Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 51A (2012) 11141-11144

ISSN: 2229-712X

Data mining with artificial neural network in a scenerio of a bank

Sunil Kumar Arora
MCA Department, IIMT Engineering College, Meerut-250001,U.P., India.

ARTICLE INFO

Article history:
Received: 9 February 2012;
Received in revised form:
13 October 2012;
Accepted: 27 October 2012;

Keywords
Data Mining,
Artifical Neural network,
Decision tree.

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. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.
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 neural networks and questions their position as a preferred tool by data mining practitioners.

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". Consider the following example of a financial institution failing to utilize their data-warehouse. Another example of where this institution has failed to utilize its data-warehouse is in cross-selling insurance products (e.g. home, 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. There are many tools available to a data mining practitioner. These include decision trees, various types of regression and neural networks. 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.

Data mining techniques

Induction Techniques

Induction techniques develop a classification model from a set of records -- the training set of examples. The training set may be a sample database, a data mart, or an entire data warehouse. Each record in the training set belongs to one of many predefined classes, and an induction technique induces a general concept description that best represents the examples to develop a classification model. The induced model consists of patterns that distinguish each class. Once trained, a developed model can be used to predict the class of unclassified records automatically. Induction techniques represent a model in the form of either decision trees or decision rules. These representations are easier to understand, and their implementation is more efficient than those of neural network or genetic algorithms.

Neural Networks

Neural networks constitute the most widely used technique in data mining. They imitate the way the human brain learns and use rules inferred from data patterns to construct hidden layers of logic for analysis. Neural networks methods can be used to develop classification, regression, link analysis, and segmentation models. A neural net technique represents its model in the form of nodes arranged in layers with weighted links between the nodes. There are two general categories of neural net algorithms: supervised and unsupervised.
Supervised neural net algorithms such as Back propagation (Rumelhart, Hinton, and Williams, 1986) and Perceptron require predefined output values to develop a classification model. Among the many algorithms, Back propagation is the most popular supervised neural net algorithm. Back propagation can be used to develop not only a classification model, but also a regression model.

Unsupervised neural net algorithms such as ART (Carpenter and Grossberg, 1988) do not require predefined output values for input data in the training set and employ self-organizing learning schemes to segment the target data set. Such self-

Tele:
E-mail addresses: mtechsunilarora@gmail.com

organizing networks divide input examples into clusters depending on similarity, each cluster representing an unlabeled category.

Genetic Algorithms

Genetic algorithms are a method of combinatorial optimization based on processes in biological evolution. The basic idea is that over time, evolution has selected the "fittest species." For a genetic algorithm, one can start with a random group of data. Afitness function can be defined to optimizing a model of the data to obtain "fittest" models.

Logistic Regression

Logistic regression is a special case of generalized linear modeling. It has been used to study odds ratios (e pj', j = 1, 2,..., k as defined in the following), which compares the odds of the event of one category to the odds of the event in another category, for a very long time and its properties have been well studied by the statistical community.

Clustering

Clustering techniques are employed to segment a database into clusters, each of which shares common and interesting properties. The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster. Clusters are also created to support the other types of DM operations, e.g. link analysis within a cluster.

Associated Discovery

Given a collection of items and a set of records containing some of these items, association discovery techniques discover the rules to identify affinities among the collection of items as reflected in the examined records.

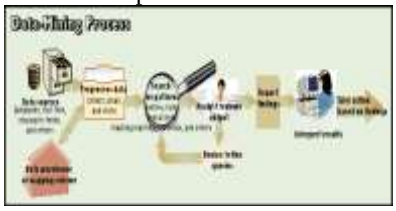
Sequence Discovery

Sequence discovery is very similar to association discovery except that the collection of items occurs over a period of time. A sequence is treated as an association in which the items are linked by time. When customer names are available, their purchase patterns over time can be analyzed

Visualization

A picture is worth thousands of numbers! Visual DM techniques have proven the value in exploratory data analysis, and they also have a good potential for mining large databases. Visualizations are particularly useful for detecting phenomena hidden in a relatively small subset of the data. This technique is often used in conjunction with other DM techniques: features that are difficult to detect by scanning numbers may become obvious when the summary of data is graphically presented. Visualization techniques can also guide users when they do not know what to look for to discover the feature. Also, this technique helps end users comprehend information extracted by other DM techniques. examples of visualization technique s that have been extended to work on large data sets and produce interactive displays

Neural Networks In Data Mining:In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.



Using it as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases is that there is actual manipulation and cross-fertilization of the data helping users makes more informed decisions.

Neural networks essentially comprise three pieces:

The architecture or model; the learning algorithm; and the activation functions. Neural networks are programmed or "trained" to ". . . store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill-defined problems; in summary, to estimate sampled functions when we do not know the form of the functions." It is precisely these two abilities (pattern recognition and function estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining. As data sets grow to massive sizes, the need for automated processing becomes clear. With their "model-free" estimators and their dual nature, neural networks serve data mining in a myriad of ways. Data mining is the business of answering questions that you've not asked yet. Data mining reaches deep into databases.

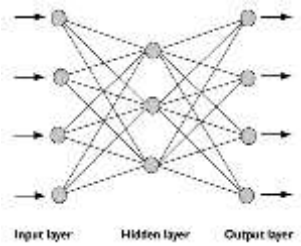
Data mining tasks can be classified into two categories: Descriptive and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea. Predictive data mining allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered form the database.

Data mining models can be categorized according to the tasks they perform: Classification and Prediction, Clustering, Association Rules. Classification and prediction is a predictive model, but clustering and association rules are descriptive models. The most common action in data mining is classification. It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules. Similar to classification is clustering. The major difference being that no groups have been predefined. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have. The next application is forecasting. This is different from predictions because it estimates the future value of continuous variables based on patterns within the data.

Neural networks, depending on the architecture, provide associations, classifications, clusters, prediction and forecasting to the data mining industry. Financial forecasting is of considerable practical interest. Due to neural networks can mine valuable information from a mass of history information and be efficiently used in financial areas, so the applications of neural networks to financial forecasting have been very popular over the last few years. Some researches show that neural networks performed better than conventional statistical approaches in financial forecasting and are an excellent data mining tool. In data warehouses, neural networks are just one of the tools used in data mining.

ANNs are used to find patterns in the data and to infer rules from them. Neural networks are useful in providing information on associations, classifications, clusters, and forecasting. The back propagation algorithm performs learning on a feed-forward neural network. One of the simplest feed forward neural networks (FFNN), such as in Figure, consists of three layers: an

input layer, hidden layer and output layer. In each layer there are one or more processing elements (PEs). PEs is 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 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 back propagation, to train the network. Back propagation 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 back propagation has finished, the forward process starts again, and this cycle is continued until the error between predicted and actual outputs is minimized.

2. ARTIFICIAL NEURAL NETWORKS:

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase

Neural Network Topologies:

Feedforward neural network: The feedforward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers.

Recurrent network: Recurrent neural networks that do contain feedback connections. Contrary to feedforward networks, recurrent neural networks (RNs) are models with bi-directional data flow. While a feedforward network propagates data linearly from input to output, RNs also propagate data from later processing stages to earlier stages.

Training Of Artificial Neural Networks:

A neural network has to be configured such that the application of a set of inputs produces (either 'direct' or via a relaxation process) the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to 'train' the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. We can categorize the learning situations as follows:

• Supervised learning or Associative learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs can be provided by an external teacher, or by the system which contains the neural network (self-supervised).

• Unsupervised learning or Self-organization in which an (output) unit is trained to respond to clusters of pattern within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli.

Reinforcement Learning: This type of learning may be considered as an intermediate form of the above two types of learning. Here the learning machine does some action on the environment and gets a feedback response from the environment. The learning system grades its action good (rewarding) or bad (punishable) based on the environmental response and accordingly adjusts its parameters.

The Back Propagation Algorithm:

Back propagation, or propagation of error, is a common method of teaching artificial neural networks how to perform a given task. The back propagation algorithm is used in layered feed forward ANNs. This means that the artificial neurons are organized in layers, and send their signals "forward", and then the errors are propagated backwards. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the back propagation algorithm is to reduce this error, until the

ANN learns the training data. Summary of the technique:

1. Present a training sample to the neural network.
2. Compare the network's output to the desired output from that sample. Calculate the error in each output neuron.
3. For each neuron, calculate what the output should have been, and a scaling factor, how
much lower or higher the output must be adjusted to match the desired output. This is the local error.
4. Adjust the weights of each neuron to lower the local error.
5. Assign "blame" for the local error to neurons at the previous level, giving greater responsibility to neurons connected by stronger weights.
6. Repeat the steps above on the neurons at the previous level, using each one's "blame" as its error.

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. The following are examples of where neural networks have been used.

Accounting

- Identifying tax fraud
- Enhancing auditing by finding irregularities

Finance

- Signature and bank note verification
- Risk Management
- Foreign exchange rate forecasting

- Bankruptcy prediction
- Customer credit scoring
- Credit card approval and fraud detection
- Forecasting economic turning points
- Bond rating and trading
- Loan approvals
- Economic and financial forecasting

Marketing

- Classification of consumer spending pattern
- New product analysis
- Identification of customer characteristics
- Sale forecasts

Human resources

- Predicting employee's performance and behavior
- Determining personnel resource requirements

Design problems:

There are no general methods to determine the optimal number of neurones necessary for solving any problem.

It is difficult to select a training data set which fully describes the problem to be solved.

Solutions To Improve Ann Performance:

Designing Neural Networks using Genetic Algorithms

Neuro-Fuzzy Systems

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. Many articles, in addition to those mentioned in this paper, consider neural networks to be a promising data mining tool Artificial Neural Networks offer qualitative methods for business and economic systems that traditional quantitative tools in statistics and econometrics cannot quantify due to the complexity in translating the systems into precise mathematical functions. Hence, the use of neural networks in data mining is a promising field of research especially given the ready availability of large mass of data sets and the reported ability of neural networks to detect and assimilate relationships between a large numbers of variables. In most cases neural networks 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 moreresearchers use them, in particular those with statistical backgrounds. Thus, 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.

Due to design problems neural systems need further research before they are widely accepted in industry. As software companies develop more sophisticated models with user-friendly interfaces the attraction to neural networks will continue to grow.

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.