IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 07-11
www.iosrjournals.org

Survey on Methodologies of Data Mining using Neural Network

Mrs. Mausami Sawarkar[1], Mr. Dhiraj Rane[2]
[1](Mausami_sawarkar@rediffmail.com Dept of CS, PCE, Nagpur)
[2](dhirajrane2302@gmail.com, Dept. of CS, GHRIIT, Nagpur, India)

Abstract: *The traditional data mining algorithms are hard to apply on noisy data, redundant information, incomplete data and sparse data in database, or the application effects are not good. But neural network have many virtues such as robustness, parallelism and anti-noise, so it is very effective on data mining in large and real databases. This paper expounds the process of data mining based neural network in detail, discusses the algorithms of classifying and clustering, indicates the problems at present and makes an expectation for the development*

Keywords- *Neural Network, Data Mining, Classification, Regression*

I. INTRODUCTION

Data mining is an interdisciplinary subfield of computer science.[1][2][3] It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.[1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.[1] Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.[1] Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

Data mining involves six common classes of tasks:[4]

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

An ANN is typically defined by three types of parameters:

1. The interconnection pattern between the different layers of neurons
2. The learning process for updating the weights of the interconnections
3. The activation function that converts a neuron's weighted input to its output activation.

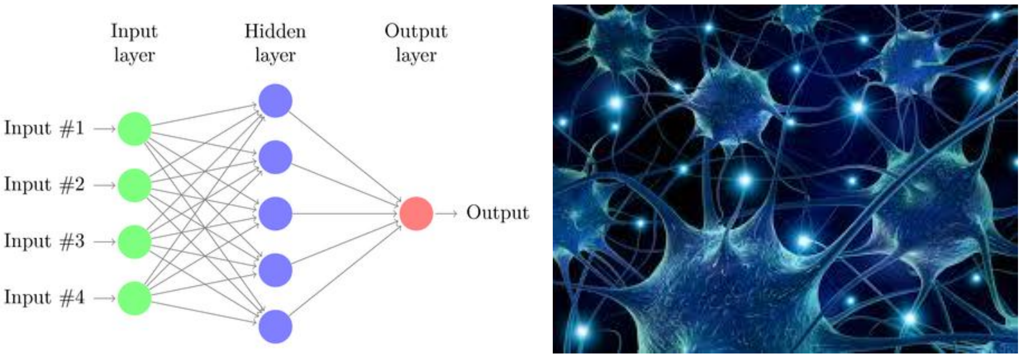

Figure 1: Artificial Neural Network

II. ARTIFICIAL NEURAL NETWORK 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 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. 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

III. STUDIES OF DATA MINING USING ANN

Data mining process can be composed by three main phases:

A. data preparation,
B. data mining,
C. expression and interpretation of the results,

Data mining process is the reiteration of the three phases. The details are shown in Fig. 2.

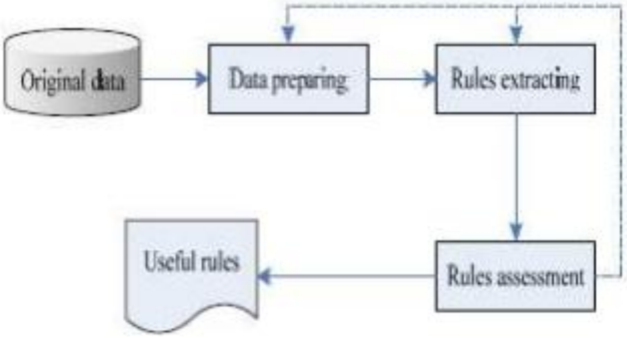

Figure 2: Data Mining using Neural Network

- Data Preparation

Data preparation is to define and process the mining data to make it fit specific data mining method. Data preparation is the first important step in the data mining and plays a decisive role in the entire data mining process.

It mainly includes the following four processes:

a) Data cleaning: Data cleansing is to fill the vacancy value of the data, eliminate the noise data and correct the
b) Inconsistencies data in the data.
c) Data option: Data option is to select the data arrange and row used in this mining.
d) Data preprocessing: Data preprocessing is to enhanced process the clean data which has been selected.
e) Data expression

Data expression is to transform the data after preprocessing into the form which can be accepted by the data mining algorithm based on neural network. The data mining based on neural network can only handle numerical data, so it is need to transform the sign data into numerical data. The simplest method is to establish a table with one-to-one correspondence between the sign data and the numerical data. The other more complex approach is to adopt appropriate Hash function to generate a unique numerical data according to given string. Although there are many data types in relational database, but they all basically can be simply come down to sign data, discrete numerical data and serial numerical data three logical data types. Fig. 2 gives the conversion of the three data types. The symbol “Apple” in the figure can be transformed into the corresponding discrete numerical data by using symbol table or Hash function. Then, the discrete numerical data can be quantified into continuous numerical data and can also be encoded into coding data

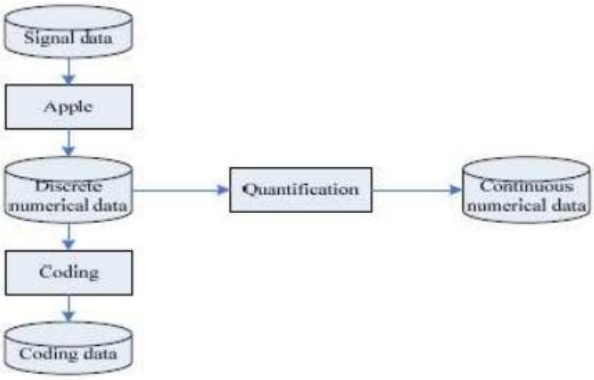

Figure 2: Data expression and conversion in data mining based on neural network

- **Rules Extracting**

There are many methods to extract rules, in which the most commonly used methods are LRE method, black-box method, the method of extracting fuzzy rules, the method ofextracting rules from recursive network, the algorithm of binary input and output rules extracting (BIO-RE), partial rules extracting algorithm (Partial-RE) and full rules extracting algorithm (Full-RE).

- **Rules Assessment**

Although the objective of rules assessment depends on each specific application, but, in general terms, the rules can be assessed in accordance with the following objectives.

1) Find the optimal sequence of extracting rules, making it obtains the best results in the given data set;
2) Test the accuracy of the rules extracted;
3) Detect how much knowledge in the neural network has not been extracted;
4) Detect the inconsistency between the extracted rules and the trained neural network.

IV. USE OF ANN FOR DATA MINING

Data mining process can be composed by three main phases:

IV.1. NEURAL NETWORK FOR CLASSIFICATION AND REGRESSION

The application domain for neural networks is extensive. Grouping similar applications in types helps to profit from previous experience when you are required to design new applications. It is usual to distinguish between 2 main types:

- classification problems
- regression problems

A data classification task is characterized by a set of records which should be assigned to one of a set of predefined categories based on the content of the records. The content is a set of variable values. In some applications, there are only few categories, minimum 2, to which a large number of inputs should be assigned. We have already met and solved one classification task, the XOR problem. From the list in Figure 1, we recognize other similar classification problems frequently used are classification of a set of medical records with symptoms of illness in categories as records for serious and less serious cases, classification of a set of digitized voices representations into a category for male and an another for female voice representations.

IV.1. PATTERN RECOGNITION

The most known application domain for neural networks is probably pattern recognition. The pattern recognition applications vary from training a neural network to uniquely identify each individual in a set of photographic images, to training a net to classify individuals in a population by gender based on pictures. Humans have a fantastic ability to perform pattern recognition without being able to give a comprehensive explanation for the 'rules' they use. We have usually no problem to distinguish between pictures of a 'cat and a dog. But try to set down the rules you use for a rule-based computer system. Another frequently investigated application is character recognition. Also in this field, the tasks vary from the very simple to the complex. A simple application is the recognition of decimal digits in a standard form, while the most challenging is the recognition of letters in handwritten messages. The approach to solving these tasks is to create an image for each character, divide each image into components by a grid. Each grid cell corresponds to a pixel of the image and is represented by an input variable. In the case of a black and white image, each pixel can be represented by a binary variable with only 2 values, 'white' or 'black'. If the image has colors, a categorical variable will be required for each pixel with as many codes as there are different colors. The whole area represented by the pixel is considered to have the same color. The resolution, the amount of detail or number of the pixels used in the application to represent the input character, is an important factor. High resolution means that details are preserved in the image, but it also means that the number of input variables is large and resource consuming, while a low resolution does not preserve as much information but is cheaper to process. It is important to find a good balance between the requirements of details and resources.

IV.2. CASE STUDY

Dr. K. Usha Rani [14] has proposed the study of heart disease diagnosis using NN. 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. In this study Neural network technique is adopted for classification of medical dataset. The experiment is conducted with Heart Disease dataset by considering the single and multilayer neural network modes. Backpropogation algorithm with momentum and variable learning rate is used to train the networks. To analyze performance of the network various test data are given as input to the network. Parallelism is implemented at each neuron in all hidden and output layers to speed up the learning process. The experimental results proved that neural networks technique provides satisfactory results for the classification task.

LIU Han-li, LI Lin, ZHU Hai-hong [14] puts forward a method that combines the learning algorithm of BP neural network with genetic algorithm to train BP network and optimize the weight values of the network in a global scale. This method is featured as global optimization, high accuracy and fast convergence. The data-mining model based on genetic neural network has been widely applied to the procedure of data mining on case information in the command centre of police office. It achieves an excellent effect for assisting people to solve cases and make good decisions. In this paper, the principles and methods of this data-mining model are described in details. A real case of its application is also presented. From this case we can draw a conclusion that the data-mining model we have chosen is scientific, efficient and practicable.

V. CONCLUSION

Compared to statistical methods, NN are useful especially when there is no a priori knowledge about the analyzed data. They offer a powerful and distributed computing architecture, with significant learning abilities and they are able to represent highly nonlinear and multivariable relationships. However, NN are not appropriated for any DM problem and the selection of network architecture for a specific problem has to be done carefully. We have not attempted to provide an exhaustive survey of the available NN algorithms that are suitable for DM. Instead, we have described a subset of the problems and constrains, selected to illustrate the breath of relevant approaches as well as the key issues that arise in applying NN in a DM setting..

REFERENCES

[1] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[2] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[3] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[4] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.

[5] Agrawal, R., Imielinski, T., Swami, A., “Database Mining: A Performance Perspective”, IEEE Transactions on Knowledge and Data Engineering, pp. 914- 925, December 1993.

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

[7] Berson, “Data Warehousing, Data-Mining & OLAP”, TMH.

[8] Haykin, S., Neural Networks, Prentice Hall International Inc., 1999.

[9] Khajanchi, Amit, Artificial Neural Networks: The next intelligence.

[10] Zurada J.M., “An introduction to artificial neural networks systems”, St. Paul: West Publishing (1992).

[11] Y. Bengio, J. M. Buhmann, M. Embrechts, and J.M. Zurada. Introduction to the special issue on neural networks for data mining and knowledge discovery. IEEE Trans. Neural Networks.

[12] M. W. Craven and J. W. Shavlik. Using neural networks for data mining. Future Generation Computer Systems, 13:211–229, 1997.

[13] ANALYSIS OF HEART DISEASES DATASET USING

[14] Dr. K. Usha Rani, Analysis of heart disease dataset using neural network approach, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011

[15] LIU Han-li, LI Lin, ZHU Hai-hong , GENETIC NEURAL NETWORK BASED DATA MINING AND APPLICATION IN CASE ANALYSIS OF POLICE OFFICE, 2010