Current Development in Artificial Intelligence.
ISSN 0976-5832 Volume 3, Number 1 (2012), pp. 1-8
© International Research Publication House
http://www.irphouse.com

Use of Data Mining & Neural Network in Medical Industry

Ina Kapoor Sharma

Lecturer B.N College of Engineering & Technology
E-mail:-ina_kapoor@yahoo.co.in

Abstract

Artificial intelligence has attracted many researchers in recent years for its usefulness in the medical field. With the increase in research interest, artificial intelligence systems have found widespread usage in the medical domain These applications arise as medical datasets usually contain a high volume of data, and it is very costly for human manual analysis and handling. The intelligent systems posses the ability to process datasets, extract useful information from them, and interpret the data at a much lower cost as compared to the manual handling. Intelligent systems like evolutionary computing, fuzzy logic, and neural networks are just some that many researchers have recently looked into.Medicine has always benefited from the forefront of technology. Technology advances like computers, lasers, ultrasonic imaging, etc. have boosted medicine to extraordinary levels of achievement. Artificial Neural Networks (ANN) is currently the next promising area of interest.Data mining is the research area involving powerful processes and tools that allow an effective analysis and exploration of usually large amounts of data. In particular, data mining techniques have found application in numerous different scientific fields with the aim of discovering previously unknown patterns and correlations, as well as predicting trends and behaviour This paper intends to bring together the most recent advances and applications of data mining and neural networks research in the promising areas of medicine and biology

Keywords: Data Mining, Artificial Neural Networks, Medicine, Medical Industry

Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help medical industry focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer medical questions that traditionally were very time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Mostly people related to medical field collects and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. .

Approaches to Data Mining Problems

The approaches to data mining problems are based on the type of information / knowledge to be mined. This paper emphasis on two important approaches that are Decision tree and Association rules.

The task of **association rule** mining is to search for interesting relationships among items in a given data set. Its original application is on “market basket data”. The rule has the form x->y, where x and y are sets of items and they do not intersect. Each rule has two measurements, support and confidence. Given the user-specified minimum support and minimum confidence, the task is to find, rules with support and confidence above, minimum support and minimum confidence.

Decision Tree Learning, used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models **are classification trees** or regression trees. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications.

In Decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making**.**

Artificial Neural Networks

Artificial Neural networks commonly referred to as **“neural networks”** is a mathematical model based on biological neural system, it an algorithm for optimization and learning based loosely on concepts inspired by research into the nature of the brain.

The brain is a highly complex, nonlinear and parallel computer. It has the capability to organize its structural constituents, known as neutrons, so as to perform certain computations many times faster than the digital computer In the general form neural network is a machine that is designed to model the way in which the brain performs a particular task or function of interest According to **Simon Haykin** " A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects

1. Knowledge is acquired by the network from its environment through a learning process
2. Interneuron connection strengths, known as a synaptic weights are used to store the acquired knowledge"

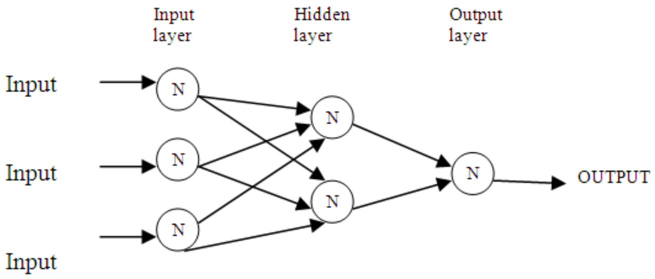

Figure 1: Neural Network

Neural Network Architecture

A neural network can be viewed as s weighted directed graph in which neurons are nodes and directed edges represent connection between neurons neuron network architecture can be classified in three classes:

Single Layer Feed Forward Networks is a layered neuron network in which neurons are organized in the form of layers, in this we have an input layer of source nodes that projects onto an output layer of neurons but not vice versa. This network is strictly a feed forward or acyclic type.

Multilayer Feed Forward Networks are network which has one or more hidden layer and whose computation nodes are correspondingly called hidden neurons. The function of hidden neurons is to intervene between the external input and network output in some useful manner.

Recurrent Networks has at least one feedback loop. It may consist of a single layer of neurons with each neuron feeding its output signal back to the inputs of all the other neurons. Recurrent networks also no hidden neurons

Training Neural Networks

Supervised Learning which incorporates an external teacher, so that each output unit is told what its desired response to input signals ought to be. During the learning process global information may be required. Paradigms of supervised learning include error-correction learning, reinforcement learning and stochastic learning.

An important issue concerning supervised learning is the problem of error convergence, ie the minimization of error between the desired and computed unit values. The aim is to determine a set of weights which minimizes the error. One well-known method, which is common to many learning paradigms, is the least mean square (LMS) convergence.

Unsupervised Learning uses no external teacher and is based upon only local information. It is also referred to as self-organization, in the sense that it self-organizes data presented to the network and detects their emergent collective properties. Paradigms of unsupervised learning are Hebbian learning and competitive learning.

It is said that a neural network learns off-line if the learning phase and the operation phase are distinct. A neural network learns on-line if it learns and operates at the same time. Usually, supervised learning is performed off-line, whereas unsupervised learning is performed on-line

Data Mining Based on Neural Networks

In more technical 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 neural networks 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.

The difference between general data mining and data mining based on neural networks is explained with the help of an example.

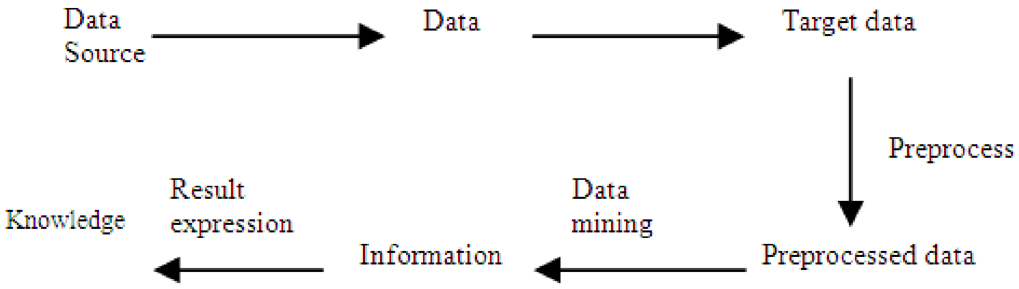

Figure 2: Data Mining without Neural Networks

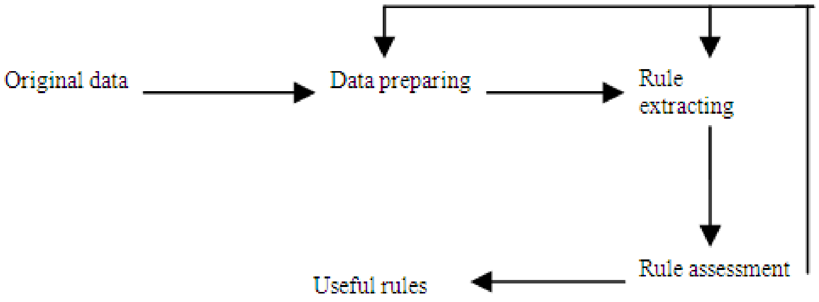

Figure 3: Data Mining Based on Neural Networks

A **genetic Algorithm** (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

Neural networks and genetic algorithms are two techniques for optimization and learning, each with its own strengths and weaknesses. The two have generally evolved along separate paths. However, recently there have been attempts to combine the two technologies.

1. That returns a rating for each chromosome given to it.
2. A way of initializing the population of chromosomes.
3. Operators that may be applied to parents when they reproduce to alter their genetic composition. Included might be mutation, crossover (i.e. recombination of genetic material), and domain-specific operators.
4. Parameter settings for the algorithm, the operators, and so forth.

Given these Five Components, a Genetic Algorithm Operates According To the Following Steps

1. The population is initialized, using the procedure in C3. The result of the initialization is a set of chromosomes as determined in C2.
2. Each member of the population is evaluated, using the function in C I . Evaluations may be normalized; the important thing is to preserve relative ranking of evaluations.
3. The population undergoes reproduction until a stopping criterion is met. Reproduction consists of a number of iterations of the following three steps
   a) One or more parents are chosen to reproduce. Selection is stochastic, but the parents with the highest evaluations are favored in the selection. The parameters of C5 can influence the selection process.
   b) The operators of C4 are applied to the parents to produce children. The parameters of C5 help determine which operators to use.
   c) The children are evaluated and inserted into the population. In some versions of the genetic algorithm, the entire population is replaced in each

cycle of reproduction. In others, only subsets of the population are replaced.

Decision Tree Learning is mostly used in statistics, data mining and machine learning. These technologies use decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are **classification trees** or **regression trees**. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.

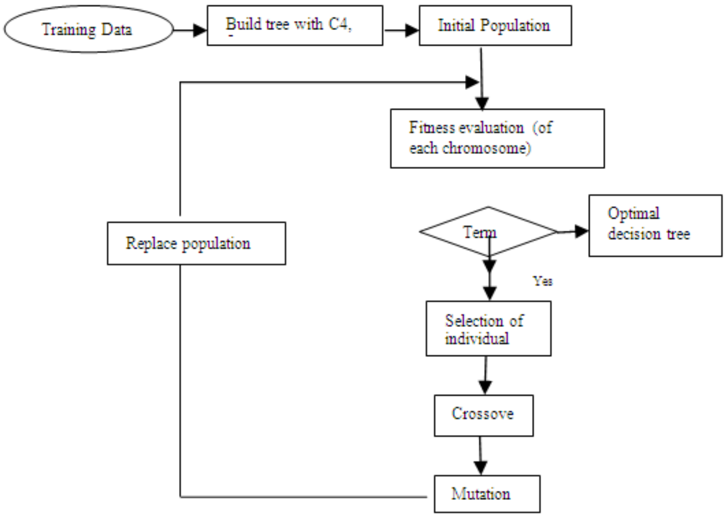

Figure: 4 Decision Tree

Our Genetic Algorithm

1. Chromosome Encoding
2. Evaluation Function
3. Initialization Procedure
4. Operators
5. Mutation
6. Crossover
7. Gradient

Mutation: A Mutation operator takes one parent and randomly changes some of the entries in its chromosome to create a child

Crossover: A Crossover operator takes two parents and creates one or two children containing some of the genetic material of each parent
Gradient: A gradient operator takes one parent and produces a child by adding to its entries a multiple of the gradient with respect to the evaluation function.

GA Procedure

```
Procedure GA
Begin
Initialize population;
Evaluate population members;
While termination condition not satisfied do
Begin
Select parents from current population;
Apply genetic operators to selected parents;
Evaluate offspring;
Set offspring equal to current population;
End
End
```

Once the fitness values of all individuals of the current population have been computed, the GA begins to generate next generation as follows:

1. Choose Individuals according to Rank Selection method
2. Use two point crossover to exchange genes between parents to create offspring.
3. Perform a bit level mutation to each offspring.
4. Keep two elite parents and replace all other individuals of current population with offspring

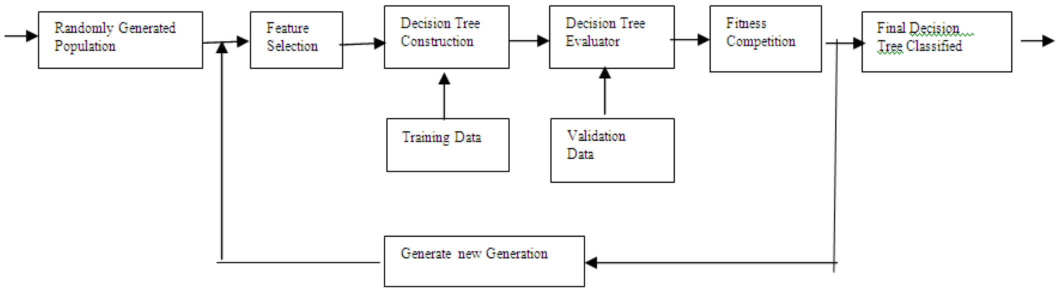

Figure 5: GA/Decision Tree Hybrid

Advantages of using Genetic Algorithms with Decision Trees

1. It is very simple to understand and easy to interpret
2. Important insights can be generated even with little hard data
3. It uses white box model

4. It can be combined with other decision techniques
5. It is modular approach which is separate from application
6. It supports multi-objective optimization
7. It is very good for noisy environment
8. It is inherently parallel so it can be easily distributed
9. It always provide an answer and gets better with time
10. It can easily exploit previous or alternate solutions
11. It has very flexible building blocks for hybrid applications

Conclusion

Classification is an important problem in the rapidly emerging field of data mining. Many problems in business, science, industry, and medicine can be treated as classification problems. Owing to the wide range of applicability of ANN and their ability to learn complex and nonlinear relationships including noisy or less precise information, neural networks are well suited to solve problems in biomedical engineering

The genetic algorithm and decision tree hybrid was able to outperform the decision tree algorithm without feature selection.

We believe that this improvement is due to the fact that the hybrid approach is able to focus on relevant features and eliminate unnecessary or distracting features. This initial filtering is able to improve the classification abilities of the decision tree.

The algorithm does take longer to execute than the standard decision tree; however, its non-deterministic process is able to make better decision trees. The training process needs only to be done once. The classification process takes the same amount of time for the hybrid and non-hybrid systems.

References

[1] Amor, N. B., Benferhat, S., and Elouedi, Z. Naive Bayes vs decision trees in intrusion detection systems

[2] Baker, J.E. Adaptive selection methods for genetic algorithms

[3] Portia A Cerny Data Mining and Neural Networks from Commercial Point of View

[4] Dr Yashpal Singh, Alok Singh Chauhan Neural networks in Data Mining

[5] Huang, Z., Pei, M., Goodman, E., Huang, Y., and Li, G. Genetic algorithm optimized feature transformation: a comparison with different classifiers.