Classification of Network State Using Data Mining

H. Singh[*], K.E. Burn-Thornton[*] & P.D. Bull[#]

[*] Data Mining Group, School of Computing, University of Plymouth, 9 Kirkby Place, Plymouth, Devon, PL4 8AA, UK. Tel: +44 (0) 1752 232711; 232621 Fax: +44 (0) 1752 232540 Emails: hsingh@soc.plym.ac.uk, kburn-thornton @ plym.ac.uk Web: http://www.tech.plym.ac.uk/soc/research/dmg

[#] Wavetek Wandel & Goltermann (WWG) Communications Test Solutions, Eurotech House, Burrington Way, Plymouth, Devon, PL5 3LZ, UK. Tel: +44 (0) 1752 765326 Fax: +44 (0) 1752 783000 Email: phil.bull@wago.de

ABSTRACT

The emergence of new transport technologies coupled with deregulation and privatisation has contributed to the contiguous growth of telecommunications networks particularly in terms of both the intricacy and size of the network. The rapid growth in network size, and intricacy, is of a concern to those who are involved in Network Management - particularly those involved with network operation, administration, maintenance and provisioning (OAM&P) functions. The integration of the evolving and emerging technologies, and systems, with legacy systems provides additional concerns for those endeavouring to ensure availability of the network resources particularly those required to meet agreed Service Level Agreements (SLAs). In this paper, we discuss the potential use of Data Mining algorithms and techniques, for classifying the Network State, and hence whether SLAs are being met, by analysing performance indicative data collected from networks using the Synchronous Digital Hierarchy (SDH) as an exemplar underlying transmission system.

Keywords
Data Mining, Network Management, Telecommunications, Service Level Agreement (SLA), Knowledge Discovery, Network State, Alarm Correlation, Synchronous Digital Hierarchy (SDH), Networks, Pro-active Management, Quality of Service (QoS).

1. INTRODUCTION

The setting up of SLAs demands a high standard of network availability and performance through improved quality management systems. The current practises of network management can overload network management architectures and thus agreed quality of service (QoS) **[1]** defined in the SLA contracts. This is due to the various processes enacted in the Maintenance Function **[2]** which involves –monitoring, and managing, extremely large amounts of data as a result of the sheer size, and complexity of the networks. Data that is collected in order to achieve QoS, is determined via analysis of the network performance indicators. This can involve analysis of voluminous samples of data collected from monitoring of network performance, and alarms used to resolve or avoid faults **[3]**, which is collected for the maintenance function. Typically, for example, performance information collected routinely for an ATM network every 15 minutes amounts to 15MB **[4]**.

The onset of overloading network operation centre is the availability of the network, which is dependent on the restoration and preventative **[5]** (e.g. re-routing the network configuration) functions enact in the network management processes. Network elements (NEs) affect each other and consequence or sequential **[6]** generation of notification messages is inevitable as a result of the occurrence of faults. A single incident, or fault, due to a particular anomaly or defect may trigger multiple generation of notification messages. These notification manifests externally as alarms and depending on the protocol used by the Management systems, an alarm is referred to as a trap or notification **[7]**, where the later using CMIP, while the former using SNMP. When considering the network complexity, which can consist of several hundred to thousands of NEs, this represents a large amount of alarms. Prominently, this can cause alarm inundation of a network operation centre and has strong antecedents in fault localisation. Hence analysis of the alarms enabling, hypothesis of the root cause of the fault to be proposed.

Performance information collected routinely and alarms, in addition to alarm log which, contains information about an alarm event entered in chronological sequence, provides a huge repository of raw data from networks. Using domain expert to analyse this data can be exhaustive and time consuming especially the classification of the state of the network and the discovery of latent trends or patterns to resolve transient problems. Consequently, alarm correlation **[8]** has been one of the many correlation techniques employed in fault localisation. Alarms, with germane

entities or trends are grouped together to form new semantics and the parenthesis is enriched to provide useful information pertaining the generation of the alarms which, may include corrective measures. Hence,

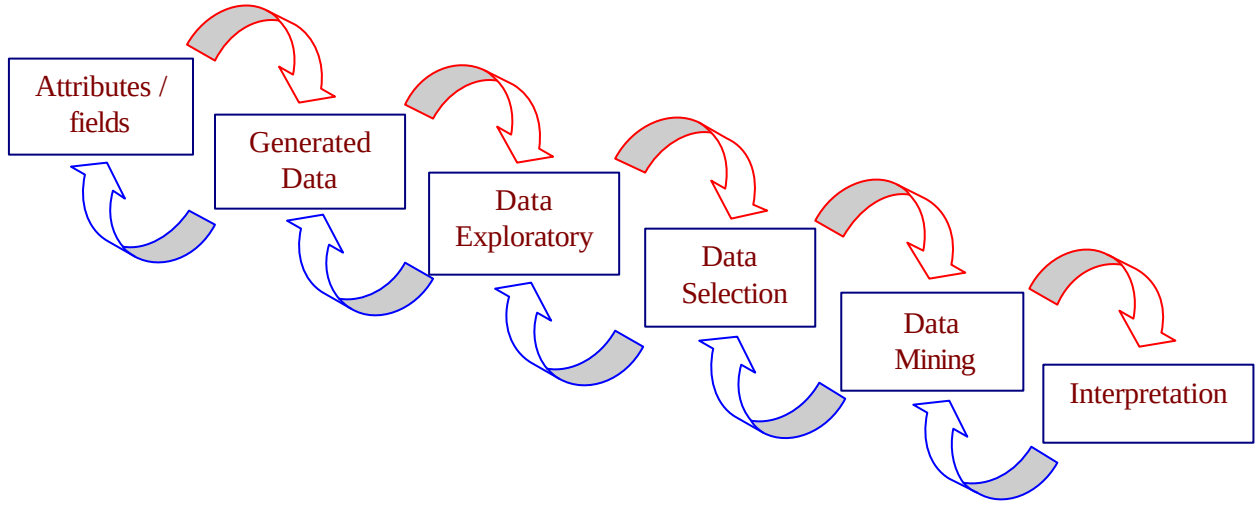

Figure1: Methodology for alarm correlation

this reduces the amount of information displayed and provides plausible information, to enable the subtle task of fault diagnosis.

In this paper we, discuss the potential of Data Mining algorithms, and techniques for the prediction of problems that are likely to persist, as well as those that are likely to degrade network performance – hence enabling classification of the network performance indicators. In section 2, we provide a conceptual introduction to Data Mining. This is followed by description of the methodology used which, is the focus of this work; the Data Mining of SDH network performance indicators in the telecommunication domain. Section 4 provides the results of this investigative work while section 5 provides the conclusion and discussion. Future work is also discussed in section 6.

2. DATA MINING

Data Mining (DM) can be described as a collection of techniques and methodologies used to explore vast amounts of data in order to find potentially useful, ultimately understandable patterns **[9]** and to discover relationships. DM is an iterative and interactive process, involving numerous steps with many decisions being made by the user. The fundamental goals of data mining are finding latent trends in data, which enables prediction and description **[10]** of the analysis phases. DM is a rapidly expanding field which, has been exploited in lucrative domains such as in the financial **[11]** business **[12]** and communications **[13]** domains although little reported work has been carried out to determine network state by analysis of the network performance indicators of alarms **[14-17]**.

2.1 Data Mining task

The initial sequence upon acquiring the required understanding of the proposed application domain or *priori* knowledge is to determine the DM task. Different algorithms are optimised based on the predefined DM task. This involves deciding whether the goals of the DM process is classification, association, or sequential **[18]**. Classification has two distinct meanings. We may aim of classifying new observations into classes from established rules or establishing the existence of classes, or clusters in data **[19]**. Association attempts to generate rules or discover correlation in data and is expressed:

- X => Y, where X and Y are sets of items.

This means that an event or transaction of database that contains X tends to contain Y.

Sequential looks at events occurring in a sequence over time or time-ordered sequences. This could be expressed through the following:

- E ? N, E is a set of event types, an event pair (A, t), where A ? E is an event type.

Where t represent the time of the event or occurrence of an event. This is followed by predefined sets of fault conditions where:

- F is a set of fault types and C ? F, C is a fault type, hence for example:

90% of the time, if the event (A, t) occurs, it is followed by fault type C.

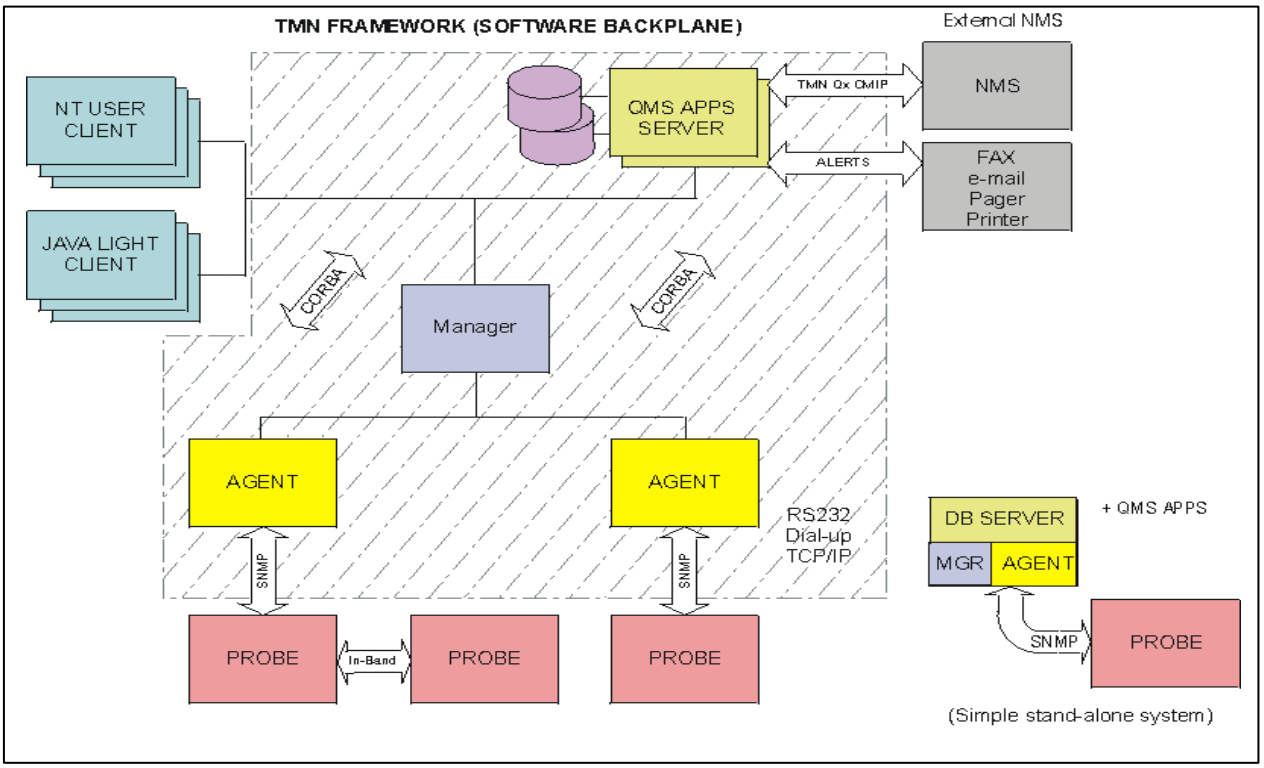

Figure 2: WWG QMS System Architecture

2.2 Techniques and methodology

The subsequent process once the DM task is defined can be derived from the four main activities; *selection, pre-processing, data mining and interpretation*, also known as post-processing.

Selection involves creating a target data set to undergo analysis, paradoxical to the assumption that the complete raw data is presented to the DM software, due to the nature of the data which may present irrelevant attributes or information pertinent to other mien of the domain. The data set selected can be focused on a subset of variables or data samples, of the proposed application domain. The recent and historical data can be found, combined, and transformed into an organised target data set repository.

Pre-processing in DM involves preparing the target data set prior to undergo analysis using the Data Mining software. It may involve converting the data into acceptable format to the data mining software, demarcation of the beginning of each message or may involve processes to discreetise continuous features for instance, as used to generate the alarm semantic. Pre-processing can be used to constrain the search space, and can make patterns or relationships in the data more visible in the later stages of the DM process. The *data mining* process involves subjecting the cleaned (containing reliable information) data to be analysed by the data mining algorithm(s) and results of the analysed (mined) data is presented to the next stage.

Interpretation involves verification of the results, hence analysing the results of the analysis which may include selecting interesting rules based on the DM task. It may also involve re-iterating some of the processes in order to provide further information. The elicited interpreted analysis can be implemented or adapted in a correlation framework or system for pro-active management applications particularly, which requires real-time data as in the telecommunications domain for fault diagnosis.

3. METHODOLOGY

The Data Mining methodology for the determination of network state is determined via analysis of the network performance indicator as depicted in ***Figure 1*** and is derived from the four basic activities carried out during Data Mining activities or processes; selection, pre-processing, data mining and interpretation. For the

focus of this work the pre-processing stages, constitute the activities of generating the data and data exploratory on the data set. These are the important stages prior to selecting the target data set, which will ITU-T G703, G704 and domain specific performance indicators.

3.1 Alarm semantic

The semantic of the alarm is based on Wavetek Wandel & Goltermann (WWG) Quality Management System (QMS). QMS fault management concept and semantic is based on ITU-T X.745 and X.733 recommendations respectively. WWG QMS system architecture and relationships between each of the main functions is as depicted in ***Figure 2***. WWG QMS enables the end-to-end visibility of network and service performance to achieve the desired levels of service quality and honour SLA commitments as conceptually depicted in ***Figure 3***.

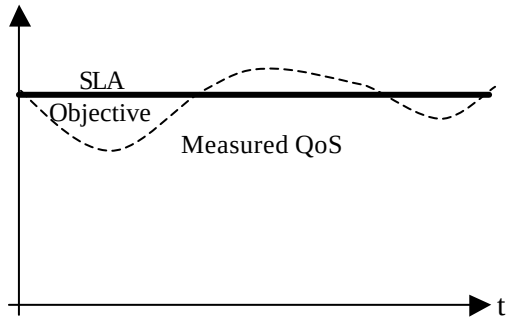

Figure 3: SLA and QoS relationship

3.2 Data Mining algorithms

The Intelligent Data Analysis (IDA) Data Mining Tool **[19]** used for the data mining process incorporates algorithms from the fields of Statistical, Machine Learning and Neural Networks. The five algorithms chosen for this investigative work are k-NN, C4.5, CN2, RBF and OC1. The generated data sets were split into ratios of 9:1, 8:2 and 7:3, hence into two parts; a training set and a testing set, which is a common technique used known as train and test. The algorithms or classifier is subjected initially with the training set and then the classification accuracy is tested using the unseen data set or testing set. The results give an indication of the error rate and the overall classification accuracy of the trained algorithms.

4. RESULTS

The result of this initial investigative work suggests that the Machine Learning algorithms, C4.5 and CN2 in particular, performed better. This results, is in comparison with the other algorithms used for this initial trials, explicit details on this is beyond the scope of this paper but will be included in future papers. C4.5 undergo further analysis to identify patterns and to test specific hypotheses. The alarm types used in this investigative work are based upon recommendation of

is a decision tree Machine Learning based algorithm and uses a modified entropy measurement to calculate the gain in information. Entropy is used to measure how informative a node, by splitting the data at various boundaries. The output result of the *data mining* process upon subjecting the data to undergo analysis by the DM tool could be in the following form:

Attribute 6 = Location 15: 1 (2.0)
Attribute 6 = Location 13:
| attribute 4 = Direction_1: 1 (5.0/2.0)
| attribute 4 = Direction_2: 2 (3.0)

This output can be paraphrased as:

If attribute 6 = Location 15 **then** 1
If attribute 6 = Location 13 **then**
  If attribute 4 = Direction_1 **then** 1
  else
  If attribute 4 = Direction_2 **then** 2

The (m/n) gives both the number of correctly and incorrect mapped data, entered in the node's leaf.

CN2 is a rule based Machine Learning algorithm, which, like C4.5 belongs to the general class of recursive partitioning algorithms. The CN2 algorithm for this investigative work was optimised to generate *unordered* rule set using the Laplace statistical significance prediction in generating the rules. The former enables the search in each class to be re-iterated, removing only covered examples of that class when a rule has been found. Hence, having found a good conjunct of attribute or complex, the rule 'if <complex> then predict <class>' is added to the end of the rule list. The latter tends to avoid the undesirable 'downward bias' of entropy **[21]**. The rule for example could be as following:

```
IF   attribute 1 = "Type_44"
  AND  attribute 6 = "Location6"
THEN  class = c3  (9 6 3)
```

The (a b c) in this case gives the training examples covered by rules in each classes, for this example three classes were used.

5. CONCLUSION AND DISCUSSION

The growth in network size, integration of multiple sub-networks and many different vendors' equipment, makes the task to elicit the required *priori* knowledge a difficult task. Consequently, the task to disambiguate the alarms based on hypothesis tends to become more knowledge intensive since, the alarm semantic are not explicit enough to provide important information

required for the diagnosis of fault and can be ambiguous. Hence, additional information is required which even to the most erudite domain expert, this proves to be a subtle task. The development of efficient and efficacious tools as well as methodology, are required in order to allow a pro-active or to compliment human interpretation of some of the functions enacted for the purpose of network management. The initial results, from this investigative work provides a framework for some of these tasks.

6. FUTURE WORK

The initial methodology, based on the results of the trials which, will be used, will be further developed and integrated into a correlation systems framework or intelligent data analysis tool. This could be incorporated, into or form, part of a network monitoring or test equipment to enable pro-active network management.

ACKNOWLEDGEMENT

The authors would like to thank Wavetek Wandel & Goltermann (WWG) Communications Test Solutions of Plymouth (UK), DTI and the EPSRC for the financial support for this project.

REFERENCES:

[1] D. Sanchez, D. Guerrero and A. Vina, "Modelling Legacy PDH Equipments from General Models to Real TMN Solutions", proceedings of MOMS'98, 186-195.

[2] A. Mahdi, K.E. Burn-Thornton and P. Bull, "Expert Diagnosis Engine for Remote Test Management of Telecommunications Network", proceedings of INC'98, ISBN 1-84104-016-8, 179-184.

[3] G. Jakobson, M. Weissman, "Alarm Correlation", IEEE Networks-1993, Vol.7, No.6, 52-59.

[4] K.E. Burn-Thornton, J. Garibaldi and A. Mahdi, "Pro-Active Network Management using Data Mining", GLOBECOM '98, 1-9.

[5] R. Gardner and D.A. Harle, "Expert Data Mining For Alarm Correlation in High-Speed Networks", proceedings of IITT EXPERTSYS'97, 145-150.

[6] H. Mannila, H. Toivonen and V.A. Inkeri, "Discovery of Frequent Episodes in Event Sequence", Data Mining and Knowledge Discovery-1997, Vol.1, No.3, 259-289.

[7] S. Aidarous and T. Plevyak, "Telecommunications Network Management into the 21st Century Techniques, Standards, Technologies and Applications", IEEE Press-1993, ISBN 0-7803-1013-6, 1-17.

[8] T.A.M. De Castro and J.M.S. Nogueira, "An Alarm Correlation System for SDH Networks", Proceedings of IT'98, 492-497.

[9] U.M. Fayyad, "Data Mining and Knowledge Discovery; Making Sense Out of Data", IEEE Expert-1996, Vol.11, No.6, 20-25.

[10] P. Adriaans and D. Zantinge, "Data Mining" Addison-Wesley-1996, ISBN 0-201-40380-3.

[11] C. Westphal and T. Blaxton, "Data Mining Solution, Methods and Tools for Solving Real-World Problems", Wiley-1998, ISBN 0-471-25384-7, 531-585.

[12] F. Giannotti, G. Manco, M. Nanni, D. Pedreschi and F. Turini, "Integration of Deduction and Induction for Mining Supermarket Sales Data", proceedings of PADD'99, ISBN 1-902426-04-5, 179-93.

[13] R. Sasisekharan and V. Seshadri, "Data Mining and Forecasting in Large-Scale Telecommunications Networks", IEEE Expert Intelligent Systems and Their Applications-1996, Vol.11, No.1, 37-43.

[14] R. Gardner and D.A. Harle, "EventFlow: A Technology for Alarm Correlation in High-Speed Networks", proceedings of INC'98, ISBN 1-84104-016-8, 171-178.

[15] S. Hajela, "Alarm Management in Telecommunications Networks", Hewlett Packard Journal-1996, Vol.47, No.5, 22-30.

[16] A.T. Bouloutas, S. Calo and A. Finkel, "Alarm Correlation and Fault Identification in Communications Networks", IEEE Transactions on Communications-1994, Vol.42, No.2/3/4, 523-533.

[17] G. Jakobson, G. Weihmayer and M. Weissman, "A Domain Oriented Expert System Shell for Telecommunications Network Alarm Correlation", proceedings of 2nd IEEE Network Management and Control Workshop-1993, Vol.2, 365-380.

[18] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining" AAAI Press / The MIT Press-1996, ISBN 0-262-56097-6, 1-31.

[19] D. Michie, D.J. Spiegelhalter and C.C. Taylor, "Machine Learning, Neural and Statistical Classification", Ellis Horwood-1994, ISBN 0-13-106360-X, 6-16.

[20] K.E. Burn-Thornton and L. Edenbrandt, "Myocardial Infarction-Pinpointing the Key Indicators in the 12 lead ECG Using Data Mining", Journal of Computers and Medicine-1998, Vol.31, No.31, 293-303.

[21] P. Clark and R. Boswell, "Rule Induction with CN2:Some recent Improvements", proceedings of the 5th European Conference (EWSK-91), 151-163.