Pattern Recognition for MCNs Using Fuzzy Linear Programming

Jing He[1,3], Wuyi Yue[2], and Yong Shi[3]

[1] Institute of Intelligent Information and Communication Technology, Konan University, Kobe 658-8501 Japan
hejing@gucas.ac.cn
[2] Department of Information Science and Systems Engineering, Konan University, Kobe 658-8501 Japan
yue@konan-u.ac.jp
[3] Chinese Academy of Sciences Research Center on Data Technology, and Knowledge Economy, Beijing 100080 P.R. China
yshi@gucas.ac.cn

Abstract. This paper presents a data mining system of performance evaluation for multimedia communication networks (MCNs). Two important performance evaluation problems for the MCNs are considered in this paper. They are: (1) the optimization problem for construction of the data mining system of performance evaluation; (2) the problem of categorizing real-time data corresponding to the data mining system by means of dividing the performance data into usual and unusual categories. An algorithm is employed to identify performance data such as throughput capacity, package forwarding rate, and response time. A software named PEDM2.0 (Performance Evaluation Data Miner) is proposed to improve the accuracy and the effectiveness of the fuzzy linear programming (FLP) method compared with decision tree, neural network, and multiple criteria linear programming methods.

1 Introduction

Performance evaluation and network planning are the key tools in a reliable multimedia communication operation. Multimedia communication networks (MCNs) to support several different traffic types have become so complex that intuition alone is not sufficient to evaluate their performance. Mathematical models of performance systems range from relatively simple ones, whose solution can be obtained analytically, to very complex ones that must be simulated [1].

An important challenge for identification mining in MCNs is the identification speed that can forward the exponentially increasing volume of traffic. The data mining system can provide the new identification service that is needed by next-generation MCNs.

Research of linear programming (LP) approaches for classification problems was initiated by [3]-[5]. [6] and [7] applied the compromise solution of multiple criteria linear programming (MCLP) to deal with the same identification mining

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part IV, LNCS 3994, pp. 509–516, 2006.

question. [8] presented an analysis for fuzzy linear programming in classification of credit card holder behavior. In [9] the identification mining of unusual patterns for MCNs based on FLP is put forward for the first time. In this paper, we present some new research work.

The unusual pattern mining process can be described as follows: given a set of n performance evaluation data, there will be K objects that are considerably dissimilar, exceptional, or inconsistent with the remaining data.

In Section 2, we describe a data mining system for performance evaluation. The subsystems for an identification mining engine based on fuzzy linear programming are presented in Section 3. In Section 4, the results of data experiments are listed out. Finally, we conclude with a brief summary in Section 5.

2 Data Mining System for Performance Evaluation

Generally, the methods to calculate network performance include analytical, numerical and simulation methods. Nowadays, emulation is the main method for performance evaluation systems for MCNs.

A data mining system for performance evaluation can be constructed with the above three methods of analytical, numerical and simulation method. Fig. 1 shows the architecture of the data mining system for performance evaluation that we present in this paper.

These components are explained in detail as follows:

Graphical user interface
This module communicates between the users and the data mining system, allowing the user to interact with the system by specifying a performance evaluation query or task, and providing information to help focus the search.

Index system for performance evaluation
The main idea of this index system comes from [1]. The details of the index can be found from the following multi-dimensional data warehouse module.

Pattern evaluation
The process of this module is shown in Fig. 2.

An index system is acceptable if (1) it is easily understood by humans, (2) it is valid on new or test data with some degree of certainty, (3) it is potentially useful, and (4) it is novel [2].

Multi-dimensional data warehouse
Before we use our on-line analytical processing (OLAP) tools, the multi-dimensional data warehouse for performance evaluation must be constructed. The snowflake schema is a variant of the star schema, where some dimension tables are normalized thereby further splitting the data into additional tables. The snowflake schema of data warehouse is shown in Fig. 3.

Pre-computation and summarization
This module, which involves data integration and data cleaning, can be viewed as an important preprocessing step for data mining. Data from operational

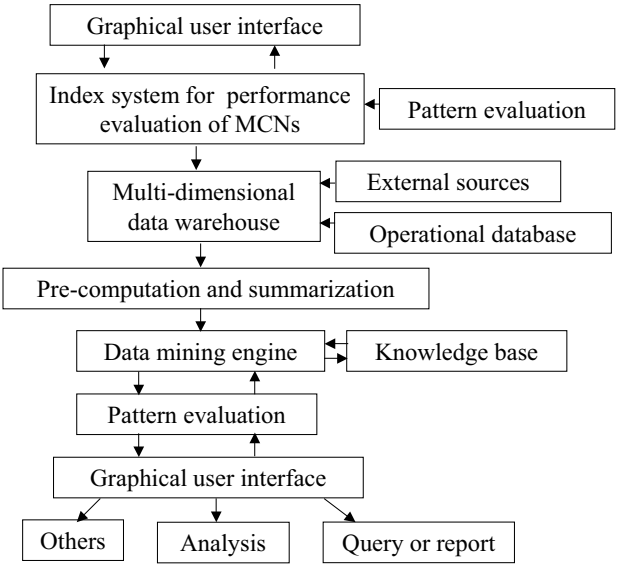

Fig. 1. Architecture of data mining system for performance evaluation

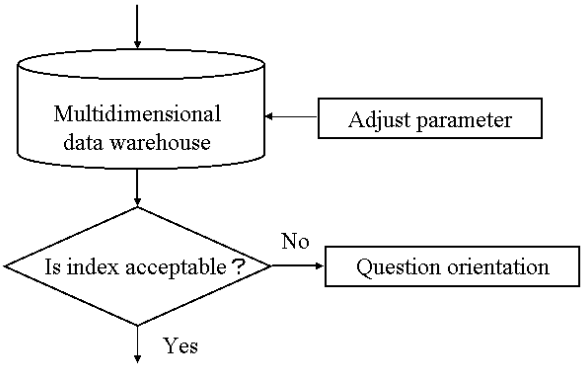

Fig. 2. Pattern evaluation process

databases and external sources (such as performance information provided by external sensors) are extracted using application program interfaces known as gateways.

3 Identification Mining Model

A basic framework of the identification mining model of unusual patterns can be presented as follows:

Given a set of r attributes about a MCN, let $\boldsymbol{A}_i = (A_{i1}, ..., A_{ir})$, $i = 1, 2, ..., n$ be training set data for the variables of every MCN, where $\boldsymbol{A}_i$ is the attributes set

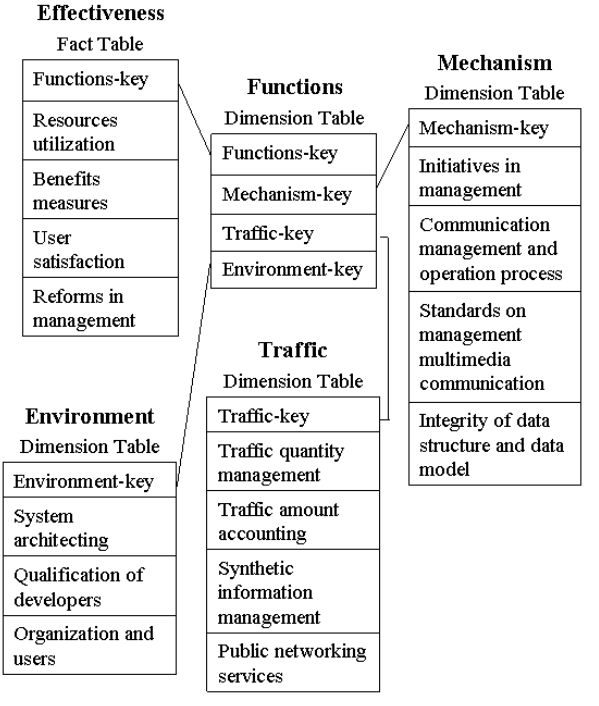

Fig. 3. Snowflake schema of data warehouse

of the ith training set, and n is the sample size. We want to determine the best coefficients of the variables $\boldsymbol{X} = (X_1, X_2, ..., X_r)^T$, where X_j is the coefficient of the variable A_{ij}, $i = 1, 2, ..., n$, $j = 1, 2, ..., r$. A boundary value b (a scalar) to separate the two classes: N (normal patterns) and M (unusual patterns), is as follows:

$$\begin{aligned} \boldsymbol{A}_i \boldsymbol{X} &\leq b, \boldsymbol{A}_i \in M, \\ \boldsymbol{A}_i \boldsymbol{X} &\geq b, \boldsymbol{A}_i \in N. \end{aligned} \tag{1}$$

To measure the separation of usual and unusual patterns, we define that α_i is the overlapping of a two-class boundary for case A_{ij} (external measurement), $i = 1, 2, ..., n$, $j = 1, 2, ..., r$. α is the maximum overlapping of a two-class boundary for case A_{ij}, $\alpha_i < \alpha$.

We define β_i to be the distance from case A_{ij} to their adjusted boundaries (internal measurement), $i = 1, 2, ..., n$, and β to be the minimum distance from case A_{ij} to their adjusted boundaries, $\beta_i > \beta$.

A model that seeks MSD (the minimal sum of the deviations of the observations from the critical value) can be written as follows:

$$\text{(M1)} \quad \text{Minimize} \sum_{i=1}^{n} \alpha_i,$$

$$\begin{aligned} \boldsymbol{A}_i\boldsymbol{X} &\le b + \alpha_i, \qquad \boldsymbol{A}_i \in M, \\ \boldsymbol{A}_i\boldsymbol{X} &\ge b - \alpha_i, \qquad \boldsymbol{A}_i \in N \end{aligned} \tag{2}$$

where $\boldsymbol{A}_i$ is given, $\boldsymbol{X}$ and b are unrestricted, and $\alpha_i \ge 0$, $i = 1, 2, ..., n$.

The alternative of the above model is to find MMD (the minimal distances of observations from the critical value are maximized). It can be written by

$$\begin{aligned} \text{(M2)} \quad & \text{Maximize } \sum_{i=1}^{n} \beta_i, \\ & \boldsymbol{A}_i\boldsymbol{X} \ge b - \beta_i, \qquad \boldsymbol{A}_i \in M, \\ & \boldsymbol{A}_i\boldsymbol{X} \le b + \beta_i, \qquad \boldsymbol{A}_i \in N \end{aligned} \tag{3}$$

where $\boldsymbol{A}_i$ is given, $\boldsymbol{X}$ and b are unrestricted, and $\beta_i \ge 0$, $i = 1, 2, ..., n$.

In a linear discriminate model, the misclassification of data separation can be described by two objects: MSD and MMD. Therefore, the main research aim in the identification of unusual patterns of MCNs is to seek the method that produces the higher detection accuracy.

Let y_{1L} be MSD and y_{2U} be MMD, the maximum value of $\sum_{i=1}^{n} \alpha_i$ is y_{1U} and the minimum value of $\sum_{i=1}^{n} \beta_i$ is y_{2L}. To explore this possibility, we propose a heuristic identification of the unusual pattern method by using the fuzzy linear programming for discovering the unusual patterns in MCNs as follows:

$$\begin{aligned} \text{(M3)} \quad & \text{Maximize } \xi, \\ & \xi \le \frac{\sum \alpha_i - y_{1L}}{y_{1U} - y_{1L}}, \\ & \xi \le \frac{\sum \beta_i - y_{2L}}{y_{2U} - y_{2L}}, \\ & \boldsymbol{A}_i\boldsymbol{X} = b + \alpha_i - \beta_i, \qquad \boldsymbol{A}_i \in M, \\ & \boldsymbol{A}_i\boldsymbol{X} = b - \alpha_i + \beta_i, \qquad \boldsymbol{A}_i \in N \end{aligned} \tag{4}$$

where $\boldsymbol{A}_i$, y_{1L}, y_{1U}, y_{2L}, y_{2U} are known, $\boldsymbol{X}$ and b are unrestricted, and α_i, β_i, $\xi \ge 0$, $i = 1, 2, ..., n$.

Method

(1) Create data warehouse for the performance evaluation of every MCN at every selected time spot.

(2) Generate a set of relevant attributes from the data warehouse, transform the scales of the data warehouse into the same numerical measurement, determine the two classes of usual and unusual patterns, as well as the classification threshold τ that is selected by the user, and the training set and the verifying set.

(3) Give a class boundary value b and use models (M_1), (M_2), and (M_3) to learn and compute the overall scores $\boldsymbol{A}_i\boldsymbol{X}$ $(i = 1, 2, ..., n)$ of the relevant attributes or dimensions over all observations repeatedly.

(4) If (M_1) exceeds the threshold τ, go to (7), otherwise go to (5).

(5) If (M_2) exceeds the threshold τ, go to (7), otherwise go to (6).
(6) If (M_3) exceeds the threshold τ, go to (7), otherwise go to (3) to consider to give another b.
(7) Apply the final learned scores $\boldsymbol{X}^*$ to predict the unknown data in the verifying set.
(8) Find the unusual patterns of the MCNs.

The FLP approach proposed in this paper is simpler and easier to get the meaningful results. For example, this FLP approach can get more meaningful solutions than the common classification approaches in the multiple criteria linear programming.

Real-time CNs data can be used to test our data mining system. Based on the above analysis, we have developed a software named performance evaluation data miner (PEDM2.0) [10]. This miner is an OLAP miner integrated with an OLAP whose mining is in relational databases. The development language is the C++ syntax based on Linux. This miner also combines with the algorithm of linear & non-linear programming in those softwares named Lingo9.0 and Lindo8.0 [11].

The FLP approach is not the only module in identification mining of the PEDM2.0. Statistics, decision tree, linear programming, multiple criteria linear programming, neutral networks are also used. The output results in the PEDM2.0 are the synthesis integration results based on different methods. A comparison study in terms of computational efficiency implementation will be discussed in the next section.

4 Data Experiments

Given a set of attributes, such as throughput capacity, package forwarding rate, response time, connection attempts, delay time, transfer rate and the criteria for "unusual" patterns, the purpose of pattern recognition for the MCNs is to find the better classifier through a training set and use the classifier to predict all other aspects of the performance of MCNs.

The frequently used pattern recognition in the telecommunication industry is still two-class separation technique. The key of two-class separation is to separate the "unusual" patterns called fraudulent activity from the "usual" patterns called normal activity and identify as many MCNs as possible. This is also known as the method of "detecting fraudulent list".

In this section, a real-time performance data mart with 65 derived attributes and 1000 records of a major CHINA TELECOM MCN database is first used to train the different classifiers [12]. Then, the training solution is employed to predict the performance of another 5000 records of MCNs. Finally, the classification results are compared with the decision tree, neural network and MCLP.

The results are shown in Table 1. Three known classification techniques have been used to run and test the 5000 records of the CHINA TELECOM MCN database. These results are compared with the FLP approach shown in Table 1. The software of the decision tree is the commercial version called C5.0 [13], while

Table 1. Identification rate comparisons on balanced 5000 records

Approaches	**Identification rate**	**Time (second)**
Decision Tree	79.39%	0.335
Neural Network	64.20%	0.201
MCLP	80.03%	0.936
FLP	81.74%	0.284

the software for both MCLP and FLP were developed at Chinese Academy of Science in China and Konan University in Japan [10].

Note that in Table 1 the column identification rate represents the rate of identifying the right unusual patterns in respective models as: Identification Rate = (Number of identified unusual patterns exactly) / (Number of unusual patterns) $\times 100\%$.

The identification time is calculated using different models. Because this data mining system is special for the MCNs, the FLP model is not the model with the fastest calculation speed, but it does have a higher calculation speed and higher identification rate.

The greater the identification rate is, the better the result is. As we see, the model that predicts best is the FLP with 81.74%. The second best model is the MCLP with 80.03%. The decision tree model has the third best prediction rate with 79.39% while the neural network is the worst one with 64.20%.

The shorter the identification time is, the better the result is. The fastest model on the identification time is the neural network with 0.201 seconds. The second is FLP with 0.284. Decision tree has the third fastest time with 0.335 seconds while MCLP is the slowest one with 0.936 seconds.

Both short identification time and high identification rate are important in identification mining calculations of MCNs. Therefore, if the data set is balanced, it is meaningful to implement FLP algorithm proposed in this paper. This conclusion, however, may not be true for all kinds of data sets because of the different data structure and data feature.

Many decision makers in MCNs often get a better result through an FLP approach. It has been recognized that in many decision making problems, instead of finding the exist "optimal solution" (a goal value), decision makers often approach a "satisfying solution" between upper and lower aspiration levels that can be represented by the upper and lower bounds of acceptability for objective payoffs. The model proposed in this paper also can be used in bioinformatics, antibody and antigen.

5 Conclusions

In this paper, an identification mining model of unusual patterns for MCNs has been presented. The construction flow of data mining systems of MCNs for performance evaluation, the snowflake schema of a data warehouse, and the algorithm of fuzzy linear programming were shown in detail. The data experi-

ments proved that the fuzzy linear programming (FLP) approach we proposed in this paper has excellent accuracy and effectiveness compared with decision tree, neural network, multiple criteria linear programming methods.

Acknowledgments

This work was supported in part by GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No. 16560350) and MEXT.ORC (2004-2008), Japan and in part by NSFC (No. 70472074, No. 70531040), 973 Project (No. 2004CB720103), Postdoctoral Science Foundation, China and BHP Billion Co., Australia.

References

1. Yue, W., Gu, J., Tang, X.: Performance evaluation index system for multimedia communication networks and forecasting for web-based network traffic. Journal of System Science and System Engineering **13** (1994) 44–50
2. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2003)
3. Freed, N., Glover, F.: Simple but powerful goal programming models for discriminant problems. European Journal of Operation Research **7** (1981) 44–60
4. Freed, N., Glover, F.: Evaluating alternative linear programming models to solve the two-group discriminant problem. Journal of Decision Science **17** (1986) 151–162
5. Glover, F.: Improve linear programming models for discriminate analysis. Journal of Decision Science **21** (1990) 771–785
6. Kou, G., Liu, X., Peng, Y., Shi, Y., Wise, M., Xu, W.: Multiple criteria linear programming approach to data mining: models, algorithm designs and software development. Journal of Operation Methods and Software **18** (2003) 453–473
7. Kou, G., Shi, Y.: LINUX based Multiple Linear Programming Classification Program: Version 1.0. College of Information Science and Technology, University of Nebraska-Omaha (2002)
8. Shi, Y., He, J., Wang, L., Fan, W.: Computer-based algorithms for multiple criteria and multiple constraint level integer linear programming. Computers and Mathematics with Applications **49** (2005) 903–921
9. He, J., Yue W., Shi, Y.: Identification mining of unusual patterns for multimedia communication networks by using fuzzy linear programming. IEICE Technical Report **DE2005-17** (2005) 11–17
10. He, J., Shi Y.: Performance Evaluation Data Miner 2.0, CAS Research Center on Data Technology and Knowledge Economy (2005)
11. http://www.lindo.com/
12. http://www.chinatelecom.com.cn/
13. http://www.rulequest.com/see5-info.html/