Investigations of Neural Networks for Data Mining Procedures

Sam Eberspacher

November 2, 2007

Abstract

Data mining is a process by which data of different types is evaluated to determine its value in a given search. The human brain is incredibly good at this process because it can determine the type of data, process the language, etc. Neural networks are a computers attempt to simulate what the brain does by creating artificial neurons which evaluate small pieces of information. By using a neural network to mine data, the results of the process will be more accurate than a straight forward procedure. In order to increase the effectiveness of the network, feedback should be provided to the network to allow it to "grow". The growth of the network would be done by adding, deleting, or modifying neurons engaged in the network.

1 Introduction

1.1 Purpose

This project is especially useful to search engines or any project that requires a significant amount of data. The data used by this approach would be more accurate than a simple algorithm because the network will learn from previous experiences. This allows the results to become even more accurate as more data is retrieved and more feedback is given to the network.

1.2 Scope of Study

The development of a neural network will be the most dominant feature of this project. Data mining is the overall purpose for which the network will be designed, but the initial input processing, and evaluation are the most important features in this project. Good input processing will be very important, because the computer can not process language like a human can. This layer of the network must translate inputs into a format that is understandable by the computer. The next layer, the processing layer, will take the refined inputs and calculate their overall value, eventually deciding whether the subject is worth keeping. By having user feedback for several of the results, the computer must then be able to refine the neurons to provide a more accurate result. In the event that input processing and evaluation require more time than I have available, the ability of the network to refine itself will be dropped from the project.

2 Background

In order to begin this project, some research was done into previous neural networks and their success. It was determined that a simple feed forward network is the least time consuming network to create and test. However, it also appeared that time would still be available to create a dynamic network which relies on some feedback. This is the most important change for the feed forward network, but also the most complex. The most interesting piece of literature that was encountered was a thesis by Peter Meijer, which outlines a process for creating a network and modifying it.

3 Procedures

The first step in this process will be research into feed forward neural networks. In order to determine whether or not the logic and structure is correct, the network will be applied to a simple function approximation problem. In order to create the initial network the input processing layer will be added and the output will reflect the results from this layer. Next, the evaluation layer will be added and the network will new be a simple feed forward neural network. The next step would be dealing with training the network to read data. Finally, if time permits, feedback would be added to the network to allow for dynamic neuron creation or modification.

3.1 Testing

This program will be tested in a number of ways to establish the viability of the network.

1. Feed-forward Testing:

   The feed-froward network will be used as a function approximation network for a revised version of the Romania problem. Given the results of a search of the map, the program will determine values between the cities that were traveled to. This will be done by isolating as many of the variables as possible and assigning values which fit the constraints of the map.

2. Data Mining Applications

   The original feed-forward network will be modified slightly to allow for searching from specific files. This will be similar to the feed-forward network because the input will have to be transformed into a heuristic file which allows for analysis

3. Dynamic Network

   At this stage of the project the final layer will be added which allows for the creation and deletion of nodes int he network based on the feedback from the user. This process will be completed during the learning stage of the searching and then applied when the user initiates a search.

I plan on using the following tools and sticking with the proposed time scale below.

3.2 Software

Computer language(s) I'll use

1. Python and C will be used to program.

4 Schedule

In the first quarter I will begin research of the network and begin on the feed-forward network. After this network is completed shortly into second quarter, the network will be modified for the evaluation of input files. If time permits, near the middle of third quarter the feedback layer and modification methods will be added into the program.

5 Expected Results

With any luck, the results of this project will generate very accurate results when given a search parameter. There are additional areas for future researchers to build on as well. The network allows for different learning algorithms to be implemented and further increase accuracy.