Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S3, 2019, pp. 30-34

Data Mining Methods for Communication Technology

V. Saraswathi Bai
Assistant Professor, Computer Science and Engineering Department, Soet,
Sri Padmavathi Visva Vidyalayam, Tirupati, Andhra Pradesh, India
E-Mail: saihumika9@gmail.com

***Abstract* - The aim of the project was to analyze the behavior of military communication networks based on work with real data collected continuously since 2005. With regard to the nature and amount of the data, data mining methods were selected for the purpose of analyses and experiments. The quality of real data is often insufficient for an immediate analysis. The article presents the data cleaning operations which have been carried out with the aim to improve the input data sample to obtain reliable models. Gradually, by means of properly chosen SW, network models were developed to verify generally valid patterns of network behavior as a bulk service. Furthermore, unlike the commercially available communication networks simulators, the models designed allowed us to capture non-standard models of network behavior under an increased load, verify the correct sizing of the network to the increased load, and thus test its reliability. Finally, based on previous experience, the models enabled us to predict emergency situations with a reasonable accuracy.**
Keywords: **Communication Network (CN), Data Mining (DM), Data Preparation, Decision Trees**

I. INTRODUCTION

Within the research the methods of knowledge discovery in databases (data mining) for behavior analysis of the real communication network have been applied. The customized data mining methodology has been used to develop models for several individual tasks we have recognized in communication network behavior. So far, we have mastered the first two phases of the methodology, the problem analysis and the data analysis and data understanding [1]. Now, we are planning to create and exploit the data mining models of data preparation and data modeling phases. We would like to interpret and apply these models to individual tasks, such as:

Classification and prediction of selected network parameters like short-term and long-term traffic load.

1. Standardized network behavior for individual net-work lines in selected intervals.
2. Comparison of traffic-load on different days of the week.
3. Discovering trends in network behavior and its usability for prediction improvement.
4. Proper network dimension verification.
5. Finding groups of line with similar characteristics using segmentation algorithm.
6. Extreme values detection.

Prediction of the increase in traffic load due to irregular users' activities. This capability will distinguish our models from common commercial simulators. The already developed models will be described in the paper and the influence of entering data quality on possible model interpretation will be discussed. For modeling the database (SQL language and Relational Database Management System) and data mining tools were used. SPSS PASW Modeler and Microsoft Access are used for experiments.

A similar problem of communication networks behavior analysis through an analysis of operational load addresses many works, such as [2], [3], [4], [5].

All of the mentioned publications work with other types of networks in terms of their principles of work; paper [2] analyzes the Internet. Similarly to our contribution, the paper focuses on the dimensioning of the network. As the evaluation parameter the percentage distribution of the traffic load was selected; the authors deal with 4 lines for the period less than 6 months. Our project verifies the dimensioning of the data on backbone network (22 lines), collected continuously since 2005. As an assessment option chosen in our project is the peak busy hour (PBH) that can express not only the intensity of losses, but also distribution of maximal values in the network over time. As in our paper the authors address the increasing trend in the load on the network. The analysis is performed only on data from four lines observed for a period of about 6months, in our opinion, data are not statistically significant enough for such type of task.

Paper [3] deals with the type of WLAN networks, publication [4] deals with Private Mobile Radio Network.

Work [5] performs an analysis of network traffic using analytical modeling techniques. They create the appropriate models based on the same simplifying assumptions as we do; it means the stationarity of processes examined at selected intervals in mean and variance. The analyzed network is a complex queuing system, which cannot be easily described mathematically, even not for a suitably chosen initial conditions and simplifying assumptions. Therefore, we have decided to apply computer simulations in our work.

Our network differs from the others according to the theory of queuing. We analyzed the network that operates as a

system with losses, works [2], [3], [4] and [5] dealt with the queuing systems. Furthermore, unlike the work of mentioned papers and unlike the commercially available communication networks simulators, the selected methods and models allow us to detect nonstandard behavior of the network in focus and moreover the models are able to predict such situation with sufficient accuracy.

The organization of this paper is as follows. In section 2 the analyzed network is described in terms of the net-work topology, the technology and the principles of work. In addition, the parameters chosen for evaluation for analysis of the network behavior are defined with the justification of their choice. In section 3 we address the process of data preparation for data mining. Individual used operations and their importance for the improvement of input data are briefly characterized. As a final output of this part the analytical table with well-defined inputs and outputs for modeling is created.

In section 4 we describe analysis tasks processed on the network, created models and simulation results are discussed here as well as the results of simulations and the resulting recommendations for network optimization. In section 5 we give our conclusions.

II. COMMUNICATION NETWORK IN FOCUS

If we want to model the behavior of a system in focus precisely, we have to be acquainted with its structure, relations of its elements, processes inside the system and their time sequence. There is a short characteristic of the communication infrastructure of the analyzed Stationary Military Communication Network (SMCN) [6].

A. The Topology of the SMCN

The communication infrastructure of the SMCN has been built on the basis of modern digital principles since the beginning of 1990s in accordance with world's trends and civilian and military standards. The principles of work are TDM (Time Division Multiplex) and PCM (Pulse Code Modulation), the protocols used are ISDN (Integrated Ser-vices Digital Network). The SMCN is mainly designed for voice transfer.

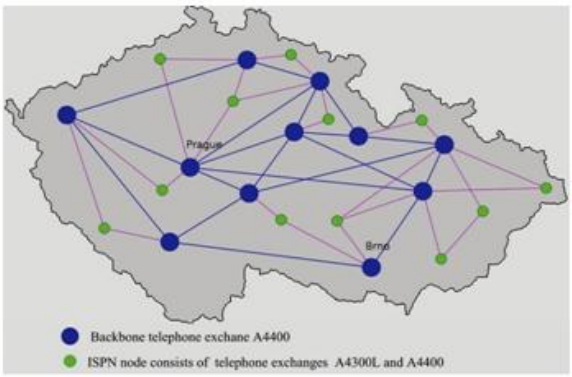

Fig. 1 A fictitious topology of the SMCN

The SMCN consists of the modern digital exchange A4400 and A4300L designed by the French firm Alcatel. They are connected by E1 flows with the business signalization ABC, version F1 and F2. The backbone consists of the digital exchange A4400 only. ISPN node is created by twenty PBX (Private Branch Exchange) A4300L or A4400 maximally. These are connected by a special ISPN signaling protocol [7]. That is why one ISPN node appears to other network as only one big PBX. The ISPN nodes are connected by two E1 flows at least to the backbone [8].

We are processing real data acquired in NMC (Net-work Management Control) supervisory centers. During the first processing we limit ourselves to a Backbone Net-work only. In Fig. 1, the fictitious topology of the network is presented. The real network nodes and connections differ in numbers and location from the ones in the figure.

B. Network Variables

We have selected permeability as a basic evaluation criterion for SMCN behavior and it can be expressed as a function of the traffic load in the individual backbone trunks. The traffic load depends on the type of the system, kind and range of services provided, users' activity, period of the day and night, days, weeks and months in a year, working hours, etc. Generally, the knowledge of the traffic load enables us to examine the use of connecting lines and their real need, to assess the quality of the telecommunication system operation, to compare the system quality with other tele-communication systems, to compare and harmonize all the parts of the connecting system. The traffic load is a function of many variables. The most frequently monitored traffic load values follow: The intensity of the carried traffic load, *Y*, can be expressed as: Data preparation is a complex and time consuming process. Data pre-processing takes significant time during the data mining project (from 20% to 60%). This step should not be omitted or underestimated, as the modeling results can distinguish from expected results.

III. DATA CLEANING

The main aim of data cleaning is filling out missing values, finding outliers, smoothing out noise and correcting data inconsistencies.

For the missing value analysis many methods can be applied, from simple ones, like ignoring records with missing attributes, filling out values manually or filling out values automatically with global constant or attribute mean, to more sophisticated, like using attribute mean according to classes or filling out the most probable value generated by regression or decision tree induction.

Noise, the random error or variance in a measured variable, should be smoothed in data source before the selected data mining algorithms are applied (some methods like neural networks are able to smooth data while processed, but in

general, the data sources should avoid such discrepancies). The binning, regression or clustering techniques are usually utilized for data smoothing.

A. Data Preparation Methods

Data obtained from real word are usually very noisy, missing, and even inconsistent. The data should be usually pre-processed to receive true and valuable models that can be used for analysis. There are a lot of data pre-processing techniques [1] and we will apply the selected ones to improve data quality provided in our domain by communication network.

1. Data Cleaning: This technique is usually applied to remove noise and correct inconsistencies.

2. *Data Integration:* It is used for merging data from multiple heterogeneous sources into one stable and coherent data storage, such as transactional relational data-base (RDBMS – Relational Database Management System) or data warehouse (EDW – Enterprise Data Warehouse). Alternatively other data store types can be utilized, like stores based on XML (extensible Markup Language) applications.

3. Data Transformation: It is another data preparation technique used for substitution of current data with new improved data better fit to modeling algorithm. Normalization is one of the transformation techniques applied for methods, where distance measures are critical (e.g. clustering or segmentation).

4. Data Reduction: This technique provides methods for the reduction of the data size. For example, it involves data aggregation, removing redundant or related data and clustering.

B. Data Integration

The data for analysis usually come from more than one data source. One real-world object can be stored in different data sources; therefore this object should be identified in all the data sources and is matched up. In our project data comes from three different sources. The main data source is a collection of the log files provided by network management system, the second data source consists of manually created data records for irregular user activities analysis. The last data source contains tables with connection between loss, offer and traffic load.

Redundancy is the next important issue. An attribute derived from other attributes or attribute correlated with another existing attribute is usually redundant for back-ground algorithms and can be removed from the table without losing data meaning. The correlation analysis, e.g. Pearson's correlation coefficient for numerical or χ^2 (chi-square) test for categorical variables, is usually applied at this step for decreasing data redundancy.

C. Data Transformation

During data transformation, the data are transformed into more appropriate form for further analysis. The basic data transformation tasks are aggregation, generalization, normalization and attribute construction. As an example, the traffic attribute is constructed according to expression (2) for each record have different capacities (30 or 60 channels), thus line 1 (60 information channels) is loaded at 50 percent at peak hours as well as line 14 or 3 (30 channels).

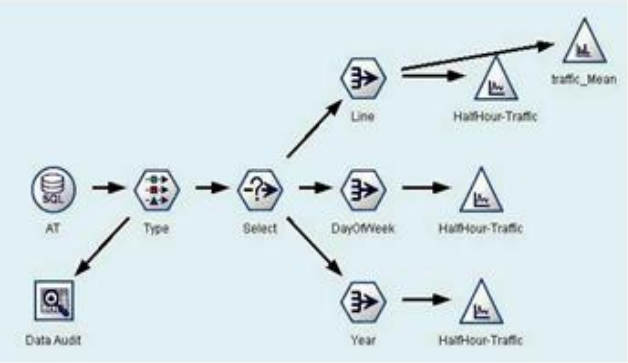

Fig. 2 The process of daily traffic load graphs creation

Probably due to organization changes, reductions in our organization and due to technology update using different network types. Permanent data transfers show a continual increase.

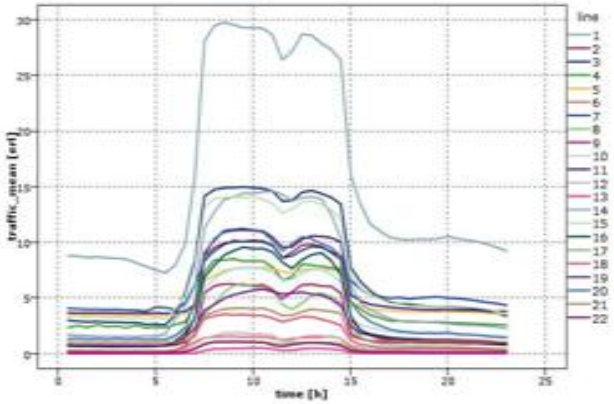

Fig. 3 The average traffic load for individual lines

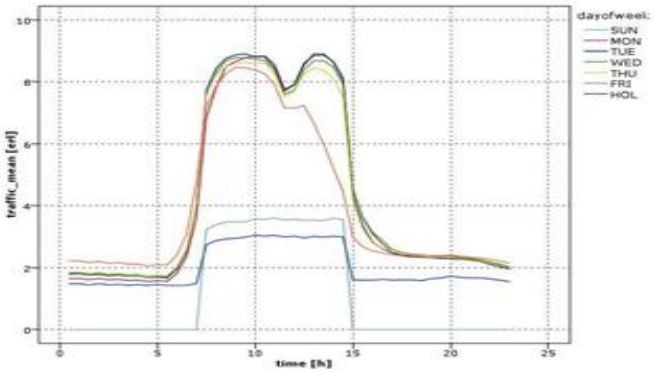

Fig. 4 The average traffic load for different type of days

It is obvious (Fig.4) that the permanent data traffic is present on the lines. The level of traffic can be uncovered from the holiday traffic load. Sundays (abbr. SUN) do not show the permanent traffic due to incorrect settings in management system only for working hours (as was discovered later). Daily traffic (traffic Mean) differs in years. Traffic load in 2006 is higher than the load in 2007 or 2008.

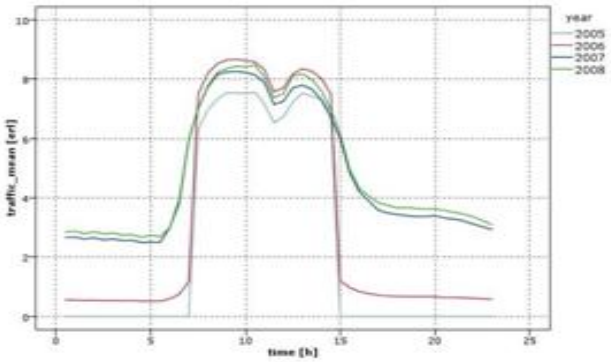

Fig. 5 The average traffic load according to years

IV. DAILY TRAFFIC LOAD COMPARISON

As an example of different behavior of traffic load, two days of week (Friday /FRI and Tuesday/TUE) were selected for comparison. The highly loaded line 1 from 2006 to 2008 was analyzed. Our hypothesis of the different traffic load and of daily minimums and maximums was verified (Fig.7). The graph shown is an output of the stream in Fig.6.

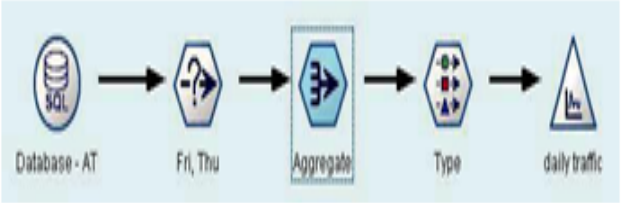

Fig. 6 Traffic load comparison – process

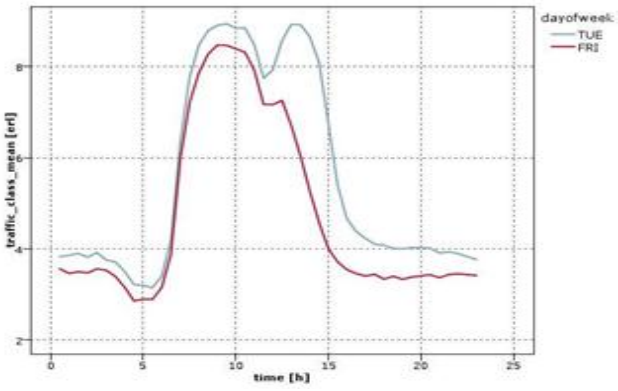

Fig. 7 Traffic load comparison - Friday vs. Tuesday

The traffic load peak on Friday ends at 1pm, whereas the peak on Tuesday ends at 3pm. The presumption of lower load on Friday was verified as well. The rules for night data transfers can also be localized from the graphs, e.g. distribution in time, traffic load size. The decrease of load at 5am is caused by the backup procedures runs. The following streams (Fig.8, Fig.10) provided more valuable results. For training and testing, the first model uses the sets of the following pattern:

The first setting for training set: line=3, year=2008, month in <1;3>.
The next settings for training set: line=3, year=2008, month in <1+*n*;3+*n*> for *n* in <1;7>. Seven different settings were tested in total.The testing set was fixed: line=3, year=2008, month=10.
Line 3 was selected, as it is one of the most loaded lines. Results are displayed in Tab. 5; values are also displayed on the Fig. 10.

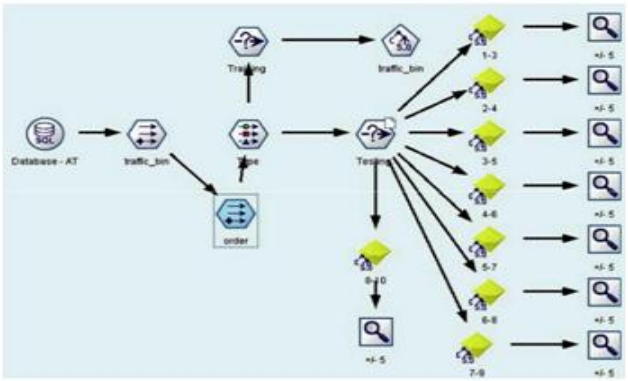

Fig. 8 Prediction for months – process

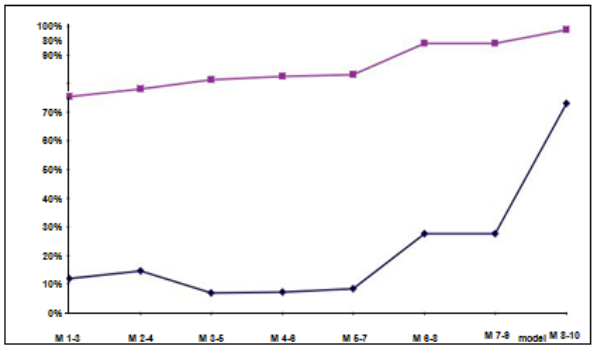

Fig. 9 Accuracy of predictions for individual models

Node traffic bin in the stream creates a discrete set (rounding by 1 erlang) for the continuous output value.

TABLE ISUMMARY OF PREDICTIONS FOR STREAM MONTHS

Model	Exactly	+/- 5
M 1-3	11,72	75,52
M 2-4	14,65	78,16
M 3-5	6,71	81,38
M 4-6	7,37	82,51
M 5-7	8,51	83,17
M 6-8	27,6	94,05
M 7-9	27,5	94,05
M 8-10	72,97	98,77

Note. Exactly in table Imeans that the rounded predicted value is equal to rounded measured value of load; +/- 5 means that predicted value is in interval +/- 5 erlangs. The second model creates a discrete set for output value per 5 erlangs, thus for 60 circuit line 12 intervals were created and for 30 circuit line 6 intervals were created in the traffic bin node. Each interval is substituted by a category value.

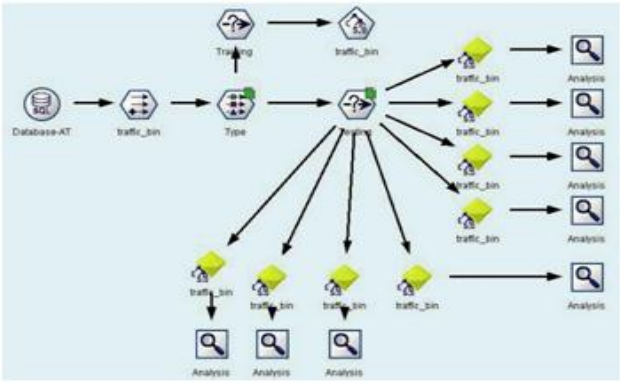

Fig. 10Interval predictions – process

V. CONCLUSION

In the paper the focus was on the analysis of a communication network using data mining techniques. The possibility to deal with a large amount of data (millions of data values) uncovers hidden dependencies in data. The entry data quality has a significant impact on further analysis. The data preparation techniques were applied to original data, and new data collections and analytic tables with pre-processed data were produced. Gradually, by means of properly chosen software, network models were developed to verify generally valid patterns of network behavior as a queuing system and the network dimensioning was confirmed. During the analysis the lines with higher loss values in comparison with the rest of lines were detected. The selected lines are not sized properly and in the future they can be vulnerable to extreme loads. Furthermore, unlike the commercially available communication networks simulators, the selected methods and models allow us to detect nonstandard behavior of the network in focus. Nonstandard network behavior is always associated with defects in the network, the operator intervention in its topology; it can also be related to natural disasters or current international situation. It always reflects the behavior of the network users. This might vary depending on the period of the day, the type of the day (holidays and working days), and also in connection with planned activities. Next, the analysis of the load increase trends has been carried out. Based on the outputs, the topology change.

REFERENCES

[1] J Han, and M Kamber, *Data Mining – Concepts and Techniques*. Morgan Kaufmann Publishers, 2006.

[2] M. O Hiba, M. S Sami and, H. M Elhag, "A backbone Internet traffic intensities and statistics in Sudan". In *3rd International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA 2006)*. Orlando (USA), 2006.

[3] M. Ghazali and M. Azminbin, "*Analysis on the Traffic Load Pattern of Unikl WLAN"*. Thesis. University Utara, Malaysia, 2007.

[4] B Can,."*Traffic Analysis and Modeling in PMR Systems"*. Thesis. Electrical and Electronics Engineering Department, Bilkent University, 2003.

[5] N.X Liu, and J.S. Baras, "Statistical modeling and performance analysis of multi-scale traffic". In *Proceedings of the 22nd AnnualJoint Conference of the IEEE Computer and Communications Societies, INFOCOM 2003*. San Francisco (CA, USA), pp.1837-1847, 2003.

[6] J Ondrasik, and A Mazalek, "New trends in ACR field communication systems", In *Proceedings of New Technologies in Radio-Communications*. University of Defense, Brno (Czech Republic), pp. 7, 2006.

[7] *Technical Documentation for A4300 and A4400*. Alcatel-Lucent,avenue Kléber - 92707 Colombes France, 1996, 2009.

[8] V Ondryhal, and Z. Vranova, "Using data mining methods for communication net behavior analysis", *EEČasopis proelektrotechnikuaenergetiku*, Vol.14, No. 5, pp. 282-286, 2008.

[9] V. Ondryhal, and Z. Vranova, "Different ways to identify trends in network traffic", In *Proc. of Networking and ElectronicCommerce Research Conference 2010*. Lake Garda (Italy), October 2010.

[10] *Rule Quest Research Data Mining Tools - Sample Applications Using See5/C5.0* [online] Cited 2011-04-12. Available at http://www. Rulequest.com/see5-examples.html.