Using a Hybrid Neural/Expert System for Data Base Mining in Market Survey Data

Victor Ciesielski and Gregory Palstra
Department of Computer Science
Royal Melbourne Institute of Technology
GPO Box 2476V
Melbourne Vic 3001 Australia
vc@cs.rmit.edu.au

Abstract

This paper describes the application of a hybrid neural/expert system network to the task of finding significant events in a market research data base. Neural networks trained by backward error propagation are used to classify trends in the time series data. A rule system then uses these classifications, knowledge of market research analysis techniques and external events which influence the time series, to infer the significance of the data. The system achieved 86% recall and 100% precision on a test set of 6 months of survey data. This was significantly better than could be achieved by a system using linear regression together with a rule system. Both systems were able to perform analysis of the test data in under 5 minutes. The manual analysis of the same data took a human expert over four working days.

Introduction

Data mining has been described as the process of finding 'nuggets' of information or knowledge in large data bases. Many traditional machine learning approaches are being re-examined for their effectiveness in finding nuggets(Fayyad & Uthurusamy 1995; Piatetsky-Shapiro & Frawley 1991). There is currently no 'universal data base mining engine' and it is unlikely that one will be developed in the foreseeable future. However, what does appear to be within reach is a set of guidelines about what kinds of techniques are good for what kinds of data bases. In this paper we present an approach that we believe will be very useful for data bases of time series data.

A Hybrid Neural/Expert Architecture

It has been well recognized that artificial neural networks and expert systems are complementary in that where one approach is weak the other is strong(Becraft, Lee, & Newell 1991; Ciesielski & Ho 1993). There has been considerable work in devising architectures which realize the benefits of both approaches. One architecture which has had considerable success (Becraft, Lee, & Newell 1991; Ciesielski & Ho 1993; Scheinmakers & Tourtetzky 1990) involves using network(s) as sub-components of expert systems as suggested by figure 1. Here the feed-forward component of a previously trained network is used to determine values of some of the variables used by the expert system. The network is trained off-line using appropriate examples and the expert system is constructed by an appropriate knowledge engineering process.

This architecture is well suited to problems which have a pattern recognition or classification component and a reasoning component.

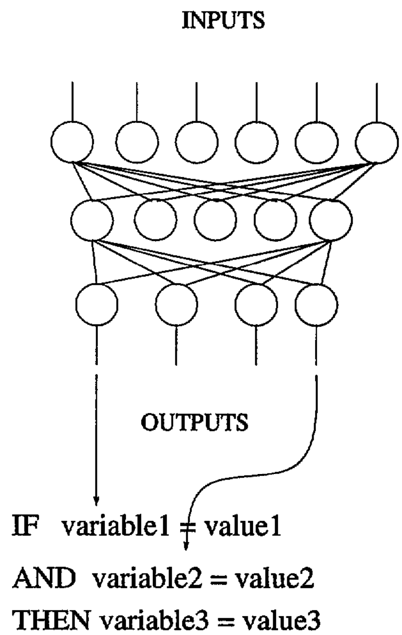

Figure 1: A Hybrid Neural/Expert Architecture

Goal

The goal of this work was to determine whether a search engine, based on the architecture described above, could be effectively used for finding significant

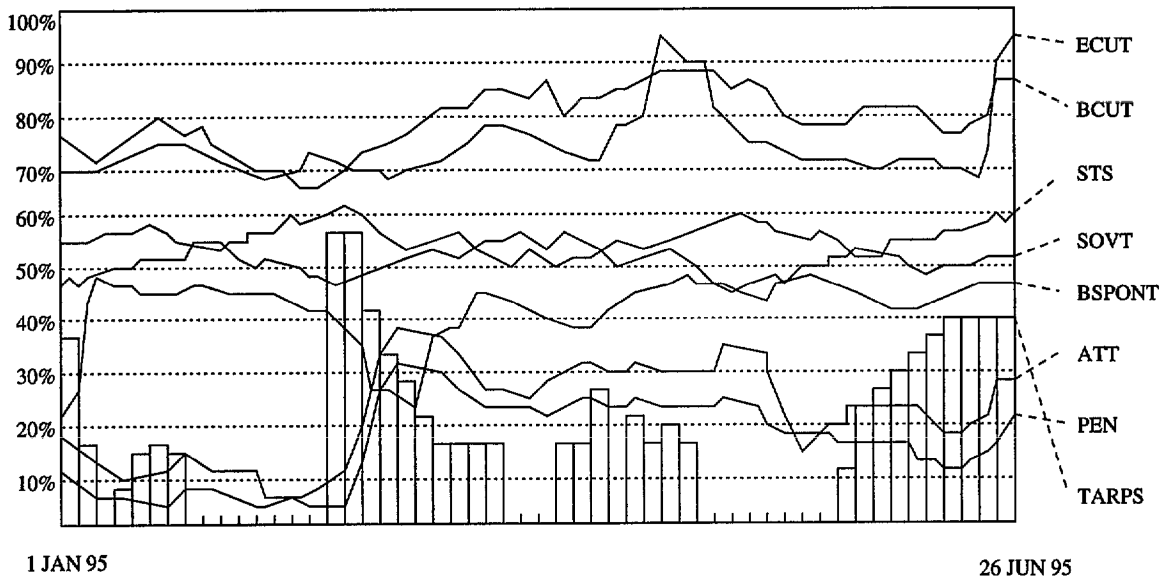

Figure 2: Graphical representation of survey data

patterns in market survey time series data.

Analysis of Market Survey Data

In our domain of market survey data the 'golden nuggets' to be found pertain to whether or not an advertising campaign is effective. Survey data analysts spend long hours analysing graphs of survey data looking for situations where an advertising campaign is not being effective. An analyst may need to look at several hundred graphs a day. Naturally the advertisers want to know about problems as soon as possible so that appropriate action can be taken. A sample of the data on which decisions are based is shown in figure 2. Much of this data comes from surveys carried out at shopping centres where the above data is collected about a product and its main competitors. The variables in the figure are (Sutherland 1993):

Executional cut through (ECUT) How well the advertisement for a product is being recalled amongst the advertising about other products.

Branded cut through (BCUT) How well the advertising for the brand is being recalled.

Attitudinal share (ATT) What brand people are likely to buy next.

Short term share (STS) The last brand bought for this type of product.

Penetration (PEN) Has the person ever bought the product.

Brand Spontaneity (BSPONT) Recall, not recognition, of the brand name.

Share of voice (SOVT) The percentage share of this product's television advertising in relation to its competitors.

Client Advertising (TARPS) The amount of television advertising for a product.

Ad executions (ADEX) The number of advertisements sent to air for the product in one week.

Net message take out (NETMESS) The percentage of people who derived the intended message from the advertising.

The graph in figure 2 ends at 26-Jun-95. On 27-Jun-95 this data would be examined to determine whether there is a problem. The analyst needs to determine the trends in selected variables prior to 26-Jun-95, consider their inter-relationships and draw a conclusion about the effectiveness of the advertising. Typically the same data will be available for dozens of products causing a 'data flood' problem. Our implementation focuses on finding, as soon as possible, the situations where an advertising campaign is not working. One such result is shown in figure 3. Note that a key aspect of the decision making is the classification of time series segments into the categories shown in table 1.

The data in figure 2 consists of rolling averages, in which the last n data points are averaged to get the value for the current week as follows:

$$p' = \frac{1}{n} \sum_{i=0}^{n-1} p_{current-i}$$

Rolling averages tend to smooth out fluctuations in the survey data, however, if n is too high the variations of interest can be obliterated. A series of preliminary experiments revealed that a 4 week rolling average was

For the four weeks leading up to the 23-Mar-1995 the advertising is cutting through but failing to register the (correct) message. Strengthen the message in the commercial, or create a new commercial.

Attribute	Classification	Conf
Executional cut through	SMALL-RISE	67%
Branded cut through	SMALL-RISE	89%
Short term share	FLAT	98%
Attitudinal Share	FLAT	73%
Penetration	BIG-FALL	78%
Unprompted brand awareness	SMALL-RISE	74%
In store promotions	SEQUENCED	
Client advertising	REMAINED	
Share of voice	89%	
Net message take out	40%	
Ad executions	3	
CONCLUSION	AD-MSG-BAD	(67%)

Figure 3: Results of an Analysis

best for this application. Co-incidentally, the analysts believe that 4 weeks is the appropriate time period in which to detect an advertising problem.

The analyst's task is to combine together the shapes and relationships of the curves with other knowledge to determine significance. For example it would be expected that sales of ice creams and soft drinks should rise in a heat wave and that a manufacturer should be alerted if this did not occur. It would be expected that the sales of a product would rise shortly after the commencement of an advertising campaign and an alert should be generated if this did not occur. It is possible for advertising for a product to actually raise the sales of a competitor's product and such situations need to be detected as soon as possible. A particularly tricky and very important aspect of determining advertising effectiveness is the problem of finding 'turning points' as soon as possible. A turning point is a change of trend, from downwards to upwards for example, and is of great importance to an advertiser spending heavily on a television advertisement.

For the purpose of this paper a 'golden nugget' or significant event is a situation where an advertising campaign is not working as expected.

The above examples suggest that the analysis task has a pattern classification component (interpreting the graphs) and a reasoning component (using knowledge of events that affect the time series) and we believed that a major part of it could be automated using a hybrid neural/expert system of the kind described earlier.

Development of the Neural/Expert Search Engine

We developed two neural/expert hybrid systems. Our first system directly used the model suggested in figure 1. A single network was used for all variables and the generated classifications were used by the expert system to determine whether that particular four week period contained a situation that should be reported. This system was unable to detect subtle differences between variables as the experts were doing so we then developed a system with separate networks for each of the eight variables. The performance of this second system was considerably better.

One Network For All Variables

The neural network used was a 4-6-6 network trained by backward error propagation. The six output units corresponded to the categories in table 1. We had successfully used such networks previously for a similar task in an intensive care ventilator monitor(Hayes & Ciesielski 1991).

String	Symbolic Meaning
SRISE	A small rise in the trend
BRISE	A large rise
SFALL	A small fall
BFALL	A large fall
FLAT	No significant change
INCONCLUSIVE	No classification possible

Table 1: Symbolic output from the neural network

The 'INCONCLUSIVE' class was used when the analysts could not place a four week pattern into one of the other five classes. This situation typically arose when there were large fluctuations in values in consecutive weeks. We introduced this class rather than force the experts to select one of the existing classes. Also, some attributes, like 'short term share', approach a threshold which effects the significance of changes in that attribute's value. A jump of four percent in 'short term share' when its value is low may be insignificant, but the same jump, when the attribute's value is close to fifty percent, is a very significant change. This situation does not affect the training of the networks, but does need to be covered in the rule system.

	1-out-of-5	Gaussian-hill
Big Rise	0.0	0.5
Small Rise	1.0	1.0
Flat	0.0	0.5
Small Fall	0.0	0.0
Large Fall	0.0	0.0
Inconclusive	0.0	0.0

Table 2: Output Encodings for 'small-rise'

The training data consisted of 600 four week intervals of survey data classified into the categories shown in table 1. A *1-out-of-5* coding was used. This resulted in rather slow training and an unacceptable error rate. We then adopted a 'Gaussian-hill' approach to coding the outputs(Pomerleau 1991). The coding for 'small rise' is shown in table 2. In this domain 'small rise' is intermediate between 'big rise' and 'flat' and has some of the characteristics of both. This is reflected in the Gaussian-hill encoding. Both training times and accuracy improved significantly (table 3).

	1-out-of-5	Gaussian-hill
6-fold X-validation	74%	80%
Mean Squared Error	0.00043	0.00030
Epochs	3500	2000

Table 3: 1-out-of-5 vs 'Gaussian-hill' output encodings

Separate Network for Each Variable

As testing proceeded, it became clear that we were still not accurately emulating the experts' classifications, with nearly 20% of test cases being incorrectly classified. Analysis of the errors revealed that within some value ranges, several attributes were seldom being classified correctly at all. Further discussion with the analysts revealed that not all attributes were analysed the same way. The same set of points could be interpreted differently, depending on which attribute they were representing. It was clear the use of a single neural network for all attributes would never provide adequate classifications and that a separate network for each variable was needed. Table 4 shows the training statistics for the single network and the average statistics for the eight dedicated neural networks. This group of networks trained at a much faster rate and performed on test cases, with an average error rate of 8%.

	One Network For All Variables	One Network Per Variable
6-fold X-validation	80%	92%
Mean Squared Error	0.00043	0.00030
Epochs	2000	1500

Table 4: Single Neural Network vs Dedicated Neural Networks

The errors that still occur seem to be a direct result of treating the attributes independently. For some attributes, like 'attitudinal share' and 'short term share' it is the relationship between their trends that is the significant factor determining the classification of these attributes. If 'short term share' is classified as flat and 'attitudinal share' is also flat, but is moving very slightly upwards and away from 'short term share', then 'attitudinal share' should be classified as rising, even though it would normally be classified as 'flat' for such a small rise. However this judgement is very difficult to make on four weeks of data. Longer term relationships are not captured by our current networks. We are currently determining whether the benefits of capturing these relationships are justified by the development costs.

The Expert System

The expert system was developed using the spiral development methodology(Giarratano & Riley 1989). In this approach the system is developed from an initial prototype by successive cycles of knowledge acquisition, coding and evaluation. We commenced the knowledge acquisition using the method of familiar tasks (Hoffman 1987). At this step the key aspects of survey data analysis were identified as was the goal of finding ineffective advertising campaigns. Subsequently we used the method of structured interviews(Hoffman 1987) which involved bi-weekly meetings with the analysts.

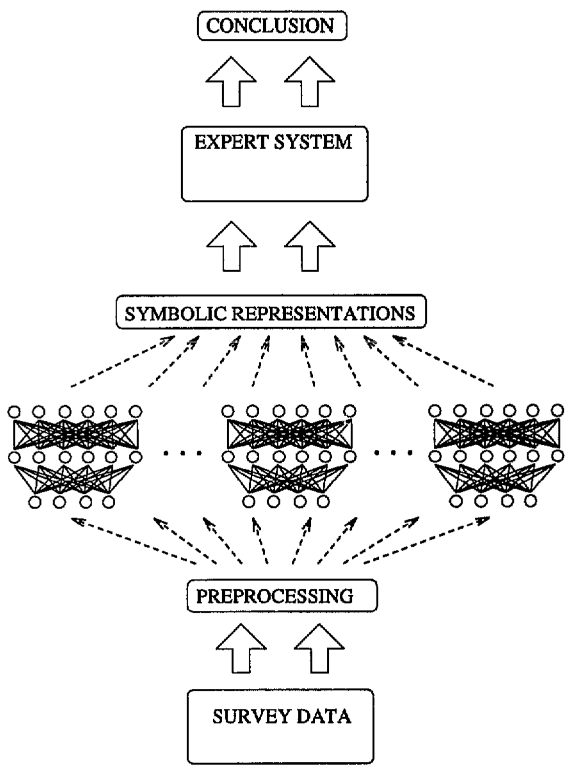

Figure 4: Hybrid System Architecture

The architecture of the developed system is shown in figure 4. The process of consultation involves the user specifying which product and survey data to examine, and the start and end date of the examination period. The hybrid system retrieves the required data from the database and applies the required rolling average. The data is then broken into four-week intervals. The eight attributes from each four-week interval are presented to their respective neural networks for classification. The symbolic representations from the neural networks, and any other facts that are required by the expert system, such as the share of voice value and the state of advertising for the interval, are added to the fact base. The rule interpreter is then invoked and a conclusion is inferred using backward chaining.

Mining Experiments and Results

Our task of finding the significant events in large amounts of survey data is analagous to the information retrieval task of finding relevant documents in a (usually very large) document collection in response to a query. Thus we have used *recall* and *precision*, the standard information retrieval measures(Salton, Allan, & Buckley 1994), to assess system performance. For our purposes recall and precision are:

$$recall = \frac{Number\ of\ significant\ events\ retrieved}{Actual\ number\ of\ significant\ events}$$

$$precision = \frac{Numberof\ significant\ events\ retrieved}{Number\ of\ events\ retrieved}$$

In general, there is a tradeoff between recall and precision. For example, requiring a higher recall results in lower precision as attempts to retrieve more significant events also result in the retrieval of more false alarms.

Testing the system presented us with some major problems. We knew from the structured interviews that the analysts were impressed with the system's ability to find significant events. However for a formal evaluation we needed to compare the system's performance with the experts on a significantly large data base. Simply taking chunks of several months of data for a number of products was not really adequate since many of these events are rare and there can be many months between occurrences. Furthermore determining the actual number of significant events in a data set is a very laborious and time consuming process since several analysts need to carefully analyse the data and agree on the significant events.

As a compromise between the need to have a large data base and the very high cost of preparing it we constructed a test set, equivalent to six months of survey data, by splicing a number of significant events (that were not involved in the training phase) into a number of places in an otherwise dull time series. The test data base was equivalent to 104 four week segments. The results are shown as the first line in table 5. The system found 25 out of 29 significant events and none of these events were false positives.

We compared the performance of our hybrid engine with a system containing a linear regression preprocessor in place of the neural network. The regression preprocessor fits a least squares line to the previous four weeks of data. The slope is then classified into

	Recall	Precision
Hybrid System	25/29 (86%)	25/25 (100%)
Expert system with linear regression	21/29 (72%)	20/21 (95%)

Table 5: Recall and precision results neural network and regression systems

the categories shown in table 1. Developing this system required determining optimal thresholds for the five categories of interest. Some of our experiments for doing this are shown in table 6.

Flat	Small Rise	Large Rise	Correct
$0 \leq m < 10$	$10 \leq m < 20$	$20 \geq m$	65%
$0 \leq m < 15$	$15 \leq m < 25$	$25 \geq m$	75%
$0 \leq m < 17$	$17 \leq m < 27$	$27 \geq m$	81%
$0 \leq m < 20$	$20 \leq m < 30$	$30 \geq m$	70%
$0 \leq m < 25$	$25 \leq m < 45$	$45 \geq m$	60%

Table 6: Regression Development

The ranges were determined in a number of knowledge acquisition cycles with the analysts. From table 6 it can be seen that the best gradient range set tested was the third row. The 81% correct classification rate here is virtually the same as the 80% achieved by the single neural network.

The second line of table 5 gives the recall and precision results for this system on the test data base, showing the hybrid system to be clearly superior. The results in the table are perhaps slightly misleading since we have compared a neural system with 8 networks with a regression system containing only a single, global, slope classification for all variables. However the analysts were not willing to repeat the tiresome process of slope determination another 8 times, particularly when they already had a system they were happy with.

While we are not in a position to compare the performance of the hybrid system with the best possible regression system, we can conclude that the development time for the hybrid system is shorter and is much preferred by the analysts. Developing the regression system requires the additional step of determining classification thresholds for the slope (table 7). This requires several iterations of frustrating structured interviews with the experts, since adjusting a threshold to make an improvement in one value typically makes another one worse. We had experienced a similar situation with the development of an intensive care ventilator monitor(Hayes & Ciesielski 1991).

The recall results indicate that the system is currently too strict in its definition of a significant event. In normal operations we would want the system to find all significant events at the cost of a moderate number of false alarms.

Neural/Expert System	Regression/Expert system
1. Collect and Classify Data	1. Collect and Classify Data
2. Train networks	2. Determine prototypical regression line
	3. Search for best thresh-holds
3. Develop rule system	4. Develop rule system
4. Put trained networks into search engine	5. Put thresh-holds into expert system

Table 7: Comparison of development steps

Conclusions

The main goal of this project was to investigate the effectiveness of a hybrid neural/expert system for the task of extracting significant events from market research data. We have developed such a system and evaluated its performance on a significant market survey data base. The hybrid neural/expert system was more accurate than a regression/expert system that could be built within the resource limitations of the project. The system promises considerable improvements in efficiency over the current manual approach. It was able to analyse the test data in under five minutes, whereas the manual analysis of the same data took an expert analyst over four working days. However the system needs some fine tuning before it is ready to be deployed. We particularly need to improve the recall at the expense of the precision, that is we would tolerate a moderate number of false positives to ensure that no significant event is missed.

The key insight that drove the design of the neural/expert search engine was viewing the analysis task as pattern classification followed by symbolic reasoning with domain knowledge. We believe that neural/expert hybrid systems of this kind will be useful for other data bases, particularly data bases of time series.

Acknowledgements

We thank Kingsley Winikoff, Max Sutherland, Roland Seidel and Dale Chant of Marketmind Technologies Pty Ltd for generous contributions of their time and expertise to this work and Angeline Pang for her work on the first prototype.

References

Becraft, W. R.; Lee, P. L.; and Newell, R. B. 1991. Integration of neural networks and expert systems for process fault diagnosis. In *The 12th International Joint Conference on Artificial Intelligence*, volume 2, 824–831. Morgan Kaufman Publishers Inc.

Ciesielski, V., and Ho, T. 1993. The synergistic combination of neural and symbolic computation in the recognition of words degraded by noise. In Rowles, C.; Liu, H.; and Foo, N., eds., *Proceedings of the Sixth Australian Joint Artificial Intelligence Conference*, 131–136. Melbourne: World Scientific.

Fayyad, U., and Uthurusamy, R., eds. 1995. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95).* Menlo Park, California: AAAI Press.

Giarratano, J., and Riley, G. 1989. *Expert Systems, Principals and Programming.* PWS-Kent Publishing Comapny, Boston.

Hayes, S., and Ciesielski, V. 1991. A comparison of an expert system and a neural network for respiratory system monitoring. Technical Report 92/1, RMIT.

Hoffman, R. R. 1987. The problem of extracting the knowledge of experts from the perspective of experimental psychology. *AI Magazine* 8(2).

Piatetsky-Shapiro, G., and Frawley, W. J., eds. 1991. *Knowledge Discovery in Databases.* Menlo Park, California: AAAI Press and The MIT Press.

Pomerleau, D. A. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation* 3(1):88–97.

Salton, G.; Allan, J.; and Buckley, C. 1994. Automatic structuring and retrieval of large text files. *Communications of the ACM* 37(2):97–105.

Scheinmakers, J. F., and Tourtetzky, D. S. 1990. Interfacing a neural network with a rule based reasoner for diagnosing mastitis. In Erlbaum, L., ed., *Proceedings of the International Joint Conference on Neural Networks. Volume 2*, 487–490. Morgan Kaufman Publishers Inc.

Sutherland, M. 1993. *Advertising and the Mind of the Consumer, What works, what doesn't and come.* Sydney: Allen and Unwin.

Weiss, S. M., and Kulikowski, C. A. 1991. *Computer Systems That learn.* San Mateo, Ca: Morgan Kaufman.