

## PRO-ACTIVE NETWORK MANAGEMENT USING DATA MINING

K.E.Burn-Thornton, J. Garibaldi & A.E. Mahdi  
 School of Electronic, Communication and Electrical Engineering  
 University of Plymouth  
 Drakes Circus  
 Plymouth  
 PL4 8AA  
 United Kingdom  
 Tel: +44 1752 232519 Fax: +44 1752 232583  
 email: k.burn-thornton@plymouth.ac.uk  
<http://www.tech.plym.ac.uk/see/staff/kbt/kbt.htm>

### Abstract

Network providers are continually looking towards the future, perhaps the year 2002, when network management systems will possess the capability of pro-actively changing network configurations and traffic routing, in real-time, in order to optimise traffic throughput in the network, and hence maximise their profit margin. However, the major barrier to the production of truly real-time pro-active network management systems appears to be the technology, or technique, which will enable the vast amount of network performance data, which is routinely collected, to be analysed, and for pro-active system changes to be made, in real-time. The rapidly emerging field of Data Mining appears to offer a solution to this problem. In this paper we discuss the potential suitability of the four main Data Mining categories of Data Mining algorithms, to the task of analysing network performance data in true real-time. Hence we provide a discussion regarding the classes of Data Mining algorithm which will provide the basis for the provision of a real-time pro-active network management system by virtue of their ability to rapidly mine vast amounts of data.

**Key Words:** Data Mining, Pro-active Management, On-line, Real-Time, Networks.

### 1 Introduction

The primary aim of this work was to determine which was the best class of Data Mining algorithms to use in order to, rapidly and accurately, mine the vast amounts of performance data that are routinely collected from a network. A secondary aim was to determine, how the algorithms could be modified so that more rapid mining could be achieved, and hence the goal of true real-time Data Mining achieved.

We begin by discussing the requirements for pro-active network management tools, and then provide a brief outline of the four major classes of Data mining algorithms and their suitability to the task in hand. The main body of this text describes the algorithms which were chosen, from each of the four categories, as well as

a description of the test data which was chosen for this investigation. This is followed by a discussion of the training and testing of the algorithms on the test data. Conclusions are then drawn with regard to the suitability of the various algorithms for mining network performance data, which is in effect large set, multivariable, multiclass data. A discussion regarding the applicability of the algorithm, and hence class of algorithms, to various techniques which may be implemented in order to increase its speed of mining is presented.

Future work is also discussed with regard to implementation of techniques for increasing the speed at which the algorithm mines, and its future potential for use in true real-time proactive network management systems.

### 2 Pro-active Network Management

Network providers are continually striving for network management systems which will possess the capability of pro-actively changing network configurations and traffic routing, in real-time, in order to optimise traffic throughput in the network, and hence maximise their profit margin[1],[2]. Such a system will also enable non-skilled operators to provide the minimal monitoring of the system, which will be required, and reduce the need for highly skilled, expensive, skilled operators so enabling the network provider to become more competitive[3],[4].

There are two barriers which currently preclude the day to day use of pro-active management systems. The first barrier is the lack of reliable, fully 'fleshed out' knowledge-bases providing information regarding the correct preventative, or palliative, action to be taken when a network presents various conditions e.g. cell loss at a given switch. The second barrier is the speed of information processing. A great deal of performance information is routinely collected for networks, typically 15MB in 15 minutes for an ATM network [5]. In order to determine events happening in the network, an operator must be able to rapidly search through the vast amounts of performance information in order to make knowledge-based changes to the network. The technique used to

process the network performance data will need to be as fast as possible.

The majority of the work which has been carried out towards the end goal of providing pro-active network management systems has been focussed on enabling pro-active management of a network by the use of knowledge-based system support[6],[7] i.e. surmounting the first barrier. We seek to overcome the second barrier, namely determining an optimal technique which enables true real-time system management to be achieved by rapidly determining information from the vast amounts of performance data.

### 3 Data Mining

Data Mining enables vast amounts of data to be mined in order to find valid, novel, and potentially useful and ultimately understandable patterns. Data mining is a rapidly expanding field which has yet to be fully exploited in many domains or the techniques have yet to be modified to address the particular problems in the domain. The lucrative domains such as financial[8] were the first domain to which Data Miners sought to ply their trade, although others are now seeking to focus the application of data mining to less financially lucrative, but the less transient domains of communications [2], in order to determine which are the appropriate techniques to modify in order to solve the particular 'gold nugget' problems in this domain.

There are four main categories of Data Mining algorithm, Statistical, Machine Learning, Neural Networks and Hybrid, each having members which are particularly suited to performing more accurately over certain characteristic data sets e.g. numerical or categorical with many (k-NN)[10] or few variables (SMART[11]) and many (ALLOC80[11]), or few (Cal5[11]), classes.

The Statistical algorithms are characterised by an underlying probability model, which provides an estimate of probability of being in a class rather than a simple classification. It is generally assumed that users of such techniques have a strong statistical background which results in expert intervention with regard to variable selection, transformation, and overall structuring of the problem.

The Machine Learning algorithms use automatic procedures based upon logical or binary operations in order to learn a task from a series of examples. The majority of these algorithms employ decision-tree and rule induction approaches in order to generate classifying expressions simple enough to be understood. These algorithms can be implemented without using expert intervention, although expert knowledge may be used.

The Neural networks algorithms consist of layers of inter-connected nodes, each node producing a non-linear function of its input. The input to a node may come from another node or directly from input data. The complete network represents a complex set of interdependencies, which may include a degree of non-linearity, allowing for general functions to be modelled. These algorithms combine the complexity of statistical techniques with machine learning.

The hybrid algorithms contain a multitude of combinations of different classes of algorithms and include the new algorithm that we are currently developing which is a hybrid of the k-NN and linear discriminant algorithm.

### 4 Data Mining Algorithms

The algorithms chosen for this investigation were: k-NN(Statistical)[10], C4.5(Machine Learning, decision tree)[13], CN2(Machine Learning, rule-induction)[14], RBF(Neural Networks)[15] and OC1(Machine Learning, decision tree algorithm)[16].

The chosen algorithms represent one from each Data Mining class and additional algorithms from the Machine Learning class so that the two types of machine learning algorithm, namely rule induction and decision tree, could be evaluated.

### 5 Data Sets

Four data sets were used in this investigation, two are artificially generated data sets. The sets were chosen because they all possess a large number of attributes and a small number of classes, something that network performance data also possesses since network performance data consist of a collection of a lot of performance parameters which equate to the network being in a 'normal' state or a 'not-normal' state. However, the size of the chosen sets rises from small, 1000 samples, to large, samples. The size of the large set is comparable to the amount of performance data which is routinely measured from a network (typically examples being 15Mbytes every 15 minutes for an ATM network [5]). These four data sets would therefore enable us to determine which of the algorithm was best at rapidly, and accurately, determining information from large, multivariate, small class data, namely network performance data. The following paragraph describes in more detail the precise nature of the four data sets.

**Data Set 1** is 'waveform' data which contains 10000 samples, 3 classes and 21 attributes.

**Data Set 2** is 'quadraped' data which contains 40000 samples, 4 classes and 72 attributes.

**Data Set 3** is 'Adult ECG' data which contains 1034 samples, 2 classes and 60 attributes.

**Data Set 4** is 'Umbilical Acid-Base' which contains 1134 samples, 2 classes and 15 attributes.

## 6 Training & Testing

The algorithms were evaluated using a block-wise random two-fold cross-validation methodology. In this methodology half of the data is picked at random, and is used to train the algorithm. When the algorithm is trained the other half of the data is used to evaluate the performance of the algorithm on 'unseen' data. This process is then repeated with the two data subsets swapped.

Tables 1-4 show the results obtained for each algorithm during its classification of Data Sets 1-4.

From the tables it can be seen that C4.5 is both quick and accurate over the whole range of data sets, especially the data set 2, 15 Mb of data, which is the data set similar to the size, and number of attributes, of network performance data commonly collected every 15 minutes for an ATM network [5].

It can also be seen, from the tables, that the variations of the k-NN algorithm have also performed equally well (accurately and quickly) on data sets 1, 3 and 4, but appear to have difficulty with data set 2. This appears to be due to the storage requirements of the algorithm, since in its basic form the algorithm stores all the examples in the training set in internal memory and then compares each novel test example against the stored cases - for this data set, approximately 40 Mb of memory is required for this storage.

The RBF neural network algorithm also performs well, although it also fails on data set 2.

However, it can be seen that the oc variations become increasingly slower in performing classification as the size, and number attributes, of the data sets increases. An interesting feature of this algorithm is that the accuracy of classification appears to increase with this algorithm as the size of data set increases.

## 7 Conclusions

Our initial investigations suggest that the C4.5 algorithm could prove to be the best algorithm to use in order to both accurately, and rapidly, mine multi-set, multivariate small class performance data, and also perform this task over the smaller size of data set. However, k-NN also look promising if the methods of comparison could be changed. In fact, the k-NN algorithm appears to be the most suitable Data Mining algorithm to use for providing the basis for pro-active system management because, although its speed is less than that of C4.5 it is more accurate for classification. In addition, the k-NN algorithm will be more adaptable to parallelisation, and hence increase its classification speed without impinging on its accuracy. In fact, both C4.5 and k-NN could be used as the currently stand to aid in pro-active system management since the maximum classification time of 204s is less than the time measurement intervals of 25 minutes. Classification of network performance data

using either algorithm would allow up to 14 minutes of time to be used in making knowledge-based changes to network configuration, before the next measurement was taken. However an increased speed of classification would enable the network provider to gather increased profits.

## 8 Future Work

We intend to investigate the degree by which parallel implementation of the k-NN algorithm and, if possible, the C4.5 algorithm can increase the speed of classification. Further work will also investigate how the classification accuracy depends on the degree of parity of the classes of the data set.

Plymouth University acknowledge the support of Nortel Technology Ltd (Europe) who are funding the project from which these results are drawn.

**Table 1 : Data Set 1 Accuracy and Speed of Classification**

Algorithm	Accuracy (%)	Time(s)
C4.5	76.54	59
CN2	68.56	1401
K-NN1	77.19	203
K-NN3	80.46	204
K-NN5	82.23	205
OC1	83.25	8623
OC2	80.16	740
OC3	77.23	66
RBF	85.96	152

**Table 2 : Data Set 2 Accuracy and Speed of Classification**

Algorithm	Accuracy(%)	Time(s)
C4.5	100.00	111
CN2	99.98	10037
K-NN1	99.99	9215
K-NN3	99.99	9219
K-NN5	100	9219
OC1	100	2837
OC2	100	854
OC3	100	255
RBF	100	3535

**Table 3 : Data Set 3 Accuracy and Speed of Classification**

Algorithm	Accuracy(%)	Time(s)
C4.5	65.28	11
CN2	63.44	80
K-NN1	62.38	11
K-NN3	67.70	11

K-NN5	65.67	12
OC1	63.54	236
OC2	66.54	82
OC3	70.21	15
RBF	67.89	69

**Table 4 : Data Set 4 Accuracy and Speed of Classification**

Algorithm	Accuracy(%)	Time(s)
C4.5	75.13	3
CN2	69.75	21
K-NN1	70.55	4
K-NN3	74.16	4
K-NN5	75.57	4
OC1	76.10	171
OC2	78.48	36
OC3	78.75	4
RBF	78.92	12

### References

- [1] GARRISON, R.: 'The bt traffic management system', *Brit. Tele- com. Eng.*, 1992, **10** (3) pp. 222-229.
- [2] AUSTIN, G.D.: 'Unlocking the value of performance monitoring data', *Telephony*, 1994, **November** pp. 48-52.
- [3] CLAPP, G.H.: 'LAN interconnects across smps', *IEEE Network Manag.*, 1992, **5** (5) pp. 25-32.
- [4] GERLA, M.: 'Flow control: a comparative study', *IEE. Trans. Comm. Flow Control*, 1980, **28** (4) pp. 553-574.
- [5] BURN-THORNTON, K.E., CATTRALL, D.M. & SIMPSON, A.: 'Functions for data mining in atm networks', 4<sup>th</sup> IFIP Conference on ATM Networks, 4<sup>th</sup> IFIP '96, July 1996, Ilkely, UK, pp. 64 -67.
- [6] LAUFMANN, S.C.: 'Towards agent-based software engineering', *IEE Proceedings - Software Engineering*, 1997, **144** (1) pp. 38 - 50.
- [7] BURMEISTER, M., HADDADI, A. & MATYLIS, G.: ' Applications of multi-agent systems in traffic and transportation', *IEE Proceedings - Software Engineering*, 1997, **144** (1) pp. 351 - 61.
- [8] SHEARER, C.: 'User-driven data mining application', UNICOM Seminar on Data Mining, July 1995, London, pp. 69 - 92.
- [9] PETERSEN, J.: 'Business applications of statistics for data mining', UNICOM Seminar on Data Mining, July 1995, London, pp. 157 - 165.
- [10] ENAS, G.G. & CHOL, S.C.: 'Choice of the smoothing parameter and efficiency of the k-nn neighbour classification', *Comput. Math. Applic.*, 1986, **12A**, pp. 235-244.
- [11] MICHIE, D., SPIEGELHALTER, D. and TAYLOR, C. , eds.: 'Machine Learning, neural and statistical classification' (Ellis Horwood, New York, 1994) 1<sup>st</sup> edn.
- [12] BREIMAN, L., FRIEDMAN J., OLSHEN, R. & STONE, C.: 'Classification and regression trees' (Wadsworth, California, 1984) 1<sup>st</sup> edn.
- [13] QUINLAN, R.: 'C4.5: Programs for machine learning' (Morgan Kaufmann, San Mateo, California, 1993) 1<sup>st</sup> edn.
- [14] CLARK, P. & NIBLETT T.: 'The CN2 induction algorithm', *Machine Learning*, 1988, **3** pp.261-28.
- [15] LEONARD, J., KRAMER, M. & UNGAR, L.: 'Using radial basis functions to approximate a function and its error-bounds', *IEEE Transactions Neural Networks*, 1992, **3** (4), pp. 624-626.
- [16] MURTHY, S., KASIF, S. & SALZBERG, S.: 'A system for induction of oblique decision trees', *Journal of Artificial Intelligence Research*, 1994, **2** pp. 1-32.