Data Mining Fault Information Reports for Prediction

Ramesh K. Rayudu
Transpower NZ Ltd,
New Zealand.

Ajay Maharaj
Transend Networks Pty Ltd,
Australia.

C-fACS
Lincoln University
New Zealand
Ramesh.Rayudu@transpower.co.nz
Phone: +64 3 3253623
Fax: +64 3 325 3839

Abstract

The paper discusses an application of datamining techniques for analysis of fault information data of transmission network in New Zealand. The analysis is mainly aimed to provide necessary prediction knowledge of a fault after an event. This prediction could be a reason for fault, faulted equipment, estimated loss of supply and estimated time for restoration. Our data used for this paper contains transmission faults of approximately 342 events spanning over three years.

In this paper, we introduce the general datamining procedure and discuss its application to analyse the fault information reports. Our solution methodology, pre-processing data selection, analysis algorithms are also discussed. Finally we present and analyse some of the results.

INTRODUCTION

Information technology has developed rapidly over the last three decades contributing towards the formation of data warehouses. The growth of this 'soft' data in computers is now leading towards the so-called 'knowledge' era. To analyse the data and identify useful information, several data analysis techniques have been proposed.

Data mining is one such process of data analysis that could be employed in extracting valid, previously unknown, and ultimately comprehensible information from large databases and use it to make crucial decisions [1]. It involves the application of several statistical and soft-computing techniques to identify meaningful relationships between data. Since the advent of SCADA systems in electric power transmission control, a large amount of data is acquired and stored in data warehouses. Several online tools

Such as state estimation, contingency analysis, and security analysis use this data. However, the data can also be used for other analysis that would assist in better management and control of power systems.

Our motivation behind this research is to device a datamining strategy to analyse power system fault information reports (FIRs). FIRs are reports that are prepared after extensive analysis of and unplanned event that occurred on the power system. Each event is analysed by engineers at a later stage and the conclusions are recorded in the reports. Since considerable amount of resources are spent on analysing and preparing these reports, our aim was to use this documented knowledge and information to provide a reasonable prediction of information in real-time to system co-ordinators when an event occurs in the future.

Data mining has been successfully proven to be effective in applications such as on-line intruder detection [2], rule discovery in alarm processing [3] and prediction of black-outs [4]. In this paper we discuss our experience of applying datamining to predict the unplanned outage information.

DATA MINING PROCESS

The general datamining process has five stages [1]:

- *Data selection* is a process where the required data is identified and gathered into a central repository.
- *Data Cleansing* is a process where the data is pre-processed and possible errors and dubious values are eliminated.
- *Feature extraction* is a process where all the attributes that are deemed to be 'interesting' are selected.

- *Model learning* involves the application of datamining techniques to discover the 'patterns'.
- *Model analysis and evaluation* is a process of visualising the 'mined' information and evaluation of the results based on the identified relationships and accuracy of prediction output.

Based on the above stated stages our datamining process has the generalised input-output patterns as shown in the following Figure 1. The input attributes depict the data available for input in real-time and the output attributes are prediction values required. The process of our research towards the illustrated mechanism is discussed below.

Data Selection

Initially we have selected around 470 events spanning over three years. Each event is described in terms of 42 attributes. Some attributes have no values in them and some have more than one value. Some attributes also include English sentences that provide descriptions of the event. Since FIR data was not configured for real-time usage and datamining, we had to go through multiple iterations of data pre-processing to extract meaningful dataset for datamining.

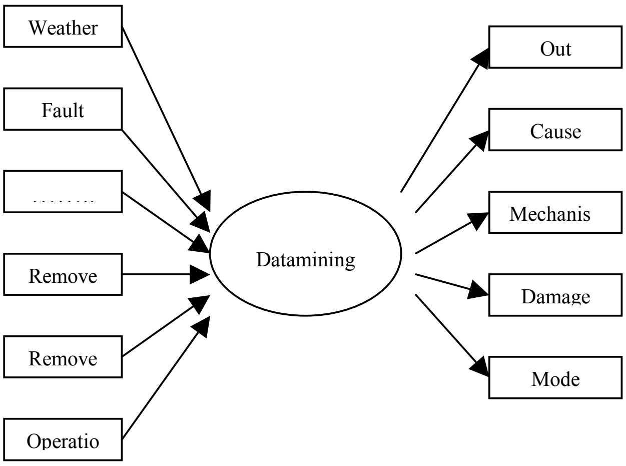

Figure 1. General input-output patterns for FIR datamining.

Data Cleansing

Our main motivational attribute of the FIR data is the time it took to restore the supply after an event. Hence, all the events that had 'no' loss-of-supply were deleted from the data. This has reduced the events to 394 out of which 52 events were used for testing and the rest were used in datamining process.

Secondly, since the original dataset had a combination of discrete, continuous and linguistic data, we also had to normalise the data based on our FIR analysis knowledge. One example is the following Table 1 that depicts the classification of loss-of-supply duration.

Table1. Loss of supply time classification

Duration time(min)	Class label
1- 5	1
6-10	2
11-20	3
21-50	4
51-100	5
101-500	6
Over 500	7

.

Feature Extraction

This process involves the identification of appropriate input and output attributes. We have selected five output attributes. These are outage duration classification, cause of the fault or event (rain snow, wind, and pollution), mechanism of the fault (error by staff, error by contractor or wilful damage), damage analysis (equipment damage and injury to humans or animals), and finally the mode of the fault (red-earth, phase to phase etc.). For inputs selection, we have analysed the data based on our experience and then by performing sensitivity analysis on the data. By this process we could delete 12 attributes that had 10% or less effect on the output prediction.

342 events with 30 attributes were used for training phase of our datamining process.

Model learning

We have used three algorithms for datamining the training data. These algorithms are:

- SSV decision tree: Decision trees algorithms are usually quite efficient and therefore worth attempting first on the data. They are preferred over other kinds of adaptive models when a logical description of collected data is required. Decision trees are built in a recurrent process by splitting the feature space into two or more parts. The splits aim for the best separation of objects belonging to different classes. At each step, they use information gain to decide separation rules.

- K nearest neighbours algorithm (KNN): This algorithm assumes all instances correspond to points in the n-dimensional space R^N. The nearest neighbours of an instance are defined in terms of the standard Euclidean distance. KNN remembers all training data and selects most similar vectors at the moment it is asked to make a prediction.
- Feature Space Mapping (FSM): FSM adopted here is a neuro-fuzzy network is based on multidimensional separable functions. The main idea is simple: components of the input and output vector define features, and combinations of these features define objects in the feature spaces. These objects are described by the joint density probability of the input/output data vectors using a network of properly parameterised transfer functions.

Model Analysis and Evaluation

The following table, Table 2, displays the accuracy of the prediction output results. For this analysis, we have chosen the threshold accuracy rate of above 50% as the average percentage of accurate prediction of these events by humans is around 50%. It is evident from the above results that no single algorithm performed well against all the outputs. Exception is the 'mode' of fault identification where all the algorithms performed reasonably well.

There are several reasons to explain the above accuracy results. Some of these are:

- Data pre-processing: It can be estimated that some of the accuracy can improved if we re-iterate the data pre-processing phase to identify other information requirements.
- Physical location of operation: The circuit breaker and equipment trips have been generalised in the experiment. It is possible that the geographical location of the equipment information could contribute towards better accuracy.
- Extensive code inputs: Each attribute has an extensive list of sub-categories. For example, the cause and mechanisms have 274 different descriptions. Since we had a set time limit for the experiment, we could not code all the descriptions.
- Some important information such as the English sentence input needs to be input in a machine-readable format.

CONCLUSION

We have presented a process to perform datamining on electric power fault information reports. We have followed the five stages of datamining and then analysed the results. We have used three different algorithms for analysis. The prediction (outputs) parameters were Outage duration in minutes, Cause and Mechanism of the fault, Expected damage to the equipment, and Mode of the fault. No single algorithm had performed well against all outputs. All algorithms for all predictors had an average accuracy of about 66.7%. We analysed some accuracy contributors for further research.

REFERENCES

1. I. H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, CA, 2000.

2- K. Ilgun, R. A. Kemmerer, and P. A. Porras, "State transition analysis: A rule-based intrusion detection approach.", *IEEE Transactions on Software Engineering*, 21(3):181-199, March 1995.

3- K. Hatonen, M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, "Rule Discovery in Alarm Databases", Report number C. No. C-1996-7, University of Helsinki, 1996.

4- P. Geurts & L. Wehenkel, "Early Prediction of Electric Power System Balackouts by Temporal Machine Learning", Technical Report WS-98-07, AAAI Press, 1998.

Table 2. Accuracy Results.

	KNN		SSV		FSM	
	Training	Test	Training	Test	Training	Test
Out duration	57.4%	37.7%	55.8%	21.4%	**63.9%**	**58%**
Cause	100%	46.8%	**76%**	**74%**	87%	35%
Mechanism	100%	45.7%	**78.7%**	**74.2%**	82.1%	48.8%
Damage	79.8%	42.8%	79%	42.3%	**86%**	**42.8%**
Mode	**74%**	**71%**	**65%**	**64%**	**97%**	**57%**