Applicability of Backpropagation Neural Network for Recruitment Data Mining

Nitu Mathuriya   Dr. Ashish Bansal

SVIT, Indore   SVITS, Indore

Abstract

One of the biggest issues in the industries and institutions is the selection of the right candidate for their organization. Recruitment is the big issue for any organization and this process consume the more time, efforts and investment of any organization. During each phase o f the recruitment process, candidates are filtered based on some performance criteria. The preliminary or initial process of recruitment in any organization is resume filtering or selection, any organization has many resumes and they have to select effective (or according to their use) resumes from lot of resumes .This preliminary process also need a educated staff (2 or 3 persons included) to collect or short out right resumes for the selection of the right candidate for the organization. The main objective of this work is initial process of the recruitment (resume screening or selection) is done using the neural network. In this research the domain knowledge is extracted through knowledge acquisition technique. A Study has been made by applying back propagation neural network. Experiments were conducted by the data collected from the engineering institute to support their hiring power. The system supports managers in the process of assessment of specialists and managing the resources using formal methods. In this paper we describe experiments using artificial neural networks to automate the pre selection of the candidate to fill in the vacancies at their organization.

Key words:
Artificial neural network, knowledge acquisition, Industry, Organization, Pre selection

1   Introduction

A framework is proposed to improve the effectiveness of any organization in resume screening. In order to help large organizations (Industry and Institutes) manage resumes effectively, and improve the efficiency of recruitment, the paper introduces a method that extract information from kinds of different data sets. The recruitment process is one of the important functions of the HR department in any organization and resume selection is the first step towards creating the good and effective staff for the organization. Resume selection process is very complex and involves much of recourses and time of the experts. This system is also useful to find out educated and effective staff for any institute and industry. The process involves lot of effort by the recruiting team and money spent for this process. Every organization has its own recruitment process and selection criteria. Recruitment of the right and effective candidate is also important for IT companies to recruit the IT engineers, as well as the engineering institute. One of the mechanisms used by the IT industries is to conduct test and group discussion, marks obtained in the semester exams and etc. On the other hand institutes has different criteria such as Marks obtained, experience, higher qualification, demo, communication skill etc according to their or universities rules. So it is very difficult to implement a similar model for those. Resume filtering is a common process for any organization so its implementation is also beneficial to recruitment department with fewer efforts. It has been observed that lot of resumes are collected only for 1 or 2 posts. The time taken to conduct interview of all the candidates consumes more time. And lot of efforts is put to analyze the profile of the all applications to determine

the ones that suit the needs of the organization. Suppose 50applicants send their resumes for the 1 post means 1 among 50 applicants who apply get selected and the ratio of applicants interviewed after resume filtering is approximate 1:10,1:20,1:25 etc. Reducing this ratio will help the organization to save the effort.

Data mining is the important approach to realize knowledge discovery. It is the process of extracting knowledge or predicting previously unknown and useful trends from large quantities of data by using the knowledge of multidisciplinary fields such as statistics, mode identify, artificial intelligence, machine learning, database and so on. The artificial neural network (ANN) is one of the most efficient techniques of data mining. It is the nonlinear auto-fit dynamic system made of many cells with simulating the construction of biology neural systems. ANN has the ability to mapping high nonlinear system, associable memory and abstractly generalization. It can make model from analyzing the mode in the data and discover the unknown knowledge. The present paper gives an engineering application of data mining based on neural networks. The back propagation (BP) neural network is used as the algorithm of data mining. In this paper, the back propagation (BP) neural network method is used as the technique of data mining to analyze the effects of structural technologic parameters on efficiency in resume filtering. In this paper, results of the experiments conducted to cluster the data with the K means clustering and classification using back propagation algorithm have been analyzed.

The work reported in this paper attempts to perform the data mining task of Data Clustering. In the ANN paradigm, typically, supervised learning based BPN are used for data clustering tasks, because of their natural propensity to (a) find similarities amongst data items and (b) to group similar data items in proximity.

Traditional approaches to Information Processing Vs. Neural Networks

1. TA: Simulate and formalize human reasoning and logic process. TA treats the brain as a black box. A focus on how the elements are related to each other and how to give the machine the same capabilities.

NN: Simulate the intelligence functions of the brain. NN focus on modelling the brain structure. NN attempts to create a system that functions like the brain because it has a structure similar to the structure of the brain.

2. TA: The processing method of TA is inherently sequential.

NN: The processing method of NN is inherently parallel. Each neuron in a neural network system functions in parallel with others.

3. TA: Learning takes place outside of the system. The knowledge is obtained outside the system and then coded into the system.

NN: Learning is an integral part of the system and its design. Knowledge is stored as the strength of the connections among the neurons and it is the job of NN to learn these weights from a data set presented to it.

4. TA: Is deductive in nature. The use of the system involves a deductive reasoning process, applying the generalized knowledge to a given case.

NN: Is inductive in nature. It constructs an internal knowledge base from the data presented to it. It generalizes from the data, such that when it is presented a new set of data, it can make a decision based on the generalized internal knowledge

5. TA: It represents knowledge in an explicit form. Rules and relationships can be inspected and altered.

NN: The knowledge is stored in the form of interconnections strengths among neurons. Nowhere in the system, can one pick up a piece of computer code or a numerical value as a discernible piece of knowledge.

1.1 Data mining Techniques

The machine learning and the statistics are two main technical approaches of data mining. The machine learning, as a broad subfield of artificial intelligence, is concerned with the development of algorithms and techniques that allow computers to learn ability to achieve the tasks of identifying, inducing, classification, predication etc. Artificial Neural

Network (ANN) and K means are the most widely applied methods in those fields. ANN is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. The most typical neural networks are the BP neural network, the Hopfield neural networks and the adaptive neural networks. As the other technical support of data mining, the statistics offers the most fundamental theory of data mining techniques based on the precise mathematical approach.

1.2 Neural Network techniques in data mining

A Neural Network is a powerful data-modelling tool that is able to capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain. Neural networks resemble the human brain in the following two ways:

A neural network acquires knowledge through learning. A neural network's knowledge is stored within inter-neuron connection strengths known as synaptic weights. The Operation of a Neural Network is controlled by three properties: The pattern of its interconnections, architecture. Method of determining and updating the weights on the interconnections, training. The function that determines the output of each individual neuron, activation or transfer function. 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 Back propagation neural network. 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. Supervised learning or Associative 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 (back propagation algorithm), reinforcement learning and stochastic learning. The input-output pairs can be provided by an external teacher, or by the system which contains the neural network (self-supervised).

1.3 Back propagation Neural Network Technique

Back propagation is a form of supervised learning for multi-layer nets, also known as the generalized delta rule. Error data at the output layer is "back propagated" to earlier ones, allowing incoming weights to these layers to be updated. It is most often used as training algorithm in current neural network applications. The back propagation algorithm was developed by Paul Werbos in 1974 and rediscovered independently by Rumelhart and Parker. Since its rediscovery, the back propagation algorithm has been widely used as a learning algorithm in feed forward multilayer neural networks. What makes this algorithm different than the others is the process by which the weights are calculated during the learning network. In general, the difficulty with multilayer Perceptrons is calculating the weights of the hidden layers in an efficient way that result in the least (or zero) output error; the more hidden layers there are, the more difficult it becomes. To update the weights, one must calculate an error. At the output layer this error is easily measured; this is the difference between the actual and desired (target) outputs. At the hidden layers, however, there is no direct observation of the error; hence, some other technique must be used. To calculate an error at the hidden layers that will cause minimization of the output error, as this is the ultimate goal.

The back propagation algorithm is an involved mathematical tool; however, execution of the training equations is based on iterative processes, and thus is easily implementable on a computer.

1.4 Algorithm

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.

Actual Algorithm:

1. Initialize the weights in the network (often randomly)

2. Repeat

  * for each example e in the training set do

  1. O = neural-net-output (network, e) ;

  Forward pass
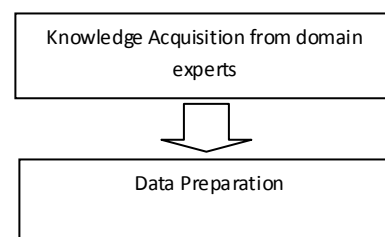
2. T = teacher output for e

3. Calculate error (T - O) at the output units

4. Compute delta_wi for all weights from hidden layer to output layer;

Backward pass

5. Compute delta_wi for all weights from input layer to hidden layer;

Backward pass continued

6. Update the weights in the network

* end

3. until all examples classified correctly or stopping criterion satisfied

4. Return (network)

2. Proposed Model

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to detect trends that are too complex to be noticed by either humans or computer techniques. The application of data mining based on neural networks can be generally divided into stages: Knowledge acquisition, data preparation, and modelling and knowledge discovery, as showed in Figure 1.

The design of the system requires the complete understanding of the problem domain. The data sets and the input attributes are determined through knowledge of an engineering college. Data preparation is the definition and expression of the mined data, which can make the mined data suitable for the algorithm. Data preparation is the most important step of data mining. Data selection means selecting data useful to data mining. It includes selections in two dimensions of line and row. First, the line dimension selection, namely field dimension, determines the neuron nodes of input and output layers of the neural network. Second, the row dimension selection, namely record dimension, determines the samples used to train and check the neural network. Before the selected data are used in the mining, they should be converted to a form acceptable, which is data expression. First, symbolic data type should be converted to a numerical one because neural network models can only deal with numerical data type. On the other hand, neural network models only accept the data between 0 and 1 or between -1 and 1 as input data. Therefore, the data should be mapped to this region according proportional transformation or others. The task and purpose should be confirmed at the beginning of data-mining stage. What algorithm would be used in mining is determined according to the characteristics of the task type. The back propagation algorithm is a widely used method in neural networks. It is extensively used in pattern recognition, character extraction, etc. After the determination of neural network models, network training can be conducted by using the selected data. This is a repeated process, in which the architecture of the neural network is determined. The aim is to determine a set of weights and threshold values that satisfy the precision requirement.

```
Knowledge Acquisition from domain
              experts
```

```
         Data Preparation
```

information and Experience. Personal information included Candidate name, address, mobile number and Email id.

Table 1: Personal information of the candidates (data set 1)

S no.	Na me	Address	Mobile no	Email id
1.	xyz	Indore	9926645442	xyz@gmail.com
2.	abc	Dewas	6548125452	abc@gmail.com
3.	prq	Indore	5468544252	prq@yahoo.com
4.	mnn	Indore	9926542584	mnn@rediffmail.com
5.	stv	Ujjain	9300258244	stv@gmail.com

Table 2: Educational qualification of the candidates (Data Set 1)

BE Year	Branc h	BE %	ME year	ME %	PHD
2005	CS	70	2010	72	Yes
2011	CS	60	No	No	No
2006	IT	72	2008	76	No
2005	EC	68	2007	72	No
2012	ME	74	NO	No	No

Figure 1 Data mining based on neural networks

Finally it is necessary that the discovered knowledge should be visualized, which can make the rules understandable easily and beneficial to applications. The constructed models were reviewed and evaluated before it is used for selection. The models were evaluated using accuracy as the criteria to assess the performance of the method.

3. Experimental Consideration

The system proposed in this paper has been implemented and evaluated with extensive experimentations on the collected data sets from engineering college for recruiting teaching staff in the college. Accuracy of the classification is used as a metric for performance. This section presents with the details of the data sets.

The data set of table 1 contains the all information of the candidate e. g. personal information, Educational

Table 3: Experience and Extra qualification of the candidates with selection output (Data Set 1)

Industry Experience	Teaching Experience	Extra Qualification (Gate card)	Desired Output
2	2	Yes	1
No	1	No	0
3	1	No	1
No	4	Yes	1

No	No	No	0

The data set contains these input tables data in a one table with the desired output. Firstly apply the data selection and data expression on these data set1.After the data pre-processing apply the Back propagation technique on these data set. The BP neural network model consists of three layers. The topology of the neural network model is shown in Figure 2.
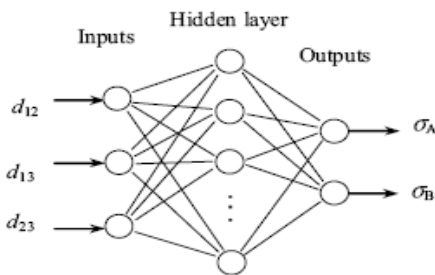

Figure 2 Architecture of Neural Network

After the sample training, a set of weights and threshold values which satisfy the precision requirement are achieved. All computer programs are completed in the environment of MATLAB. The results of data mining are shown in table 4.

Table 4: Result 1 (Output table)

S no.	Name	Address	Mobile no	Email id
1.	xyz	Indore	9926645442	xyz@gmail.com
3.	prq	Indore	5468544252	prq@yahoo.com
4.	mnn	Indore	9926542584	mnn@rediffmail.com

This output table shows the personal information of the selected candidates from the data set1.Table 4 give the information of those candidates after applying the neural network technique whose resume information deserve for the selection. This filtering done by applying the BPNN technique on training data set and finds the result.

4. Evaluation Methodology

The metric used to evaluate the clustering and classification algorithm is the accuracy. Accuracy is determined as the ratio of records correctly classified

during testing to the total number of records tested. BPNN techniques were applied with MATLAB and the accuracy of the network is depicted in table 5.It is observed that neural network techniques have better accuracy and suitable for this problem domain.

Table 5: Result of BPNN technique trained with data set1 and tested with dataset2

Training data set	Test Data set	Test data set Record	Accuracy
Dataset1	Dataset1	450 approx	92
Dataset1	Dataset2	300 approx	88
Dataset2	Dataset2	300 approx	83
Dataset2	Dataset1	450 approx	78

Dataset 1 and Dataset 2 both contain different record sets. It was observed that the accuracy of BPNN better than other traditional algorithms.

5. Conclusion and Future work

In this paper we present a neural network based approach to mining classification rules from given databases. The approach consists of different phases:

1) Constructing and training a network to correctly classify tuples in the given training data set to required accuracy,
2) Extracting knowledge through the network using BPNN

A set of experiments was conducted to test the proposed approach is using a well defined set of data mining problems. The results indicate that, using the proposed approach, high quality or useful data can be discovered from the given data sets. In future we can apply this technique to select deserving candidates for an organization. This technique is also applicable on feedback system in any organization. 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.

References

[1] Jiawei Han, Micheline Kamber. 2006. Data Mining Concepts and Techniques", Second Edition Morgan Kaufmann Publishers, San Francisco.

[2] Haykin, S., Neural Networks, Prentice Hall International Inc.

[3] L. Wang and T. Z. Sui, Data Mining Technology Based on Neural Network in the Engineering, School of Mechanical Engineering and Automation ortheastern University IEEE 2007

[4] N. Sivaram Research Scholar and V, Department of CSE ,Clustering and Classification Algorithms for Recruitment Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010

[5] "Artificial Neural Network", Wikipedia Encyclopedia, Wikipedia Foundation, Inc., 2006.

[6] Arun K Pujari ,Data mining techniques universities press India(Private limited) Eight Impression 2005