Water Treatment Plant Decision Making Using Rough Multiple Classifier Systems

Lin Feng¹, ²

¹College of Computer Science, Sichuan Normal University, Chengdu, Sichuan, China
²Sichuan Key Laboratory of Visualization Computing and Virtual Reality
Chengdu, Sichuan, China
scfengyc@126.com

Abstract

For high-dimensional water treatment plant data sets and a single rough classifier’s weak classification ability for data sets with many classes, a new computing approach, termed CMBMRCS (water treatment plant Classification Model Based on Multiple Rough Classifier Systems), is proposed. First, by combing rough sets theory, some subset of attributes is selected. Then, each simplified data set establishes a group of rough classifiers. Finally, the water treatment plant data classification result is obtained according to the absolute majority voting strategy. The experimental results illustrate the effectiveness of the proposed methods.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of the Intelligent Information Technology Application Research Association. Open access under CC BY-NC-ND license.

Keywords: Rough sets theory; attribute reduction; decision rules; water treatment plant monitoring

1. Introduction

With the rapid growth of industrialization and population, the earth has faced over the last one hundred years has considerably increased environmental pollution, and is affecting water resources, air and soil qualities. In the last decade, wastewater treatment modeling has become a standard engineering tool for wastewater treatment plant design, process optimization, operator training, and developing control strategies [1], and attracted extensive attention by many researchers. Many new wastewater treatment modeling are emerging constantly. For example, Oliveira-Esquerre et al. presented a model of combing the principal components analysis and the back propagated neural network to predict the biochemical oxygen demand (BOD) of the output stream of the biological wastewater treatment plant at RIPASA S/A Celulose e Papel [2]. Rieger et al. designed and developed a system for decision construct and decision making of wastewater issues, which are sound use of wastewater and treatment processes. Measures of common use in this respect are mostly dominated by environmental impact assessment, risk assessment, and cost benefit analysis [1]. Luo et al. proposed a soft computing approach based on the back propagation neural networks and fuzzy-rough sets, which has been applied for forecasting effluent NH3-N, COD and TN concentration...
of a real wastewater treatment plant. The fuzzy-rough sets theory is employed to perform input selection of neural network which can reduce the influence due to the drawbacks of BP such as low training speed and easily affected by noise and weak interdependency data. The model performance is evaluated with statistical parameters and the simulation results indicates that the FR-BP modeling approach achieves much more accurate predictions as compared with the other traditional modeling approaches [3].

However, model predictions can only be as good as the data fed as model input or otherwise used for calibration. Data reconciliation procedures include fault detection, fault isolation, fault identification, and preparation of a data set suitable for the modeling objective [1]. Data mining techniques can help improve model predictions by addressing each and every one of the above mentioned problems [4]. Among data mining techniques, rough sets theory, proposed by Pawlak [5], is a mathematical and suitable approach for handling imprecision, incompleteness and uncertainty in data analysis. Its efficiency has been successfully demonstrated in many applications such as attribute reduction, pattern recognition, and fault detection [6, 7, 8]. It has unique advantage to deal with high-dimensional wastewater treatment plant data.

On the other hand, the classification ability of a single rough classifier is not good for wastewater treatment plant with many classes. So, it is an interest research topic for multiple rough classifier systems. In this paper, a novel intrusion detection approach, termed CMBMRCS (water treatment plant Classification Model Based on Multiple Rough Classifier Systems), is constructed. First, some subset of attributes is selected according to rough sets theory. And then, each of reduced data sets is trained to create a rough classifier respectively. Finally, the final classification result for identifying intrusion data is obtained according to the absolute majority voting strategy. The simulation experiment is also given.

The rest of this paper is organized as follows. Section 2 gives an overview of rough sets theory. In Section 3, we give an overview of CMBMRCS. Section 4 shows the results of evaluating CMBMRCS with wastewater treatment plant data set and Section 5 offers the summary and conclusions of the research.

2. Theoretical Foundations of Rough Sets

For the convenience of later discussion, we first introduce some basic notions of rough sets theory.

**Definition 1:** An information table \( S = (U, R, V, f) \), where \( U \) is a finite nonempty set of objects, \( R \) is a finite nonempty set of attributes, \( V = \bigcup_{r\in R} V_r \), \( V_r \) is a nonempty set of values of \( r \), and \( f : U \times R \rightarrow V \) is an information function that maps an object in \( U \) to exactly one value in \( V_r \). If \( R = C \cup D \), \( C \cap D = \emptyset \), \( S \) is called a decision table or a decision information system, where \( C \) is a set of condition attributes describing the objects, and \( D \) is a decision attribute that indicates the classes of objects.

**Definition 2:** Let \( S = (U, R, V, f) \) be an information table. For any \( B \subseteq R \), there is an associated \( B \)-discernibility relation \( IND(B) \):

\[
IND(B) = \{(x, y) | (x, y) \in U \times U, \forall b \in B (f(x, b) = f(y, b))\}
\]

**Definition 3:** Let \( S = (U, R, V, f) \) be an information table, for any \( B \subseteq R \), and \( X \subseteq U \), the \( B \)-lower approximation of a set can be defined:

\[
B_-(X) = \{x \in U | [x]_B \subseteq X\}
\]

Where \([x]_{IND(B)} = \{y \in U | (x, y) \in IND(B)\}\) is the equivalence class determined by \( x \).

**Definition 4:** Let \( P \) and \( Q \) be equivalence relations of \( U \), the positive region \( POS_r(Q) \) is defined as:
\[
PO_S(Q) = \bigcup_{X \in U / Q} P^*_c(X)
\]  

**Definition 5:** Given \( S = (U, C \cup D, V, f) \), an attribute set \( B \subseteq C \) is a reduct of \( C \) with respect to \( D \) if it satisfies the following two conditions:

1. Jointly sufficient condition: \( \gamma^*_B(D) = \gamma^*_C(D) \)
2. Individually necessary condition: \( \forall b \in B, \gamma^*_{B \setminus \{b\}}(D) \neq \gamma^*_B(D) \)

### 3. Overview of CMBMRCs

CMBMRCs is shown in Fig.1. It is composed of three main modules: attribute reduction, classification rules generation and output the final classification result.

#### 3.1. Attribute Reduction Module

Attribute reduction is one of the core notions in rough sets theory. It is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data. Rough sets can be used to reduce the number of attributes contained in the data set using the data alone, requiring no additional information. In CMBMRCs, attribute reduction can provide the same quality of classification as the original that only monitor a small amount of attributes from network packets. Many approaches of attribute reduction have been developed in the area of rough sets. In rough sets theory, a subset of attributes defines an equivalence relation. Based on the equivalence partition, one can induce a set of positive rules and a set of boundary rules, respectively. In machine learning, this is commonly referred to as the problem of feature selection. In rough set analysis, the problem is called attribute reduction, and a selected set of attributes for rule induction is called a reduct. Intuitively speaking, an attribute reduct is a minimal subset of attributes whose induced rule sets have the same level of performance as the entire set of attributes, or a lower but satisfied level of performance [9].
3.2. Classification rules

An important application of rough sets is to induce decision rules, in terms of rough classifiers that indicate the decision class of an object based on its values on some condition attributes in the decision table. Decision rules are expressed by the form if [condition] then [consequent]. In rough sets, decision rules can be generated by an inductive learning principle, and can be fall into two categories: certain rules and possible rules. Certain rules are generated from the lower approximation sets, while possible rules are generated from the upper approximation sets [4].

In rough sets, the process by which the maximum number of condition attribute values is removed without loosing essential information is called value reduction. The resulting rules are optimal because their conditions maximally general or minimal length. In this paper, we use rules inductive learning algorithm (value reduction) to generate a set of rules.

3.3. Classification Process

For identification data, we can give an approach based on [10].

Approach: Classification process for identification data.
Input: identification intrusion detection data \( x \);
Output: classification result of \( x \).

Step1. \( x \) input rough classifiers 1, rough classifiers 2, ..., rough classifiers \( n \) respectively.
Step2. For \( i = 1 \) to \( n \)

Calculate the output result of rough classifiers \( i \).
Step3. The final classification result of \( x \) is obtained according to the absolute majority voting strategy.

4. Experimental testing and analysis

In order to evaluate the CMBMRCS approach, a water treatment plant database [11, 12] was chosen. The dataset is a set of historical data charted over 521 days, with 38 different input features measured daily. Each day is classified into one of thirteen categories depending on the operational status of the plant. The purpose of experiment is to compare the classification performance of single rough classifier method with multiple rough classifiers method. The evaluation criterion is the classification accuracy. In the experiment, we use the Nguyen improved greedy algorithm to discrete continuous-valued attributes and use the general attribute value reduction to generate decision rules. The minority priority matching strategy is adopted for testing data. The experimental results are showed in Table 1.

| The times of experiment | Multiple rough classifiers | Single rough classifier |
|-------------------------|----------------------------|-------------------------|
|                         | Classification accuracy (%)| Classification accuracy (%)|
| 1                       | 78.43                      | 75.67                   |
| 2                       | 77.52                      | 75.88                   |
| 3                       | 75.79                      | 73.24                   |

The results of experiment show that multiple rough classifiers method is the higher classification accuracy than single classifier method. Consequently, multiple rough classifiers have better performances for the complex water treatment plant data classification problems.

5. Conclusions

Over the past ten years, data mining techniques have been successfully applied in the fields including marketing, manufacturing, process control, fraud detection, and network management. In this paper, we
proposed a CMBMRCS approach. The experimental results illustrate that the proposed method has the higher classification accuracy than single classifier method, and could meet the demand of accuracy for the water treatment plant data classification problems with many classes.

6. Acknowledgment

The authors thank the anonymous referees for their good comments. This paper is in part supported by the Scientific Research Fund of Sichuan Provincial Education Department under Grants No.09ZC079, and the Key Research Foundation of Sichuan Normal University, respectively.

References

[1] L. Rieger, I. Takaes, K. Villez, et al., “Data reconciliation for wastewater treatment plant simulation studies—planning for high-quality data and typical sources of errors,” Water Environment Research, vol. 82, no. 5, pp. 426-433, 2010.

[2] K. P. Oliveira-Esquerre, M. Mori, and R. E. Bruns, “Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis,” Brazilian Journal of Chemical Engineering, vol. 19, no. 4, pp. 365 - 370, 2002.

[3] F. Luo, R. H. Yu, Y. G. Xu, et al., “Effluent quality prediction of wastewater treatment plant based on fuzzy-rough sets and artificial neural networks,” International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 09), IEEE Press, Aug. 2009, vol. 5, pp. 47-51, doi:10.1109/FSKD.2009.494.

[4] L. Feng and Y. L. Yuan, “An intrusion detection architecture based on multiple rough classifiers integration,” Journal of Experimental & Theoretical Artificial Intelligence, unpublished.

[5] Z. Pawlak, “Rough sets,” International Journal of Information and Computer Sciences, vol. 11, no. 5, pp. 341-356, 1982.

[6] L. Feng, G. Y. Wang, and X. X. Li, “Knowledge acquisition in vague objective information systems based on rough sets,” Expert Systems, vol. 27, no. 2, pp. 129-142, 2010.

[7] L. Feng, G.Y. Wang, and T.R. Li, “Knowledge acquisition from decision tables containing continuous-valued attributes,” Acta Electronica Sinic, vol. 37, no. 11, pp. 2432-2438, 2009.

[8] Q. Wu and Z.T. Liu, “Real formal concept analysis based on grey-rough set theory,” Knowledge-Based Systems, vol. 22, no. 1, pp. 38-45, 2009.

[9] Y. Y. Yao and Y. Zhao, “Attribute Reduction in Decision-Theoretic Rough Set Models,” Information Sciences, vol. 178, no. 17, pp. 3356-3373, 2008.

[10] L. Shang, J. G. Wang, W. S. Yao, et al., “A classification approach based on evolutionary neural networks,” Journal of Software, vol. 16, no. 9, pp. 1577-1583, 2005.

[11] Q. Shen and R. Jensen, “Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring,” Pattern Recognition, no. 37, pp. 1351–1363, 2004.

[12] C.L. Blake and C.J. Merz, UCI Repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA, 1998, http://www.ics.uci.edu/mllearn/MLRepository.html.