clusters and the penalty factor C can be determined by setting a priori maximum permitted rejection rates of the error on the clusters. We consider two simulation samples as examples to illustrate the influence of these two parameters on the clustering results.

For the first sample (as shown in Fig. 3), when q was small, there was only one cluster. When q increased, the clustering boundary fit the data more tightly and splits into more clusters. For the second sample (as shown in Fig. 4), when parameter C was larger than 1, all of the samples fit into the generated cluster including some noise elements and outliers. When parameter C was equal to 1, there were two outliers that were excluded from the cluster, and the cluster fit more tightly to the samples.

In conclusion, the parameter q of the Gaussian kernel determined the number of clusters (i.e. the number of the hyper-rectangle rules); and the penalty factor C determined the size of the clusters (i.e. the size of the hyper-rectangle rules). The choice of an appropriate parameter C can avoid the over-fitting of the rules because of the isolated points and noise elements.

2.3. Hyper-rectangle rule

The nature of the hyper-rectangle rule extraction method based on SVC clustering is based on the principle of constructing a hyper-rectangle that covers the input space based on the SVC classification hyper-plane. Each

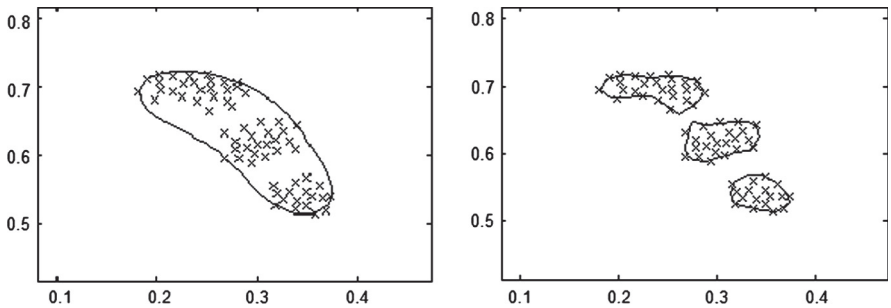

Fig. 3. Effect of the scale parameter q on the number of the clusters.

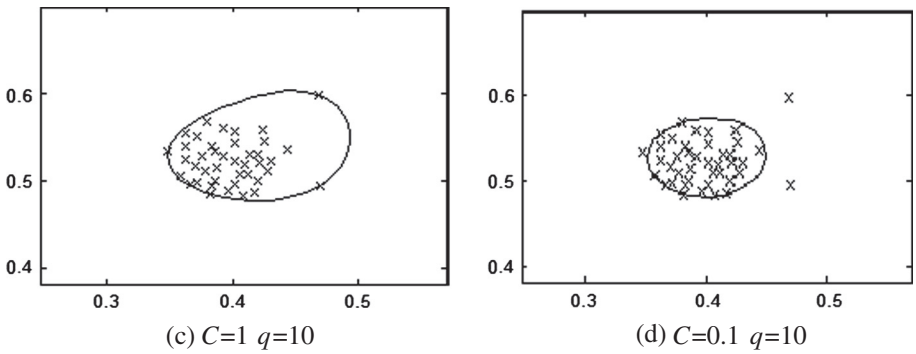

Fig. 4. Effect of the parameter C on cluster compactness.

hyper- rectangle H^{j,L_j} is defined by the interval of all its attributes $x_1 \in [x_1^L, x_1^U] \cap \cdots \cap x_i \in [x_i^L, x_i^U] \cap \cdots \cap x_N \in [x_N^L, x_N^U]$, where L_j is the class label and x_i^L and x_i^U from the range of component i of sample x. When H^{j,L_j} is projected onto coordinate axes, the following if-then rules below are obtained.

$$R^{j,L_j}: \text{ if } x_1 \in [x_1^L, x_1^U] \cap \cdots \cap x_i \in [x_i^L, x_i^U] \cap \cdots \cap x_N \in [x_N^L, x_N^U] \text{ then class} L_j \tag{16}$$

Each discovered rule should have a measure of certainty that assesses the validity of the rule. There are two objective measures, one is the confidence that acts as the measure of reliability or accuracy, and it represents the strength or quality of a rule; the other is the support degree which represents the percentage of data samples that satisfy the extracted rule. In order to control the number and the validity of the hyper-rectangles, we defined the support and the confidence of H^{j,L_j} as follows:

$$\text{conf.}(R^{j,L_j}) = \frac{\text{sample covered by } H^{j,L_j} \text{ with class label } L_j}{\text{sample covered by } H^{j,L_j}} \tag{17}$$

$$\text{supp.}(R^{j,L_j}) = \frac{\text{sample covered by } H^{j,L_j} \text{ with class label } L_j}{\text{sample with class label } L_j} \tag{18}$$

Rules that satisfy both a user-specified minimum confidence threshold (MCT) and minimum support threshold (MST) are called as the strong association rules, and are considered as interesting rules. On the contrary, rules with low support probably represent noises, or exceptional cases.

2.4. Rule-based recognition methods

2.4.1. Distance method

The distance method entails that the test example will be recognized as the class labeled by the hyper-rectangle rule that is closest to the test example. Each hyper-rectangle H^{j,L_j} with class label L_j is represented by its lower-left corner H^{j,L_j}_{lower} and upper right corner H^{j,L_j}_{upper}. The distance between H^{j,L_j} and a sample $X = (x_1, \cdots, x_N)$ is defined as follows:

$$D\left(X, H^{j,L_j}\right) = \sqrt{\sum_{i=1}^{N}\left(w_{fi} \times \left(d_i\left(X, H^{j,L_j}\right)\right)^2\right)} \tag{19}$$

where w_{fi} is the weight of the ith feature. In general, the w_{fi} is equal to 1, and

$$d_i(x, H^{j,L_j}) = \begin{cases} x_i - H^{j,L_j}_{\text{upper},i} & \text{if} \quad x_i > H^{j,L_j}_{\text{upper},i} \\ H^{j,L_j}_{\text{lower},i} - x_i & \text{if} \quad x_i < H^{j,L_j}_{\text{lower},i} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where $H^{j,L_j}_{\text{lower},i}$ is the ith element of H^{j,L_j}_{lower}.

2.4.2. Range method

The range method entails that the sample is recognized directly according to the range of the rules. For each hyper-rectangle H^{j,L_j}, its interval is $[x_{j1}^L, x_{j1}^U] \cap \cdots \cap [x_{ji}^L, x_{ji}^U] \cap \cdots \cap [x_{jN}^L, x_{jN}^U]$, where x_{ji}^L and x_{ji}^U from the range of the ith component of the sample x belonging to the jth class. For the sample $X = (x_1, \cdots, x_N)$;, if $x_1 \in [x_{j1}^L, x_{j1}^U] \cap \cdots \cap x_i \in [x_{ji}^L, x_{ji}^U] \cap \cdots \cap x_N \in [x_{jN}^L, x_{jN}^U]$, then the sample is recognized as the jth class.

2.5. Rule simplification

In order to make the rules more concise and easier to explain, the extraction rules should be simplified further by using such means as rules combination, dimension reduction and interval extension.

2.5.1. Rules combination

When the distribution of the sample is very complex, the number of rules will be very large, making them very difficult to understand. In order to make the rules easier to explain, the extracted rules should be simplified further by combining the hyper rectangle rules with closer distance and smaller supports into rules with bigger support. In this process, the threshold is used to limit the minimum confidence of the generated rules. In the process of combining rules, the overlapping degree of two different hyper rectangles with the same class label is measured from the area of the two overlapping hyper-rectangles. The bigger the area, the smaller is the distance.

For example, as shown in Fig. 5, according to the nearest neighbor strategy, while $H^{2,1}$ is the nearest neighbor hyper-rectangle of $H^{1,1}$ the two smaller hyper-rectangles $H^{1,1}$ and $H^{2,1}$ are combined into a larger hyper-rectangle $H^{a,1}$, and then $H^{a,1}$ is combined with its nearest neighbor $H^{3,1}$. This process is repeated until there are no more hyper-rectangles that can be combined. In the process of the rules combinations, if the confidence of every new generated hyper-rectangles is less than the given minimum confidence threshold MCT, the combination will be cancelled. The combination process is shown in Fig. 5.

2.5.2. Rules reduction

For if-then rules, the more the attributes of a rule is, the more difficult it is to understand. The attributes are selected by using GA, in order to further enhance the comprehensibility of the rules, however, another approach is to use rule reduction, which is done in the following two steps: interval extension and dimension reduction. entails conversion of the closed interval of the attribute to an open interval; dimension reduction entails eliminating one-

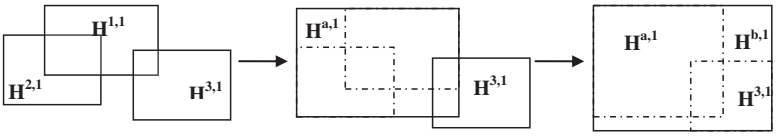

Fig. 5. Hyper rectangle combination process.

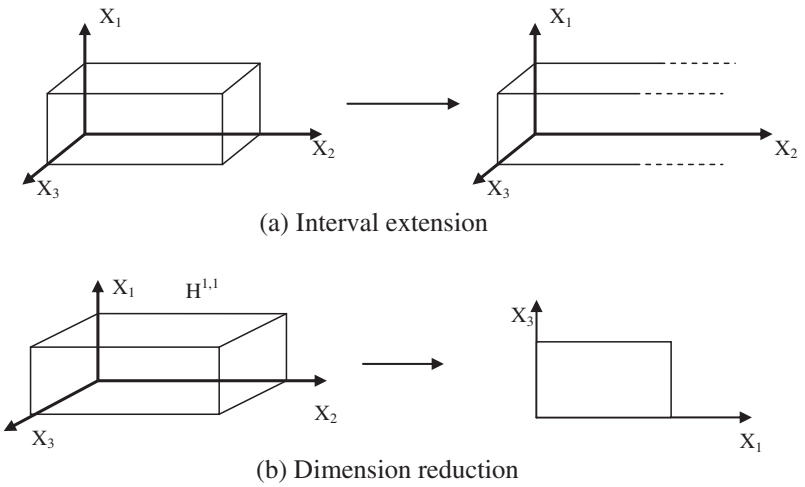

Fig. 6. Hyper-rectangle rules reduction.

dimensional attributes from the rule antecedent. After rule reduction, the rule becomes easier to understand and is more representative. If the confidence of the rules after reduction is within than the MCT, the reduction process will be cancelled.

The reduction process of the hyper-rectangle $H^{1,1}$ in a three dimensional space is shown in Fig. 6. After interval extension, the closed interval of attribute x_2 is broken, and the corresponding if-then rules become as follows.

$$R^{1,1}: \ x_1 \in [x_{1L}, x_{1U}] \cap x_2 > x_{2L} \cap x_3 \in [x_{3L}, x_{3U}] \tag{21}$$

If $conf\,(R^{1,1})$ is greater than the MCT value even after eliminating the attribute x_2 from the rule $R^{1,1}$, then the rule $R^{1,1}$ can be re-written as

$$R^{1,1}: \ x_1 \in [x_{1L}, x_{1U}] \cap x_3 \in [x_{3L}, x_{3U}] \tag{22}$$

If the confidence of the new rules is less than the MCT, the process will be cancelled.

3. Verification analysis based on the UCI data

Six datasets from the UCI machine learning database were used to verify the effectiveness of our proposed method. The datasets and their features are listed in Table 1, and the optimal feature combinations are also listed which is obtained by feature selection method based on the GA. The SVC is the rule extraction method from SVC without feature selection, and the GA_SVC is the rule extraction method from SVC with feature selection, as shown in Table 2. From Table 2, it can be seen clearly that the recognition rate of rules extracted from GA_SVC method is much higher than the one from SVC method, which fully shows the importance of the GA feature selection.

The proposed new method, the C4.5 decision tree method and the BayesNet method were compared using 10-fold cross-validation, and the results are presented in Table 3. From Table 3, it can be seen that the proposed method had the highest accuracy for Act, Iris and Hepatitis data sets, which demonstrates its superior generalization performance.

Iris is a classical pattern classification dataset, which is often used to evaluate the performance of the new algorithms. The selected features are presented in Table 1 and the MCT and MST were set as 0.8 and 0.1 respectively. Some hyper-rectangle rules for the Iris dataset with different parameters C and q are given in Tables 4–6. We can see that when $q = 10$, there are only three rules which have larger support. With increasing q, more rules are obtained and they have smaller support. As discussed above, the smaller the parameter C is, the less support of the rule has. From Table 5 we can see that support values of the three rules are respectively reduced to 0.72, 0.63 and 0.67.

Finally, the rules with the highest recognition rate (as shown in Table 5) are further simplified. The first step is rule combination; the results of this combination are listed in Table 7. The next step is the rules reduction, including

Table 1
Datasets and their features.

Data set	Number of training samples	Number of test samples	Number of features	The optimal attributes combination	Fitness
ACT	126	14	6	1 1 0 1 0 1	0.49
Ecoli	302	34	7	1 0 0 0 1 1 0	0.37
Iris	135	15	4	0 0 1 1	0.69
Glass	193	21	9	0 0 1 1 0 0 0 0 0	0.24
Hepatitis	139	16	19	0000010010110110010	0.44
Wine	160	18	13	0000010001000	0.71

Table 2
The effect of Genetic Algorithm feature selection on recognition rate.

	ACT	Ecoli	Iris	Glass	Hepatitis	Wine
SVC	0.8643	0.67	0.82	0.4975	0.7599	0.8733
GA_SVC	0.8929	0.7581	0.98	0.5111	0.8667	0.9375

Table 3
Comparison of recognition results.

	ACT	Ecoli	Iris	Glass	Hepatitis	Wine
GA_SVC	**0.8929**	0.7581	**0.98**	0.5111	**0.8667**	0.9375
C4.5	0.8857	**0.8423**	0.96	0.6589	0.8129	0.9382
BayesNet	0.8571	0.8125	0.9267	**0.7477**	0.8323	**0.9888**

interval extension and dimension reduction, the results are presented in Table 8. From Tables 7 and 8, we can see that the combined and reduced rules are more concise and easier to understand. In addition, the recognition rate can also increase to the extent.

Table 8
Rules reduction results.

Hyper-rectangle rules	[supp., conf.]	According to the range method
1. If $x_4 \leq 0.6$ then *class* 1	[0.91, 1]	0.9867
2. If $x_4 \in [1, 1.6]$ then *class* 2	[0.8, 0.97]	
3. If $x_3 \geqslant 4.8$then *class* 3	[0.78, 0.97]	

4. Aero-engine spectral oil diagnosis knowledge rules extraction

Spectral oil analysis is an important mean of aero-engine wear fault diagnosis, and the expert system is the effective way to implement the diagnosis. At present, the expert system knowledge acquisition is mainly based on human experts experience knowledge, and very difficult to realize the automatic acquisition. Therefore, the realiza-

Table 4
Rules for Iris dataset with $C = 0.5$ and $q = 10$.

Hyper-rectangle rules	[supp., conf.]	According to the distance method	According to the range method
1. If $x_3 \in [1.2, 1.9] x_4 \in [0.1, 0.5]$ then *class* 1	[0.91, 0.97]	0.9533	0.88
2. If $x_3 \in [3.5, 5.0] x_4 \in [1, 1.8]$ then *class* 2	[0.76, 0.94]		
3. If $x_3 \in [4.8, 6.3] x_4 \in [1.5, 2.5]$ then *class* 3	[0.78, 0.91]		

Table 5
Rules for Iris dataset with $C = 0.5$ and $q = 20$.

Hyper-rectangle rules	[supp., conf.]	According to the distance method	According to the range method
1. If $x_3 \in [1.3, 1.9] x_4 \in [0.1, 0.6]$ then *class* 1	[0.91, 1]	0.98	0.8513
2. If $x_3 \in [3.7, 4.9] x_4 \in [1, 1.7]$ then *class* 2	[0.78, 0.97]		
3. If $x_3 \in [4.8, 5.3] x_4 \in [1.5, 2]$ then *class* 3	[0.31, 0.82]		
5. If $x_3 \in [5.3, 6.1] x_4 \in [2.1, 2.5]$ then *class* 3	[0.2, 1]		

Table 6
Rules for Iris dataset with $C = 0.1$ and $q = 10$.

Hyper-rectangle rules	[supp., conf.]	According to the distance method	According to the range method
1. If $x_3 \in [1.3, 1.7] x_4 \in [0.1, 0.5]$ then *class* 1	[0.72, 0.99]	0.9733	0.8267
2. If $x_3 \in [3.5, 4.9] x_4 \in [1, 1.6]$ then *class* 2	[0.63, 1]		
3. If $x_3 \in [4.8, 6.0] x_4 \in [1.5, 2.5]$ then *class* 3	[0.67, 1]		

Table 7
Rules combination results.

Hyper-rectangle rules	[supp., conf.]	According to the distance method	According to the range method
1. If $x_3 \in [1.3, 1.9] x_4 \in [0.1, 0.6]$ then *class* 1	[0.91, 1]	0.98	0.8767
2. If $x_3 \in [3.7, 4.9] x_4 \in [1, 1.7]$ then *class* 2	[0.82, 0.97]		
3. If $x_3 \in [4.8, 6.1] x_4 \in [1.5, 2.5]$ then *class* 3	[0.78, 0.97]		

tion of the automatic acquisition of expert system knowledge is especially important to the expert system fault diagnosis. The main means to realize knowledge acquisition is to use data mining method to extract knowledge rules automatically from a large number of data.

In view of this, the new method proposed in this paper is applied to extract knowledge of actual aero-engine spectral oil data. Taking a military aero-engine spectral oil data as an example, the dataset contains 237 samples of 10 aero-engines in normal condition and wear condition. A part of the dataset is listed in Table 9. The 7 kinds of elements (Fe, Al, Cu, Cr, Ag, Ti and Mg) contents respectively corresponds to the A1–A7. Wear state F is divided into three classes: 1—normal; 2—bearing wear and 3—bearing wear and cage fracture. Wear state F is taken as decisions attribute *D* of the example.

In the 237 samples, the number of samples with class "1" is 230, the number of samples with class "2" is only five, and the number of samples with class "3" is the least, only two, therefore the number of fault samples is especially few, leading to sample serious imbalance. Therefore the typical SMOTE algorithm of re-sampling algorithms is used to resample fault samples.

After re-sampling, the numbers of samples with class "2" and "3" are all expanded to 100. The first step is to select features by using GA, and the code of the optimal attributes combination is: 1100100, fitness is 0.78, which means the Fe, Al and Ag three elements are selected to extract rules. In the SVC training, the penalty factor *C* and the Gaussian Kernel parameter *q* are got by using 10-fold cross validation method, finally, the optimal *C* and *q* are obtained, and they are respectively 0.5 and 2, the MCT and MST are set 0.9 and 0.1 respectively. Table 10 lists the extraction rules from GA_SVC and average recognition rate by 10-fold cross validation with the two recognition methods discussed in section 1.5. In both cases, the recognition

Table 9
A part of spectral oil data.

Fe(A1)	Al(A2)	Cu(A3)	Cr(A4)	Ag(A5)	Ti(A6)	Mg(A7)	F(D)
0.50	0.00	0.30	0.00	0.10	0.50	2.00	1
1.60	0.00	0.60	0.00	0.10	0.60	2.90	1
2.60	0.00	0.90	0.20	0.20	0.70	3.50	1
2.30	0.00	0.60	0.10	0.20	0.50	4.80	1
2.60	0.00	0.60	0.20	0.20	0.60	4.40	1
15.60	0.50	2.40	1.40	0.50	1.10	7.20	2
3.20	0.00	0.70	0.30	0.20	0.70	5.10	1
4.80	0.00	1.50	0.20	0.10	1.00	6.10	1
23.90	1.80	9.80	1.10	1.80	1.90	9.30	3

Table 10
Rules for spectral data set from GA_SVC.

Hyper-rectangle rules	[supp., conf.]	According to the distance method	According to the range method
1. If Fe ∈ [0.2, 5.9]Al ∈ [0, 0.9]Ag ∈ [0, 1] then *class* 1	[1, 1]	0.975	0.9175
2. If Fe ∈ [8.1, 11.8]Al ∈ [0, 0.7]Ag ∈ [0.5, 0.9] then *class* 2	[0.58, 1]		
2. If Fe ∈ [12.7, 15.6]Al ∈ [0.26, 0.6]Ag ∈ [0.5, 0.89] then *class* 2	[0.21, 1]		
3. If Fe ∈ [19.02, 23.9]Al ∈ [0.6, 2.21]Ag ∈ [0.4, 1.8] then *class* 3	[1, 0.97]		

Table 11
Rules combination results.

Hyper-rectangle rules	[supp., conf.]	According to the distance method	According to the range method
1. If Fe ∈ [0.2, 5.9]Al ∈ [0, 0.9]Ag ∈ [0, 1] then *class* 1	[1, 1]	0.975	0.9175
2. If Fe ∈ [8.1, 15.6]Al ∈ [0, 0.7]Ag ∈ [0.5, 0.9]then *class* 2	[0.79, 1]		
3. If Fe ∈ [19.02, 23.9]Al ∈ [0.6, 2.21]Ag ∈ [0.4, 1.8]then *class* 3	[1, 0.97]		

Table 12
Rules reduction results.

Hyper rectangle rules	[supp., conf.]	According to the range method
1. If Fe <= 5.9 then Normal	[1, 1]	0.97
2. If Fe >=8.1 and Fe<=15.6 then Axial bearing wear	[0.79, 1]	
3. If Fe >= 19.02 then axial bearing wear and cage fracture	[1, 0.97]	

rates are all more than 90%, which show that the extracted rules have very good quality.

The combined rules, the reduced rules and the recognition rate are listed in Tables 11 and 12 respectively. By comparing the Tables 10–12, we can see that the combined and reduced rules have better understand ability and interpretability, and more conducive to engineering application.

5. Conclusions

(1) A new approach of rule extraction from Support Vector Machine is proposed in this paper. In this method, the first step is to select the feature of the sample data by using Genetic Algorithm. Then SVC algorithm is adopted to get the Clustering Distribution Matrix of the sample data. Finally, hyper-rectangle rules are constructed on the base of the Clustering Distribution Matrix. In order to make the rules more concise, easy to be explained, hyper-rectangle rules are simplified further by using rules combination, dimension reduction and interval extension.

(2) The UCI machine datasets are used to test the proposed algorithm and the comparison with the C4.5 decision tree method and the BayesNet method is carried out. The results show that the proposed method has better generalization performance.

(3) Aero-engine spectral oil diagnosis expert systems knowledge acquisition is carried out by using the proposed method. Taking practical aero-engine spectral oil data as an example to extract knowledge rules, verification results show that this method can well satisfy the engineering need, and can be used as an effective tool of aero-engine wear fault diagnosis expert system knowledge rules extraction. It will effectively improve the intelligent level and knowledge acquisition ability of the expert system.

Acknowledgements

Authors are very grateful to the L.Q. Song. and L.B. Chen engineers of Beijing Aeronautical Technology Research Center, who provide the actual aero-engine oil spectral data for this paper. The work is supported by the National Science Foundation of China (Grant No: 61179057).

References

[1] JiaWei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001.
[2] H. Mannila, Data Mining: Machine Learning, Statistics, and Databases, in: Eight International Conference on Scientific and Statistical Database Management, Stockholm, June 18–20, 1996.
[3] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, Fast discovery of association rules. Chapter 12, in: Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 307–328.
[4] L. Fu, Knowledge Discovery Based on Neural Networks, Communications of the ACM 42 (11) (1999) 47–50.
[5] W. Ziarko, Discovery through Rough Set Theory, Communications of the ACM 42 (11) (1999) 55–57.
[6] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106.
[7] V. Vapnik, Statistical Learning Theory, Wiley, New York, NY, 1998.
[8] G. Fung, S. Sandilya, R. Rao, Rule extraction from linear support vector machines, in: Proc. 11th Int'l Conf. Knowledge Discovery and Data Mining, 2005.
[9] H. Nunez, C. Angulo, A. Catala, Rule-extraction from support vector machines, in: Proc. European Symp. Artificial, Neural Networks, 2002, pp. 107–112.
[10] Y. Zhang, H. Su, T. Jia, J. Chu, Rule extraction from trained support vector machines, in: Proc. Ninth Pacific–Asia Conf. Advances in Knowledge Discovery and Data Mining, 2005, pp. 61–70.
[11] X. Fu, C. Ongt, S. Keerthit, G. Hung, L. Goh, Extracting the knowledge embedded in support vector machines, in: Proc. IEEE Int'l Conf. Neural Networks, 2004, pp. 291–296.
[12] Y. Zhang, Z. Li, Y. Tang, K. Cui, DRC-BK: Mining Classification Rules with Help of SVM, LNAI 3056, in: H. Dai, R. Srikant, C. Zhang (Eds.), Springer, 2004, pp.191–195.
[13] N. Barakat, J. Diederich, Learning-based rule-extraction from support vector machines: performance on benchmark data sets, in: N. Kasabov, Z.S.H. Chan (Eds.), Proc. Conf. Neuro-Computing and Evolving Intelligence, 2004.
[14] N. Barakat, J. Diederich, Electric rule-extraction from support vector machines, International Journal of Computational Intelligence 2 (1) (2005) 59–62.
[15] L.B. Jack, A.K. Nandi, Feature selection for ANNs using genetic algorithms in condition monitoring[A], in: ESANN'1999 Proceedings-European Symposium on Artificial Neural Networks[C], Bruges(Belgium), 1999, 313318.
[16] A. Ben-Hui, D. Hom, H.T. Sidgelman, Support vector clustering, Journal of Machine Learning Research (2001) 125–137.