2012 International Conference on Applied Physics and Industrial Engineering

Risk Assessment of Communication Network of Power Company Based on Rough Set Theory and Multiclass SVM

Xi He¹, Wei Wang²*, Xinyu Liu¹ and Yong Ji¹

¹Henan Electric power research institute
²Zhengzhou Information Science and technology Institute
Zhengzhou, Henan 450002, China

Abstract

This paper proposes a new risk assessment method based on the attribute reduction theory of rough set and multiclass SVM classification. Rough set theory is introduced for data attribute reduction and multiclass SVM is used for automatic assessment of risk levels. Redundant features of data are deleted that can reduce the computation complexity of multiclass SVM and improve the learning and the generalization ability. Multiclass SVM trained with the empirical data can predict the risk level. Experiment shows that the predict result has relatively high precision, and the method is validity for power network risk assessment.

© 2011 Published by Elsevier B.V. Selection and/or peer-review under responsibility of ICAPIE Organization Committee. Open access under CC BY-NC-ND license.

Keywords: risk assessment, rough set, multiclass SVM

1. Introduction

Communication network of power company is a private network which is used for electric power system operation and management, its security and reliability directly impact on stability of the power network. In order to effectively defense against accidents and ensure the secure and stable operation of the power system, the method for vulnerability analysis should be effective, in order that the vulnerable parts of the power communication system can be identified correctly. Effective risk assessment of power communication network is of benefit to protect assets and prevent malicious events. The risk management of the electric power communication system is of great significance to the improvement of operation security and the management level of the whole network.

Information security risk assessment is the risk assessment theory in the field of computer information security. From the angle of scientific risk assessment methods, the procedure can be used to give a scientific, impartial, systematic and comprehensive evaluation of information system with the security attrib-
utes including confidentiality, integrity and availability. The process of evaluation should be implemented in accordance with information security technology standard and management standard formulated by the state. The theory focuses on assessment of threat assets and the probability of security incidents due to the vulnerability.

Information security risk assessment measures are mainly divided into qualitative analysis, quantitative analysis and the analysis method associate with qualitative and quantitative analysis. Qualitative analysis usually concentrates on the losses brought by threat, but ignores the probability of events. The measure presents the relative level mainly through textual description and descriptive numerical range that is widely used. And the typical methods are as follows, such as FTA, ETA, FMEA and Delphi Method and so on. Quantitative analysis calculates risk levels with two fundamental elements that are the probability of threat events and probable losses, and then the relative level is presented by exact calculations. Typical methods of this sort are Factor Analysis, Clustering Methodology, Time Series Model, Regression Model, Decision Tree Method and so forth. The assessment method is subjectivity, randomicity and fuzzification, because the problems of risk assessment have the attributes of complexity, nonlinearity, uncertainty and dynamic nature. Both qualitative analysis and quantitative analysis have their limitations on information security risk assessment. Associative analysis measure can partly overcome the subjectivity that can be suitable for processing nonlinear and uncertain problems. The representative methods are: Fuzzy-AHP, Information Entropy Method, Fuzzy Wavelet Neural network.

Combined with the attribute reduction of rough set theory and multiclass SVM classification, this paper proposes an associative analysis measure for risk assessment. First collected data are processed by the arithmetic based on the attribute reduction of rough set theory, and redundant features are deleted from the decision table without loss of any effective information. Then data are trained based on multiclass SVM classification. Eventually the multiclass classifier is applied to automatic predict assessment of risk levels.

2. Rough Set Theory

Rough set theory offers an effective approach to reduce the data dimensions of information system. By deleting redundant components and attributes from knowledge dataset, unnecessary data are deleted and the key attributes are reserved. In the stage of data preprocessing, we delete redundant features (attributes, objects and attribute values, etc) based on rough set theory, thereby greatly accelerating the computing speed of the system.

Definition 1 A formula for an information system can be expressed as $S = \langle U, C, D, V, f \rangle$, where $U$ is the discussion field, that is, a collection of research objects; $C \cup D = R$ is an attribute set, and subset $C$ and $D$ are called condition attribute set and resulting attribute set respectively; $V = \bigcup v_r, r \in R$ is a set of attribute values, and $v_r$ shows the domain of attribute of a certain attribute $R_r$; Define an information function $f$, that is, $f : U \times R \rightarrow V$. It specifies the attribute value of each object in $U$. Therefore, a form data sheet of a knowledge system $S$ is sometimes referred to as decision table.

Definition 2 For every set of attributes $P \subseteq R$ and two objects $X, Y \subseteq U$, $X$ and $Y$ are indiscernible if and only if $f(X, a) = f(Y, a) (a \in P)$. The expression is as follows:

$$ind(P) = \{(X, Y) \subseteq U : \forall a \in P, f(X, a) = f(Y, a)\} \quad (1)$$

Assuming a given knowledge base $k = (U, R)$, for every subset $X \subseteq U$ and equivalence relation $R \subseteq ind(k)$, $X$ can be classified according to the description of elementary set $R$. For every set $X \subseteq U$, the measurement of classification mentioned above hinges on two precise sets, namely the lower and the upper approximations of a set. Consider two subsets,

$$R(X) = \bigcup \{Y_i \subseteq U : ind(R) : Y_i \subseteq U\} \quad (2)$$
They are respectively called R-lower approximation of $X$ and R-upper approximation of $X$.

Basic rules of reduction algorithm are described as follows: Information system $S$ is described in the Definition 1, where the value of $n$ is $|U|$. And a matrix that has $n \times n$ dimensions is defined:

$$a^*(x,y) = \{a \in R : f(x,a) \neq f(y,a), w(x,y)\}$$

where every element is described as follows: for every $x, y \in U$, we have, $w(x,y)$ satisfies $x \in POS_C(D)$ and $y \notin POS_C(D)$, or $x \notin POS_C(D)$ and $y \in POS_C(D)$ and $(x, y) \notin ind(D)$. $a^*(x,y)$ represents a set of attributes which is the mark that distinguishes $x$ from $y$. And a function $\Delta^*$ is defined. For every attribute $a \in R$, specify Boolean variable $a$. If $a(x,y) = \phi$, then specify Boolean constant 1.

$$\Delta^* = \prod_{(x,y) \in U} a^*(x,y)$$

Conjunction in minimal disjunctive form of $\Delta^*$ is all the $D$ reductions of condition attribute set $C$.

3. Theory of Multiclass SVM

Based on the structural risk minimization principle, SVM is a new machine learning method presented by Vapnik which is well applied in the area of text classification, image identification, handwritten numeral recognition and bioinformatics and the like. Traditional or standard SVM is originally designed for binary classification problem. In this paper, multiclass SVM model is introduced in terms of actual demand of risk assessment.

K-class multiclass classification can be expressed as: there is a certain unknown dependency of variable $y$ and $x$, that is, the two variables follow certain unknown joint probability $F(x,y)$; the optimal function $f(x,w_c)$ is established according to the given $N$ independent identity distribution samples $(x_i, y_i), i = 1, 2, \cdots, N$ (where $x_i \in \mathbb{R}^{(d)}$ is a d-dimension vector, and $y_i \in \{1, 2, \cdots, k\}$ indicates the class of $x_i$). And carry out an assessment of the dependency, thus minimizing the risk.

There are generally two ways of extending SVM to multiclass classification: direct approaches and decomposition approaches. Solving parameters of separation hyperplane of multiple SVM is converted to optimization problem in direct approaches. Multiclass classification is directly implemented by solving the optimization problem. While in the decomposition approaches, multiclass classification is implemented by constructing a series of standard binary SVM classifiers in some way and then grouping them together, among which there are two methods including one-to-many and one-to-one. This paper adopts revised one-to-one decomposition approaches based on one-to-many approaches. One-to-one approach is used to construct $k(k-1)/2$ SVM classification models for k-class multiclass classification:

$$f_{ij}(x) = (w_{ij} \cdot \Phi(x_i)) + b_{ij} (i < j \leq k)$$

Samples belonging to class $i$ are flagged as +1, whereas those belonging to class $j$ are flagged as -1 when training for $f_{ij}(x)$. One-to-one multiclass SVM method is equivalent to solving the $k(k-1)/2$ quadratic optimization problem below which is still the standard binary classification problem:

$$\min_{w_j, b_j, \epsilon_j} \frac{1}{2} \|w_j\|^2 + C \sum_{i=1}^{N} \epsilon_j$$

subject to

$$(w_{ij} \cdot \Phi(x_i)) + b_{ij} \geq 1 - \epsilon_{ij}, y_j = i$$

$s.t.$ $$(w_{ij} \cdot \Phi(x_i)) + b_{ij} \leq 1 - \epsilon_{ij}, y_j = i$$

$\epsilon_{ij} \geq 0, i, j = 1, 2, \cdots, k; i < j; r = 1, 2, \cdots, N$$
In classification we use a voting strategy. If \( \text{sgn}(f_j(x))=1 \), then \( \text{vote}(i) \) of class \( i \) add 1, otherwise \( \text{vote}(j) \) of class \( j \) plus 1. Classify the new sample \( x \) by using the \( k(k-1)/2 \) standard SVM decision function \( f_j(\bullet) \), and record the corresponding votes. Eventually thanks to the comprehensive utilization of \( \text{vote}(m) \), the decision function is gained as follows:

\[
f(x) = \arg \max_{m=1,2,\ldots,k} \text{vote}(m)
\]  

4. Procedure of Risk Assessment

Risk analysis is comprised of three fundamental factors—asset, threat and vulnerability. Asset is valuable information or resources of some organizations. Asset can be estimated from sensitivity, importance and criticality; threat is the possibility of occurrence of unforeseen events that are harmful to the system or organization, with the attributes described by subject of threat, objects influenced frequency of occurrence, motives and consequences; vulnerability is the systematic defect used by threat, which increases the possibility of attack on system0. The attribute is severity of weakness of assets. We construct following risk analysis model, as is seen in Fig. 1:

Fig. 1. Risk Analysis Model

Table I. Specific Indicators of Risk Analysis Model

| Main Class               | ID | Indicator                                      | ID | Indicator                                      |
|--------------------------|----|-----------------------------------------------|----|-----------------------------------------------|
| Asset Identification     |    |                                               |    |                                               |
| A1 Electric Power Supply Facilities | A2 | Host Equipment                                |
| A3 Network Devices       | A4 | Peripherals                                   |
| A5 Anti-virus and Intrusion Detection System | A6 | Identity Authentication and Recognition System |
| A7 Monitoring Equipment  | A8 | Auditing System                               |
| A9 Database              | A10| Document Information                          |
| A11 System Management Software | A12| Software of Office Automation                 |
| Treat Identification     |    |                                               |    |                                               |
| B1 Software Error        | B2 | Hardware Failure                              |
| B3 Influence of Physical Environment | B4 | Disoperation                                  |
| B5 Management Not in Place | B6 | Malicious Code                                |
| B7 Exceed or Abuse Authority | B8 | Cyber Attack                                  |
| B9 Physical Attack       | B10| Breach of Confidence                          |
| Vulnerability Identification |    |                                               |    |                                               |
| C1 Boundary Protection   | C2 | External Access Control                       |
| C3 Installing the Patch  | C4 | User Password                                 |
| C5 Network Security      | C6 | Data Integrity                                |
| C7 Protocol Security     | C8 | Hostile Attack                                |
| C9 DoS                   | C10| Device Failure                                |
| C11 Behavioral Denial    | C12| Physical Environment                          |
This paper details the indicator with regards to the risk analysis model. And specific measures on three fundamental factors are shown in the table above.

In the paper, steps to risk assessment are described in details as below:

step1: Index system establishing. An evaluation index system which includes layers, criterion layers and indicator layers is proposed.

step2: Data Collecting. Sample data are collected according to index system. Acquisition of sample data is the premise and the foundation to the normal operation of risk assessment. Those data come from questionnaire and acquisition of documentation and network operation management activities by auxiliary tools.

step3: Data preprocessing. First of all, the dimensionality of sample data is processed by rough set theory, and the data are reduced, thus dimensionality could better fulfill the model of SVM multiclass classifier.

step4: Design for classifier. In the process, the selection of SVM kernel function and setting parameters are the main operation. Different kernel function constructs SVM of the different structure. Set parameters of multiclass SVM to determine the most suitable kernel function in the training model, after that data are imported for training the model, thus leading to a SVM model of high performance.

step5: Classification decision. The SVM is trained by the test data imported, and the model gains relatively high performance after training, soon the proper classification of test data along with classification accuracy of the model will be accomplished. Test data could be used to classify risk levels of real communication network automatically provided that predict data meet the required condition.

Table II. Comparison between Assessed Values and Empirical Values

| Data number | Machine predict | Expert advice | Data number | Machine predict | Expert advice |
|-------------|-----------------|---------------|-------------|-----------------|---------------|
| 1           | 1 VH            | 1 VH          | 18          | 2 H             | 3 M           |
| 2           | 2 H             | 1 VH          | 19          | 2 H             | 3 M           |
| 3           | 1 VH            | 1 VH          | 20          | 3 M             | 3 M           |
| 4           | 1 VH            | 1 VH          | 21          | 3 M             | 3 M           |
| 5           | 1 VH            | 1 VH          | 22          | 3 M             | 3 M           |
| 6           | 2 H             | 1 VH          | 23          | 4 L             | 3 M           |
| 7           | 2 H             | 2 H           | 24          | 4 L             | 3 M           |
| 8           | 2 H             | 2 H           | 25          | 3 M             | 3 M           |
| 9           | 2 H             | 2 H           | 26          | 3 M             | 3 M           |
| 10          | 3 M             | 2 H           | 27          | 3 M             | 3 M           |
| 11          | 3 M             | 2 H           | 28          | 4 L             | 4 L           |
| 12          | 2 H             | 2 H           | 29          | 4 L             | 4 L           |
| 13          | 2 H             | 2 H           | 30          | 4 L             | 4 L           |
| 14          | 2 H             | 2 H           | 31          | 5 VL            | 4VL           |
| 15          | 1 VH            | 2 H           | 32          | 5 VL            | 5 VL          |
| 16          | 2 H             | 2 H           | 33          | 5 VL            | 5 VL          |
| 17          | 2 H             | 2 H           | 34          | 4 L             | 5 VL          |
5. Experiment

Considering actual state of some certain provincial communication network of electric power and the index system mentioned above, we select 612 samples from all the collected samples as training samples, and 34 samples from the statistical data gained in 2009 as testing samples. The 612 samples are analyzed and their dimensionalities are reduced based on rough set theory.

Owing to the data format of LIBSVM software0 used in training process, data whose dimensionality has been reduced need to be transformed to the format that is suitable for LIBSVM. This paper can classifies the risk levels as very high, high, medium, low and very low level. Select 5×4/2=10 SVM for multiclass classification of risk levels. This paper selects RBF as kernel function of SVM, with C = 200 and \( \sigma = 2 \) as parameters.

Comparison results between assessed values gained by the automatic evaluation tool and empirical values given by experts are shown in the following table. Experiments indicate that risk assessment method presented in this paper is mainly in agreement with empirical method given by experts. And it is rather convenient and automatic for the effective risk assessment of communication of electric power by machine learning.

6. Conclusion

This paper proposes a method based on the attribute reduction of rough set theory and multiclass SVM classification for risk assessment of communication network of certain power company. Collect dataset according to index system and then do preprocessing of data reduction based on rough set theory and reduce dimensionality of attributes, thus lowering the complexity of effective SVM multiclass classification algorithm. Multiclass SVM machine learning is introduced to train the sample data, thus obtaining an effective SVM multiclass classification model. And make decision analysis on test data using the trained model. Experiments demonstrate the method has high precision, it illustrates effectiveness of method for risk assessment of the Power Grid.

References

[1] Dengguo Feng, Yang Zhang, Yuqing Zhang: Survey of information security risk assessment. Journal of China Institute of Communications, vol. 25, No.7, 2004, pp. 10–18.

[2] Visintine V: An Introduction to Information Risk Assessment. SANS Institute 2003.

[3] Pawlak Z: Rough Sets. International Journal of Computer and Information Science, vol. 11, No.5, 1982, pp. 341–356.

[4] Pawlak Z: Rough set theory and its applications to data analysis. Journal of Cybernetics and System, vol. 29, No. 7, 1998, pp. 661–688.

[5] Ziarko W: Variable precision rough set model. Journal of Computer and System Sciences, vol.46, No.1, 1993 pp. 39–59.

[6] Vladimir N V: An overview of statistical learning theory. IEEE Transaction Neural Networks, vol. 10, No.5, 1999 pp. 988-999.

[7] Cristianini N, Taylor J S: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge Univeristy Press, New York ,2000.

[8] GB/T 20984-2007 IT Security IS Risk Assessment Criterion. China Standards Press, Beijing ,2007.

[9] Hong Fan, Dengguo Feng, Yafei Wu: Methods and Applications of the Information Security Risk Assessment. Qinghua University Press, Beijing ,2006.

[10] Chih-Chung Chang and Chih-Jen Lin: LIBSVM:a Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin