A Support Vector Machines Security Assessment Method Based on Group Decision-marking for Electric Power Information System

Xiaorong Cheng, Yan Wei, Xin Geng
School of Computer Science and Technology
North China Electric Power University
Baoding, China
woaitianjin886@163.com

Abstract—In accordance with the characteristics and the special demands of electric power information system, this paper designs a support vector machines (SVM) risk assessment method based on group decision-marking. According to security technology indices of electric power information system, the method chooses the mode of expert scoring, based on the group decision-marking, to calculate integrated value of each index, which is as a training sample used to train SVM, and it forecasts risk level for the system. Finally, it verifies the correctness of the method by analyzing results of the examples of the electric power information system security assessment. The experiment shows that the method can not only forecast the current risk level of the electric power system with a high accuracy rate, but also reduce the influence of the subjective factors in some degree.

Keywords risk assessment; support vector machines; group decision-marking; electric power information system

I. INTRODUCTION

With the developing of the electric power automation level, the power system depends on the power information network protecting its safety, reliability and efficiency increasingly. Any security problems from this information network are likely to affect power system security, stability and economic operation, thereby affecting the reliable power supply and people’s daily lives [1]. Therefore, in-depth study of security risk assessment methodology for power system information network and formulate a safety measure for electric power system information network is much significant.

Risk assessment analyses maintenance, management, operation of information systems, identifying existing vulnerabilities and the threat of the vulnerability used possibly, assessing whether implement and maintain appropriate security measures, estimating existing risk, influence of risks and possibility of occurrence of risks. And thus managers may choose security measures which can make the risks to the acceptable level. The common risk analysis methods are AHP, Delphi method, Fuzzy Math method and so on [2].

In accordance with the characteristics and the special demands of electric power information system, this paper designs a support vector machines (SVM) risk assessment method based on group decision-marking, as shown in Fig.1. According to security technology indices of electric power information system, the method chooses the mode of expert scoring, based on the group decision-marking, to calculate integrated value of each index as a training sample which is used to train SVM. By training SVM, it divides the power information system into three risk levels, High, Middle, Low, and forecasts risk level for the system by assessment data.

Figure 1. A SVM risk assessment method
II. SUPPORT VECTOR MACHINES

A. SVM Theory

SVM was originally introduced by Vapnik [3] and his colleagues, and it was used to settle the issue of pattern recognize at first. While traditional statistical theory keeps to empirical risk minimization (ERM), SVM satisfies structural risk minimization (SRM) based on statistical learning theory (SLT), whose decision rule could still obtain small error to independent test sampling.

The basic idea of SVM is to map the input data points to a high-dimensional feature space and find a hyper-plane. The algorithm is chosen in such a way to maximize the margin between the separating hyper-plane and data.

For the training dataset:
\[ \{(x_i, y_i)\} \]
where \( x_i \) represents condition attribute and \( y_i \) represents class attribute.

As shown in Fig. 2, SVM optimize the classification boundary by separating the data with the maximal margin hyper-plane. SVM optimize the classification boundary by separating the data with the maximal margin hyper-plane.

Let’s consider a linear instance of a classification problem. Given a data set \( D = \{(x_i, y_i)\} \), the classification function is approximated by the following function:
\[ f(x, a) = \omega \cdot x + b. \]  
(1)

Coefficients \( \omega \) and \( b \) are estimated by minimizing the regularized risk function:
\[ \min_{\omega, b} \frac{1}{2} ||\omega||^2. \]  
(2)

s.t. \( y_i ((\omega \cdot x_i) - b) \geq 1, i = 1, 2, ..., n. \)  
(3)

According to Lagrange principle, the above problem can be transformed to solve its antithesis problem. Finally we can get the decision function as follows:
\[ f(x) = \text{sgn}\left\{\sum_{i=1}^{n} a_i y_i (x_i \cdot x) + b^+\right\}. \]  
(4)

The practical data are usually nonlinear separable. Therefore, Vapnik and Cortes put forward C-SVM in 1995, it induces positive slack variables set \( \xi = (\xi_1, \xi_2, ..., \xi_n) \), which is necessary to allow misclassification. In this way, function (3) can be converted to:
\[ y_i ((\omega \cdot x_i) + b) \geq 1 - \xi_i, i = 1, 2, ..., n. \]  
(5)

When samples match (3), \( \xi_i = 0 \), otherwise \( \xi_i > 0 \), it means that this is a nonlinear point. The C-SVM can be shown as follows:
\[ \max_{a} W(a) = \sum_{i} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j x_i \cdot x_j. \]  
(6)

s.t. \( \sum_{i} y_i a_i = 0, 0 \leq a_i \leq C, i = 1, 2, ..., n. \)  
(7)

where \( C \) is a nonnegative constant that determines the tradeoff between the training error and the model flatness.

Another method to solve the nonlinear question, we can map the input data x to a high-dimensional feature space F via a nonlinear mapping \( \phi \). For computational convenience, introduce the so-called kernel function with this form:
\[ K(x, x_i) = \phi(x)^T \phi(x_i). \] And so, all computations are carried on via kernel function in the input space. We can show the decision-making function as:
\[ f(x) = \text{sgn}\left\{\sum_{i} a_i K(x_i, x) + b\right\}. \]  
(8)

The kernel function in common use can be shown as:
(1) Q-polynomial kernel function:
\[ K(x, x_i) = (x \cdot x_i + 1)^q. \]  
(9)

(2) RBF kernel function:
\[ K(x, x_i) = \exp\left\{-\frac{\|x - x_i\|^2}{\sigma^2}\right\}. \]  
(10)

B. Determining the Parameter

The accuracy of an SVM model is largely dependent on the selection of the model parameters. There are two methods for finding optimal parameter values, a grid search and a cross-validation search [4]. A grid search tries values of each parameter across the specified search range using geometric steps. Grid searches must be evaluated at many points within the grid for each parameter. For example, if a grid search is used with 10 search intervals and an RBF kernel function is used with two parameters, then the model must be evaluated at 10×10=100 grid points. A cross-validation search, firstly divide the training set into k subsets of equal size, sequentially one subset is tested using the classifier trained on the remaining k-1 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified.
III. CONSTRUCTING THE SVM TRAINING SET

A. Risk Index System

Risk assessment for power information system is a pertinence work, and the selected risk indices should reflect the real situation of the enterprise and be representative for highlighting the characteristics of the power system. At the same time, it is necessary to make indices operational which will facilitate the comparisons of data. Hence, in the above guiding principles, and taking power information system’s complexity into full account, this paper chooses security technology indices of electric power system as risk index system, and divides these indices into five items. There are physical security, network security, computer system security, application security, data security, and they are divided into 18 sub-items, as shown in Table 1.

B. Data Pretreatment

Before constructing SVM training set, we need preprocess scores by experts to reduce the influence of the subjective factors. In this paper, for the eighteen risk indices, we adopt the mode of expert scoring, we assign each index weight Wi and expert weight We by using the group decision-making. This can reduce the subjectivity and the negative impact of the decision-making. This can reduce the subjectivity and the negative impact of the difference to some extent. So we get the value Ui of each index, 

\[ U_i = \sum_{j=1}^{m} P_{i,j} \times W_{e,j} \times W_i, \]  

where \( P_{i,j} \) is scores of index i marked by expert j, \( W_{e,j} \) is weight of expert j, \( W_i \) is weight of index i, \( U_i \) is integrated value of index i marked by experts, while m is numbers of experts.

For each index weight \( W_i \), the concept of QoC in variable precision rough set is used to calculate weight, that is, directly mines the importance among indices from decision data to reduce the impact of man-made factors on results [5]. When electric power information system is assessed, the architecture shown in Table 1 is adopted, in which risk indices are mutual independent and index i under item j is recorded as \( P_{ij} \). After each expert marks for every index, the importance of \( P_{ij} \) corresponding to \( D_i \), namely

\[ r^\beta(P_{ij}, D_i) = \frac{\text{card}(\text{POS}^\beta_{P_{ij}}(D_i))}{\text{card}(U)}, \]  

are obtained. In the above formula, \( P_{ij} \) is scores of risk index \( P_{ij} \) marked by experts, \( D_i \) is scores of item which is corresponding to index \( P_{ij} \), \( \beta \) is the correct classification rate, and \( \text{card}(U) \) is the number of lower indices corresponding to \( D_i \), while \( \text{card}(\text{POS}^\beta_{P_{ij}}(D_i)) \) is the number of \( P_{ij} \) included in decision-making attribute \( D_i \) under the condition of correct classification rate. Judgment matrix is structured by AHP:

\[
\begin{bmatrix}
  a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\
  a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  a_{k1} & a_{k2} & a_{k3} & \ldots & a_{km}
\end{bmatrix},
\]

(13)

\( a_{k,i,j} \) is the element of A about i row and j column, showing the importance between index \( P_i \) and \( P_j \) under the item k, namely,

\[ a_{k,i,j} = \frac{r^\beta(P_{k,i}, D_k)}{r^\beta(P_{k,j}, D_k)} = \frac{\text{card}(\text{POS}^\beta_{P_{k,i}}(D_k))}{\text{card}(\text{POS}^\beta_{P_{k,j}}(D_k))}, \]  

(14)

A is a complete consistency judgment matrix because of \( a_{k,i,j} \times a_{k,j,h} = a_{k,i,h} \), this shows that, when the importance of inter-indices relying on the concept of AHP and VPRS is calculated, there is no need to check the consistency of the judgment matrix, which simplifies the process of the calculating. The index weight in A is defined as following:

\[ W = (W_{P1}^{P1}, W_{P2}^{P2}, \ldots, W_{Pn}^{P1}, \ldots, W_{Pn}^{Pn}) \]  

(15)

\[ W_{Pj}^{Pi} = \frac{W_{Pi}}{W_{Dk}} \]  

For each expert weight \( W_e \), which is defined by eigenvalue of judgment matrix, is used to revised subjective weight to reflect both subjective and objective situations [6]. Specific practices are as follows: assuming k experts mark for each of the bottom indices shown in Table 1 based on the actual situation, and matrix A, B and C are got, while \( \alpha, \beta \) and \( \gamma \) are the eigenvectors corresponding to the max eigenvalue of the matrixes, then it is believed that k experts’ weight are obtained by \( \alpha + \beta + \gamma \) unitized.

TABLE I. SECURITY TECHNOLOGY INDICES

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I01</td>
<td>Physical Location</td>
<td>I05</td>
<td>Dual-network Isolation</td>
<td>I09</td>
</tr>
<tr>
<td></td>
<td>I02</td>
<td>Steal and Destroy</td>
<td>I06</td>
<td>Structural Safety</td>
<td>I10</td>
</tr>
<tr>
<td></td>
<td>I03</td>
<td>Fireproofing</td>
<td>I07</td>
<td>Border Integrity</td>
<td>I11</td>
</tr>
<tr>
<td></td>
<td>I04</td>
<td>Temperature and Humidity Control</td>
<td>I08</td>
<td>Network Equipment Protection</td>
<td></td>
</tr>
</tbody>
</table>

The maximum value of every index is 5, the minimum is 0.
Through the above-mentioned methods, we can get integrated value of every index by experts, and transform every time scores into \( R_k = [U_{k,0}, U_{k,1}, \ldots, U_{k,17}] \), where \( U_{k,i} \) is scores of \( U \) at the \( k \)th time. \( R_k \) can be used to train SVM as a train sample. Training set consists of many training samples. So, we can get the following matrix \( R \),

\[
R = [R_0, R_1, \ldots, R_k]'
\]

For example, the result that 18 indices are marked by four experts is shown in Table 2.

After defining parameter \( W_i \) and \( W_o \) according to (11), we can get the integrated value of 18 indices marked by four experts in Table 3. Now, after Data preprocessing, \( R_0 \) can be used to train SVM as a train sample.

IV. EXPERIMENTAL RESULTS

Put the risk index matrix generated from (17) as a training set and according to the risk level, divide the each result into three types: high, middle, and low, and 1 represents high risk, 2 represents middle risk, and 3 represents low risk. After training, SVM is divided into 3 risk types, that is, 3 represents high risk, 2 represents middle risk, and 1 represents low risk. When forecasting the risk level of electric power information system, it's only to calculate the integrated value of 18 types of risk indices, as an SVM input sample, the output level gets the risk level of this system immediately.

At experiment, use the Libsvm as software platform, and choose 20 samples as the SVM training set and 6 samples as the test set. Using radial basis function as kernel function, and 5 fold cross validation to determine optimal parameters in Fig 3.

Finally define this RBF kernel function's the optimal parameters \( C = 32, \gamma = 2^4 \), and the accuracy rate is 90.0%. Using the model to test the 6 samples of the test set, the results of accuracy rate reached 100%, training and testing accuracy are both met the requirements. Assuming the every index in a risk assessment prediction sample of a electric power information systems is \([0.25, 0.30, 0.20, 0.20, 0.44, 0.20, 0.07, 0.59, 0.50, 0.35, 0.18, 0.06, 0.07, 0.20, 0.32, 0.03, 0.09, 0.38]\), the output result is 2 after prediction, that is, the risk level of the system is middle. As a result, after getting 18 kinds of risk indices of a system, using the risk assessment method based on SVM mentioned in this paper, we can predict the system's risk level to provide an objective evaluation index to policy-makers, and then take appropriate preventive measures.

TABLE II. DECISION-MARKING TABLE OF SECURITY TECHNOLOGY INDICES

<table>
<thead>
<tr>
<th></th>
<th>I_{01}</th>
<th>I_{02}</th>
<th>I_{03}</th>
<th>I_{04}</th>
<th>I_{05}</th>
<th>I_{06}</th>
<th>I_{07}</th>
<th>I_{08}</th>
<th>I_{09}</th>
<th>I_{10}</th>
<th>I_{11}</th>
<th>I_{12}</th>
<th>I_{13}</th>
<th>I_{14}</th>
<th>I_{15}</th>
<th>I_{16}</th>
<th>I_{17}</th>
<th>I_{18}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert1</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Expert2</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Expert3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Expert4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

TABLE III. INTEGRATED VALUE OF THE EIGHTEEN INDICES

<table>
<thead>
<tr>
<th></th>
<th>I_{01}</th>
<th>I_{02}</th>
<th>I_{03}</th>
<th>I_{04}</th>
<th>I_{05}</th>
<th>I_{06}</th>
<th>I_{07}</th>
<th>I_{08}</th>
<th>I_{09}</th>
<th>I_{10}</th>
<th>I_{11}</th>
<th>I_{12}</th>
<th>I_{13}</th>
<th>I_{14}</th>
<th>I_{15}</th>
<th>I_{16}</th>
<th>I_{17}</th>
<th>I_{18}</th>
</tr>
</thead>
<tbody>
<tr>
<td>R_0</td>
<td>0.25</td>
<td>0.26</td>
<td>0.21</td>
<td>0.18</td>
<td>0.44</td>
<td>0.19</td>
<td>0.06</td>
<td>0.57</td>
<td>0.51</td>
<td>0.34</td>
<td>0.13</td>
<td>0.09</td>
<td>0.08</td>
<td>0.22</td>
<td>0.22</td>
<td>0.07</td>
<td>0.08</td>
<td>0.24</td>
</tr>
</tbody>
</table>

V. CONCLUSION

This paper designs a SVM security assessment method based on group decision-marking for electric power information system. Through index weights and expert weights, reduce the impact of the evaluating result from human factors and overcome the subjectivity of assignment directly. And then, according to (11), calculate the integrated value of each index as SVM training set. At last, through analysis and calculate the samples of power information system risk assessment, get an objective and valid conclusion and prove this method feasible. The managers of power information system can use this method to evaluate the risk level of the system regularly, thereby managing the risks more scientifically.

REFERENCES