Data mining and Neural Networks from a Commercial Perspective

Portia A. Cerny
Data Analyst & Modeller
Aim Proximity
Auckland, New Zealand
Student of the Department of Mathematical Sciences
University of Technology Sydney, Australia
portiac@aimproximity.co.nz, portia.a.cerny@uts.edu.au

Abstract

Companies have been collecting data for decades, building massive data warehouses in which to store it. Even though this data is available, very few companies have been able to realize the actual value stored in it. The question these companies are asking is how to extract this value. The answer is Data mining.

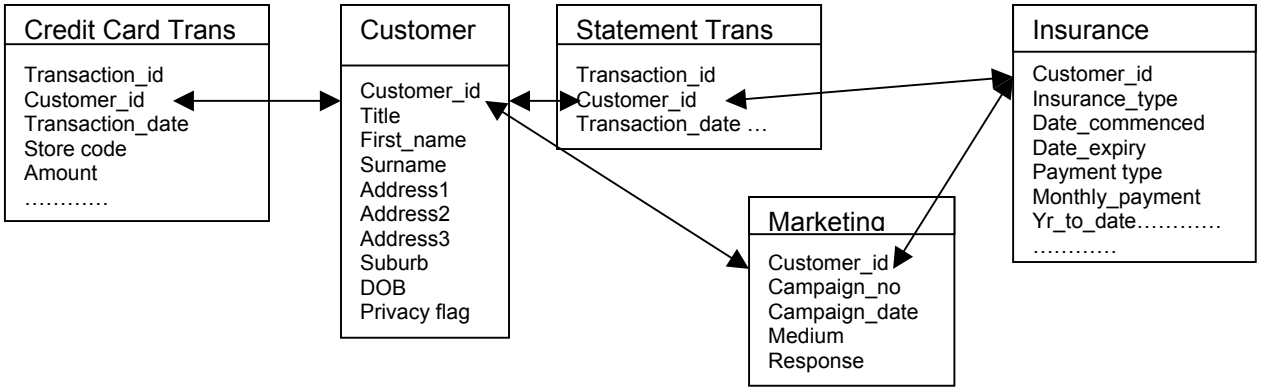
There are many technologies available to data mining practitioners, including Artificial Neural Networks, Regression, and Decision Trees. Many practitioners are wary of Neural Networks due to their black box nature, even though they have proven themselves in many situations.

In our current research we are attempting to compare the aforementioned technologies and determine if Neural Networks outperform more traditional statistical techniques. This paper is an overview of artificial neural networks and questions their position as a preferred tool by data mining practitioners.

1 INTRODUCTION

Data mining is the term used to describe the process of extracting value from a database. A data-warehouse is a location where information is stored. The type of data stored depends largely on the type of industry and the company. Many companies store every piece of data they have collected, while others are more ruthless in what they deem to be "important". The data-warehouse is usually a relational database, that is, a database with multiple tables that are linked together using a number of unique record identifiers or keys such as customer_id in the example below.

Figure 1: Part of a relational database for a financial institution



Consider the following example of a financial institution failing to utilize their data-warehouse. Income is a very important socio-economic indicator. If a bank knows a person's income, they can offer a higher credit card limit or determine if they are likely to want information on a home loan or managed investments. Even though this financial institution had the ability to determine a customer's income in two ways, from their credit card application, or through regular direct deposits into their bank account, they did not extract and utilize this information.

Another example of where this institution has failed to utilise its data-warehouse is in cross-selling insurance products (e.g. home, contents, life and motor vehicle insurance). By using transaction information they may have the ability to determine if a customer is making payments to another insurance broker. This would enable the institution to select prospects for their insurance products. These are simple examples of what could be achieved using data mining.

Four things are required to data-mine effectively: high-quality data, the "right" data, an adequate sample size, and the right tool.

1. **Quality**: A database is only as useful as the data it contains. For example, if half of the fields have greater than 80% missing values, or there are discrepancies in the data, it is difficult to extract value.

   A common problem for most companies is duplication. It is possible to have data on the same person without knowing that Janice Brown from Ponsonby is also Janice Smith from Mt Albert.

   The preferred data is machine-recorded data because human error is reduced and consistency is maintained. Credit card data is an example of machine-recorded data.

2. **The "right" data**: What is the point of an insurance company collecting information regarding a person's favourite ice-cream flavour? It is important that appropriate and relevant data is captured.

   An important piece of data that is often not captured is age. Knowing the age of a person can drive marketing campaigns. An insurance company is more likely to offer life insurance to a 35 year-old than to an 85 year-old. Similarly for banking institutions, offering credit cards to 16 year olds is not viable, as most do not have a regular source of income. Knowing a person's age is one of the first steps in developing a "personality" for the customer.

3. **Sample size**: For any statistical analysis a sufficiently large dataset is essential. There is however the question of how much data is enough. One client of Acxiom Australia Pty Ltd, Golden Casket Lottery Corporation Ltd, has a loyalty programme of over 1 million customers who generated 1.2 billion transactions over a two-year period. It would be impossible to use the entire sample due to the current computing limitations; hence the use of a good random sampling algorithm was particularly useful here (refer to Numerical Recipes in C, Chapter 7 [7]).

4. **The right tool:** There are many tools available to a data mining practitioner. These include decision trees, various types of regression and neural networks.

1.1 Decision Trees

A decision or classification tree represents a series of rules that can be expressed in a spoken language and converted easily into programming languages such as SQL [1]. For a full discussion on decision trees please refer to Berry and Lindoff chapter 12 [1].

A common use for decision trees is in private health insurance. The procedure to accept or reject an application for membership requires strict rules that a health insurance company (HIC) must follow. If an application is rejected the HIC must provide an adequate reason for rejection to avoid claims of discrimination. By using a decision tree it is possible to determine which customers to accept or reject given certain non-discriminatory criteria.

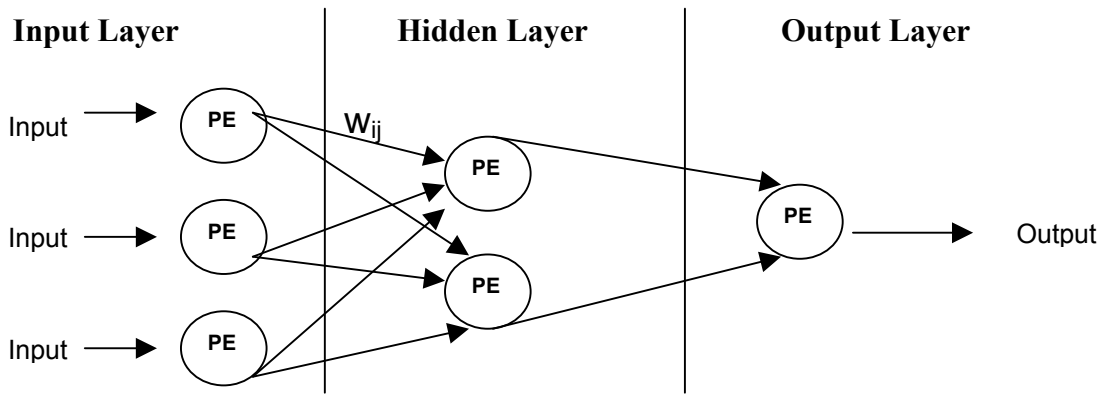
1.2 Various types of Regression

Regression is the analysis, or measure, of the association between a dependent variable and one or more independent variables. This association is usually formulated as an equation in which the independent variables have parametric coefficients that enable future values of the dependent variable to be predicted. Two of the main types of regression are: linear regression and logistic regression. In linear regression the dependent variable is continuous and in logistic it is either discrete or categorical. For logistic regression to be used, the discrete variable must be transformed into a continuous value that is a function of the probability of the event occurring, (Olivia Parr Rud) [10]. Regression is used for three main purposes: (1) description, (2) control and (3) prediction [6]. Neter et alia provides a detailed discussion on different types of regression.

1.3 Neural Networks

There are two main types of neural network models: supervised neural networks such as the multi-layer perceptron or radial basis functions, and unsupervised neural networks such as Kohonen feature maps. A supervised neural network uses training and testing data to build a model. The data involves historical data sets containing input variables, or data fields, which correspond to an output. The training data is what the neural network uses to "learn" how to predict the known output, and the testing data is used for validation. The aim is for the neural networks to predict the output for any record given the input variables only.

Figure 2:
Example of a simple feedforward neural network



One of the simplest feedforward neural networks (FFNN), such as the one in Figure 3, consists of three layers: an input layer, hidden layer and output layer. In each layer there are one or more processing elements (PEs). PEs are meant to simulate the neurons in the brain and this is why they are often referred to as neurons or nodes. A PE receives inputs from either the outside world or the previous layer. There are connections between the PEs in each layer that have a weight (parameter) associated with them. This weight is adjusted during training. Information only travels in the forward direction through the network - there are no feedback loops.

The FFNN according to Tchaban et alia [14] is "ideally suited", although not restricted to, problems with "many inputs, and a single output as it has a direct structure and can be easily trained". The simplified process for training a FFNN is as follows:

1. Input data is presented to the network and propagated through the network until it reaches the output layer. This forward process produces a predicted output.
2. The predicted output is subtracted from the actual output and an error value for the networks is calculated.
3. The neural network then uses supervised learning, which in most cases is backpropagation, to train the network. Backpropagation is a learning algorithm for adjusting the weights. It starts with the weights between the output layer PE's and the last hidden layer PE's and works backwards through the network.
4. Once backpropagation has finished, the forward process starts again, and this cycle is continued until the error between predicted and actual outputs is minimised.

Ripley [8,9] provides a thorough discussion on neural networks and is recommended reading.

Neural networks are becoming very popular with data mining practitioners, particularly in medical research, finance and marketing. This is because they have proven their predictive power through comparison with other statistical techniques using real data sets. Fadalla et alia [3] examined forty papers where which neural networks have been used. In most cases neural networks were compared to other statistical techniques, which included various types of discriminant analysis and regression. The majority were based on stock market forecasting and bankruptcy prediction. Fadalla et alia provides a summary of the results from ten of these papers and concludes that in most comparative studies neural networks outperform statistical-econometric models.

Table 1:
Taken from Lisboa et alia, *Business Applications of Neural Networks* pp xvi [5]

Advantages of Neural Networks	**Disadvantages of Neural Networks**
High Accuracy: Neural networks are able to approximate complex non-linear mappings.	**Poor Transparency:** Neural networks operate as "black boxes".
Noise Tolerance: Neural networks are very flexible with respect to incomplete, missing and noisy data.	**Trial-and-error design:** The selection of the hidden nodes and training parameters is heuristic.
Independence from prior assumptions: Neural networks do not make *a priori* assumptions about the distribution of the	**Data hungry:** Estimating the network weights requires large amounts of data, and this can be very

data, or the form of interactions between factors.	computer intensive.
Ease of maintenance: Neural networks can be updated with fresh data, making them useful for dynamic environments.	**Over-fitting:** If too many weights are used without regularisation, Neural network become useless in terms of generalisation to new data.
Neural network overcome some limitations of other statistical methods while generalizing them.	There is no explicit set of rules to select the most suitable Neural network algorithm.
Hidden nodes, in supervised Neural networks can be regarded as latent variables.	Neural networks are totally dependent on the quality and amount of data available.
Neural networks can be implemented in parallel hardware.	Neural networks may converge to local minima in the error surface.
Neural networks performance can be highly automated, minimizing human involvement.	Neural networks lack classical statistical properties. Confidence intervals and hypothesis testing are not available.
Neural networks are especially suited to tackling problems in non-conservative domains.	Neural network techniques are still rapidly evolving and they are not yet robust

Another possible advantage is the additional parameters created by the hidden nodes. The topic of another paper should be "Is it fair to compare two models where the number of parameters are significantly different?"

According to Berry & Lindoff, there are two main drawbacks to using a neural network. The "first is the difficulty in understanding the models they produce". Their "black box" nature means that unlike a decision tree, there are no easily extractable rules that can be used to show how the classification or prediction was made. The second is "their particular sensitivity to the format of incoming data: different data representations can produce different results; therefore, setting up the data is a significant part of the effort of using them" [1]. The second drawback, as well as some of the disadvantages listed in the Table 1 are drawbacks of most modelling techniques and should not be restricted to neural networks.

2 Review of literature reporting Neural Network Performance.

There are numerous examples of commercial applications for neural networks. These include; fraud detection, telecommunications, medicine, marketing, bankruptcy prediction, insurance, the list goes on. Most literature relating to the use of neural networks, including some of those listed below, are pro neural networks. The reason for this imbalance is possibly because people are unlikely to publish cases where using neural networks have not been successful.

The following are examples of where neural networks have been used. Where possible we have chosen examples where neural networks have been compared to another statistical method such as a decision trees or logistic regression. The main methods of comparison are lift charts, rating accuracy on training data, prediction accuracy, and misclassification rate on testing data.

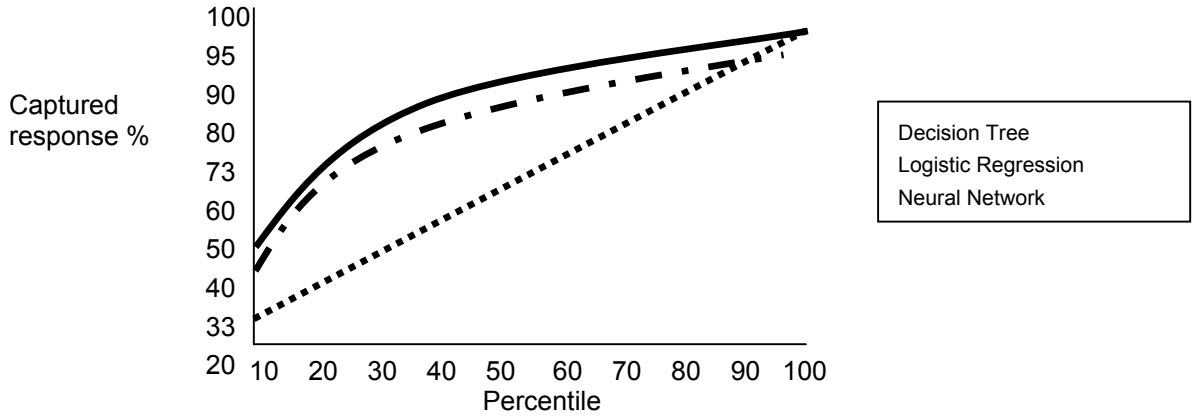
2.1 Insurance Industry (Kate Smith et alia) [12]

The article, "*An analysis of customer retention and insurance claim patterns using data mining: a case study"* [12], compares three types of models used to classify insurance policy holders as likely to renew or terminate their policies. This information is used to gain an understanding of customer retention patterns. The models, which were built using SAS Enterprise Miner Software, were: a decision tree (chi-squared test with .02 significance level, max number branches =1, max depth of tree = 10.), logistic regression, and a neural network (3-layer feed-forward neural network with 29 inputs, 25 hidden neurons and a single output neuron). In this situation the neural network marginally outperformed the logistic regression and both outperformed the decision tree.

The entire data set is comprised of 20,914 policyholders. A subset of 4,183 policyholders was used to test the models. Figure 4 is a lift chart which illustrates the results from the subset. The *x*-axis represents 100% of the sample (4,183 policy holders) and has been sorted according to their likelihood of terminating as predicted by the model. Of these 4,183 policyholders, *m* are known to have terminated their insurance policy. The *y*-axis represents the percentage of captured response. If all *m* people are identified then 100% of response has been captured. By mailing 0% of the sample they will capture 0% response. By mailing 100% of the sample they will capture 100% of response. However it is not usually cost effective to mail the entire sample. In this case if only 10% of the sample is to be examined, the neural network would discover almost 50% of all terminating cases compared to 40% with logistic regression and 28% with the decision tree.

The following lift chart (Figure 4) is of our own making and represents the results given in the article. The chart "depicts the percentage of the total terminations that would be discovered if only a certain percentage of policyholders where examined". [12].

Figure 3:
Lift Chart representing percentage of total terminations



2.2 Predicting Bankruptcy (Rick Wilson et alia) [16]

Rick Wilson et alia, in the article titled "Bankruptcy prediction using neural networks" , compares the capabilities of neural networks and multivariate discriminant analysis to predict a firm going bankrupt. They use three different compositions of training data

and testing data (50/50, 80/20, 90/10). They provide a number of result tables comparing performance, looking at the reduction-in-error of classifications, predictive validity of classifications, training composition effect on classifications and some others, all of which demonstrate that the neural networks outperformed discriminant analysis in prediction accuracy. Overall they are impressed with the capabilities of neural networks, and feel that research is still in its infancy. They consider their very promising results as representing the "lower bound" of what neural networks can actually do.

2.4 Neural Networks and Operations Research (Eva Wilppu) [15]

Eva Wilppu cites some examples where neural networks have been used optimisation problems. El Ghaziri (1991) used a neural network for routing problems and found that the results from the self-organizing map (eg unsupervised neural network such as Kohonen map) were as good as provided by any other algorithm. (Fujimura et alia (1997)) studied the travelling salesman problem also using a self-organizing map and found that this specific neural network was able to solve the travelling salesman problem in a shorter time than standard operations research applications. Lozano et alia (1998) used a self-organizing map to solve a location-allocation problem and unlike many of the other papers listed, they were not overtly in favour of the NN. They state however, that from their results neural networks can be considered an alternative to classical operations research applications.

2.5 Neural Networks and Oncology (Schwarzer et alia) [11]

Not all applications of neural networks have been successful. Based on a critical review of forty-three articles, Schwarzer et alia detail seven common mistakes made when using or reporting on the use of neural networks in oncology. The first six mistakes, as listed below, should be considered when discussing neural networks in general, the seventh is specific to survival data. The seven common mistakes are:

1. Mistakes in estimation of misclassification probabilities.
2. Fitting of implausible functions.
3. Incorrectly describing the complexity of a network.
4. No information on complexity of the network.
5. Use of inadequate statistical competitors.
6. Insufficient comparison with statistical methods.
7. Naive application of artificial neural networks to survival data.

Their conclusion is that "we are far away [from] demonstrating the value of neural network for medical research in oncology". From Table 2, one of the disadvantages of using a neural network is because they are data hungry. It seems clear that one of the main reasons for poor results in oncology is because the training sets were so small, ranging from 45 to 1600 records. Oncology provides an example of where neural networks have not been successfully used and in most cases it is due to a lack of data and how the experiment was conducted.

2.6 Golden Casket [4]

In 2000 Acxiom Australia Pty Ltd built three predictive models for Golden Casket Lottery Corporation Ltd in Queensland, Australia. Acxiom is a data services company where the author worked until July 2000. Golden Casket has a loyalty programme with over 1 million members (700,000 active each quarter). Using a random sample of 100,000 members, three models were built: a future value model and a cross-sell model and behavioural segmentation model. Both the future value, and cross-sell models were developed using a neural network. In this case the neural network tool used was HNC's Marksman, a feed forward neural network using backpropagation with one hidden layer.

The future value model predicts the worth a customer will have at a later date. In this case six months of historical data was used to predict the value of a customer one month in advance. This time lag of one month was allowed for the model to be run, the campaign to be developed and the mail to be delivered. Models with a two and three month lag time were created, however the larger the time lag the lower the performance of the model.

Figure 4:
Results from future value model [4]

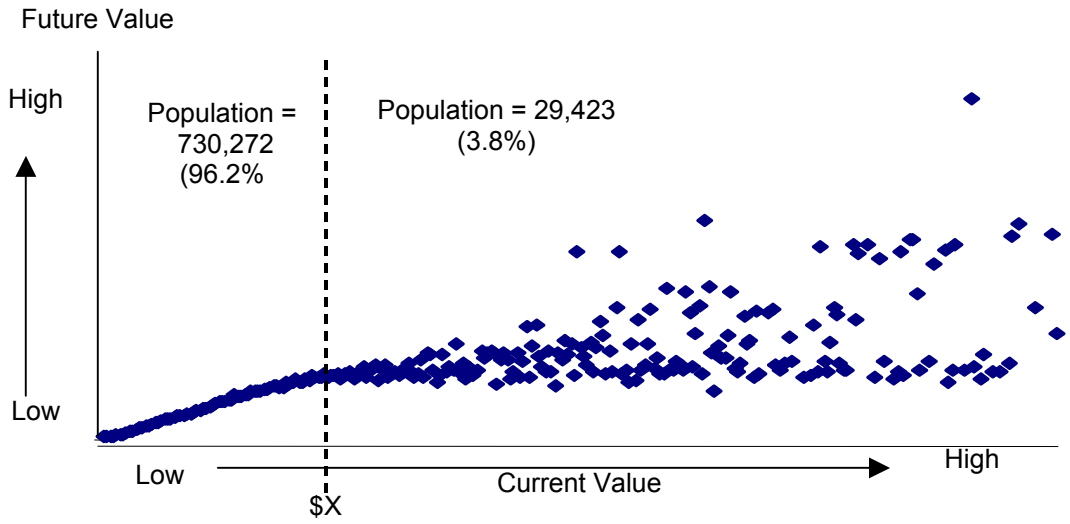


Figure 4 represents the current value of a customer on the x-axis, and the actual future value of the customer on the y-axis. $X is a dollar value and splits the customer base into two groups. The first group contains 96.2% of the population. Here where the current value is less than $X the model predicts with 96.8% accuracy. For customers in the second group (3.8% of population), where customer value is greater than $X, the model predicts with 23% accuracy. For any customer the model predicts future value with 71% accuracy.

The cross-sell model predicts the game a customer is most likely to take up. At the time the model was built there were nine different games a customer could play.. The neural network cross-sell model gives each game a probability (the sum of all game probabilities =1). The highest probability score represents what the customer is probably already playing. Using the probability scores for all games a cross-sell opportunity or reinforcement strategy was chosen. Table 2 represents some possible customer results from the model.

A reinforcement strategy was chosen for Martha is because she is a one game person. Her probability scores for the other games are almost zero, which means that the

likelihood of her taking up another game is very small. In this case Golden Casket will encourage her to continue to play Game 1. Tom also receives reinforcement but on three games. This is because he has a good spread of games. Harry plays Game 4, and is quite likely to play Game 5 as well, hence Golden Casket will choose to cross sell Game 5.

Table 2:
Cross-sell model probability scores

Customer	Game1	Game2	Game3	Game4	Game5	Game6	Game7	Game8
Martha	90%	0	0	5%	2%	0	3%	0
Tom	33%	33%	33%	0	0	0	0	0
Harry	0	0	0	65%	32%	0	0	0

The models were used to drive Golden Casket's direct marketing campaigns. In order to test the power of the models customers were selected using a random sampling tool, the results from the neural network models, and a previously developed segmentation. The results were as follows:

- The random sample had a 1-2% response rate
- The segmentation model had a 10-12% response rate
- The neural network prediction models had a 20-30% response rate

Golden Casket not only benefited from using the models in their direct marketing campaign, but also in their call centre, where each customer screen contains a "cross-sell" field which the customer service representative could act upon when talking to that customer.

One point they made, which is also commonly stated in most literature on neural network, was "neural networking is very powerful in predicting future behaviour but it doesn't tell you why". Golden Casket use these neural network models because it has enabled them to triple incremental revenue from direct marketing and make significant savings in other areas.

Conclusion

There is rarely one right tool to use in data mining; it is a question as to what is available and what gives the "best" results. From the author's commercial experience at Acxiom and Aim Proximity, the results from neural networks are very promising. Many articles, in addition to those mentioned in this paper, consider neural networks to be a promising data mining tool. In most cases they perform as well or better than the traditional statistical techniques to which they are compared.

Resistance to using these "black boxes" is gradually diminishing as more researchers use them, in particular those with statistical backgrounds. As software companies develop more sophisticated models with user-friendly interfaces the attraction to neural networks will continue to grow.

Acknowledgements

We would like to extend out particular thanks to Golden Casket for graciously allowing us to publish their results and Andrew Mason, Peter Wright and Murray Smith for their encouragement.

References:

[1] Berry, J. A., Lindoff, G., *Data Mining Techniques*, Wiley Computer Publishing, 1997 (ISBN 0-471-17980-9).

[2] Bradley, I., *Introduction to Neural Networks*, Multinet Systems Pty Ltd 1997.

[3] Fadalla, A., Lin, Chien-Hua. "An Analysis of the Applications of Neural Networks in Finance"*, Interfaces* 31: 4 July- August 2001 pp 112-122.

[4] Golden Casket, *AC Pan Pacific Conference Presentation 2000.*

[5] Lisboa, P. J. G, Edisbury, B., Vellido, A., *Business Applications of Neural Networks*, World Scientific, Singapore, USA, UK, (ISBN 981-02-4089-9).

[6] Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W., *Applied Linear Regression Models* 3rd Ed. 1996, Irwin, USA (ISBN 0-256-08601-X).

[7] *Numerical Recipes in C: The art of Scientific Computing*, Cambridge University Press, 1992 (ISBN 0-521-43108-5).

[8] Ripley, B. D., Can Statistical Theory Help Us Use Neural Networks Better? *Interface* 97. *29th Symposium of the Interface: Computing Science and Statistics.*

[9] Ripley, B. D., *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996, UK (ISBN 0-521-46086-7).

[10] Rud, O. P, *Data Mining Cookbook*, Wiley Computer Publishing 2001, USA (ISBN 0-471-38564-6).

[11] Schwarzer, G., Vach, W., Schumacher, M., "*On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology"*, URL citeseer.nj.nec.com/44173.html.

[12] Smith, K. A., Willis, R. J., Brooks, M., "An analysis of customer retention and insurance claim patterns using data mining: a case study*", Journal of the Operations Research Society* (2000) Vol 51, pp 532-541.

[13] Steinbeg, D. *CART™ Classification and Regression Trees: A Tutorial*, Salford Systems.

[14] Tchaban, T., Griffin, J. P., Taylor, M. J., *A comparison between single and combined Backpropagation Neural Networks in the Prediction of Turnover*, url: http://www.citeseer.nj.nec.com/188602.html.

[15] Wilppu, E., "Neural Networks and Logistics" *TUCS Technical Report No 311* April 1999, Turku Centro for Computer Science (ISBN 952-12-0558-X).

[16] Wilson, R. L., Sharda, R., "Bankruptcy prediction using neural networks", *Decision Support Systems* 11 (1994) pp545-557.