IJITKM Special Issue (ICFTEM-2014) May 2014 pp. 19-26 (ISSN 0973-4414)

Neural Networks in Data Mining From Past to the Future

[1]Er.Mohit Kumar Vats,[2]Er.Misba Hashmi
[1,2]Dept. of Computer Science & Engg.
[1]mail@mksvats.com,[2]misbahashmi9@gmail.com

ABSTRACT: Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives (massive date warehouse) and summarizing it into useful information. Due to the importance of extracting knowledge/ information from the large data repositories, data mining has become an essential component in various fields of human life. 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. Neural-network methods are not commonly used for data-mining tasks, however, because they often produce incomprehensible models and require long training times. This paper is an overview of artificial neural networks and questions their position as a preferred tool by data mining practitioners.
Keywords: Data Mining, Historical Trends, Current Trends, Future Trends, Artificial Neural Networks (ANN).

1. INTRODUCTION

The data collected from different applications require proper mechanism of extracting knowledge/information from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining. The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [2]. 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.

The various application areas of data mining are Life Sciences (LS), Customer Relationship Management (CRM), Web Applications, Manufacturing, Competitive Intelligence, Retail/ Finance/ Banking, Computer/ Network/ Security, Monitoring/ Surveillance, Teaching Support, Climate modeling, Astronomy, and Behavioral Ecology etc. In this article, we provide an introduction to the topic of using neural-network methods for data mining. Neural networks have been applied to a wide variety of problem domains to learn models that are able to perform such interesting tasks as steering a motor vehicle, recognizing genes in uncharacterized DNA sequences, scheduling payloads for the space shuttle, and predicting exchange rates. Neural Networks have not often been applied in data-mining settings, in which two fundamental considerations are the comprehensibility of learned models and the time required to induce models from large data sets. We discuss new developments in neural-network learning that effectively address the comprehensibility and speed issues which often are of prime importance in the data-mining community. Specifically, we describe algorithms that are able to extract symbolic rules from trained neural networks, and algorithms that are able to directly learn comprehensible models. This paper is organized as follows historical perspectives of data mining, current trends in data mining section and future trends of data mining using ANN.

2. HISTORICAL, CURRENT & FUTURE TRENDS OF DATA MINING

The building blocks of data mining is the evolution of a field with the confluences of various disciplines, which includes database management systems(DBMS), Statistics, Artificial Intelligence(AI), and Machine Learning(ML).

2.1 Data Trends

In initial days, data mining algorithms work best for numerical data collected from a single data base, and various data mining techniques have evolved for flat files, traditional and relational databases where the data is stored in tabular representation. Later on, with the confluence of Statistics and Machine Learning techniques, various algorithms evolved to mine the non numerical data and relational databases.

2.2 Computing Trends

The field of data mining has been greatly influenced by the development of fourth generation programming languages and various related computing techniques. In, early days of data mining most of the algorithms employed only statistical techniques. Later on they evolved with various computing techniques like AI, ML and Pattern Reorganization. Various data mining techniques (Induction, Compression and Approximation) and algorithms developed to mine the large volumes of heterogeneous data stored in the data warehouses.

2.3 Current Trends

The ever increasing complexities in various fields and improvements in technology have posed new challenges to data mining; the various challenges include different data formats, data from disparate locations,

advances in computation and networking resources, research and scientific fields, ever growing business challenges etc.

The following table depicts various currently employed data mining techniques and algorithms to mine the various data formats in different application areas. The various data mining areas are explained after the table1.

a. Mining the Heterogeneous data
b. Utilizing the Computing and networking Resources.
c. Research and scientific computing trends.
d. Business Trends.

2.5 Future Trends Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments. Ever increasing technology and future application areas are always poses new challenges and opportunities for data mining, the typical future trends of data mining includes Standardization of data mining languages Data preprocessing Complex objects of data Computing resources Web mining Scientific Computing Business data.

Table 1: Current Data Mining areas and techniques to mine the various Data Formats

Data mining type	**Application Areas**	**Data Formats**	**Data mining Techniques/Algorithms**
Hypermedia data mining	Internet and Intranet Applications.	Hyper Text Data	Classification and Clustering Techniques
Ubiquitous data mining	Applications of Mobile phones, PDA, Digital Cam etc	Ubiquitous Data	Traditional data mining techniques drawn from the Statistics and Machine Learning
Multimedia data mining	Audio/Video Applications	Multimedia Data	Rule based decision tree classification algorithms
Spatial Data mining	Network, Remote Sensing and GIS applications	Spatial Data	Spatial Clustering Techniques, Spatial OLAP
Time series Data mining	Business and Financial applications.	Time series Data	Rule Induction algorithms.

The following table presents the comparative statement of various data mining trends from past to the future.

Table 2: Data mining Trends Comparative Statement

Data mining trends	**Algorithms/ Techniques employed**	**Data formats**	**Computing Resources**	**Prime areas of applications**
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored in traditional databases	Evolution of 4G PL and various related techniques	Business
Current	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization	Heterogeneous data formats includes structured, semi-structured and	High speed networks, High end storage devices and Parallel, Distributed computing etc	Business, Web, Medical diagnosis etc…
Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects includes high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi-instance objects, Multi-	Multi-agent technologies and Cloud Computing	Business, Web, Medical diagnosis, Scientific and Research analysis fields (bio, remote sensing etc…), Social networking

In the next section, we consider the applicability of neural- network methods to the task of data mining

3. Data mining and 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.

3.1 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.

3.2 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

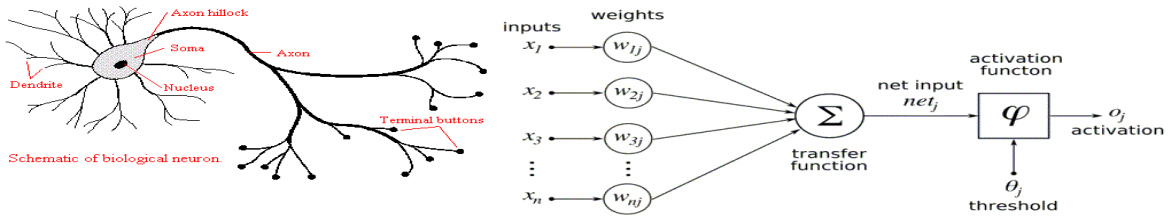


Fig1: Schematic View of biological neuron & Schematic View of Artificial Neural Networks.

- **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 .

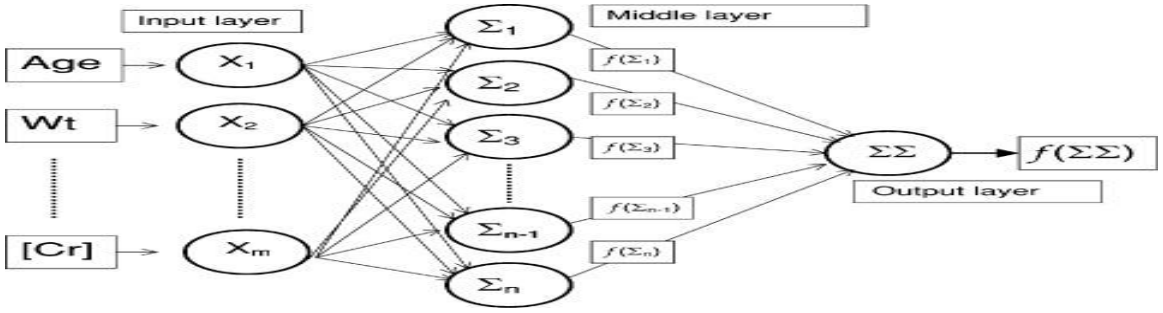


Fig2: Architecture of Artificial Neural Networks Used in Data Mining & Other Application.

4. An Introduction to Data Mining

Data Mining can be defined as extracting or mining knowledge from large amount of data (Han and Kamber, 2001). It is considered as one of the steps in knowledge discovery process where valuable information is extracted from large data bases. It can also be defined as a process of extracting valid, previously unknown, non-trivial and useful information from large databases (Rao, 2003). Shaw et al (2001) have classified data

mining tasks into five categories as Dependency analysis, class identification, concept description, deviation detection and data visualization summarized below.

1. Dependency Analysis: It can be defined as a process of finding associations between entities and finding relationships among them like market basket analysis.
2. Class Identification: It can be defined as grouping various entities into classes. It includes two types of identification tasks –mathematical taxonomy and concept clustering (Shaw et al., 2001).
3. Concept Description: In concept description groups are made on the basis of the domain knowledge and databases, without any compulsory descriptions. It includes tasks like data summarization and data comparison.
4. Deviation Detection: It includes tasks like finding changes in data and anomaly detection that is finding actions that are different form the benign actions.
5. Data Visualization: It includes finding and analyzing different patterns which are complex in nature. It can be used to explore the databases and can be used alone or in combination with any of the above mentioned tasks.

There are various Data mining techniques that have been proposed in literature so far. Some of them are mentioned below:

1. Classification: It is a technique which is based on identification and formation of classes based on certain criteria and is predefined. It is used to predict future actions using various techniques like decision trees, neural networks and memory based reasoning.

2. **Neural Networks:** Neural Networks follow predictive model which are based on biological modeling capability

3. Decision Trees: It is a tree like structure where the leaf node represents or predicts the decision and the non-leaf node represents the various possible conditions that can occur. It includes algorithms like CART and CHAID.

4. Clustering: Clustering is used to group data items into clusters which are not predefined. It is basically of two types- Hierarchical and Non Hierarchical. It includes algorithms like K-means Clustering and DBSC.

5. Association: Association is aimed at finding relationships among data sets and entities that are present in the data bases. A classic example of association technique is market basket analysis and includes algorithms like Apriori and dynamic item set counting.

6. Genetic Algorithms: Genetic Algorithms are based on the concept of evolution of genes that is carrying the features from one stage to other and follow optimization techniques for the same.

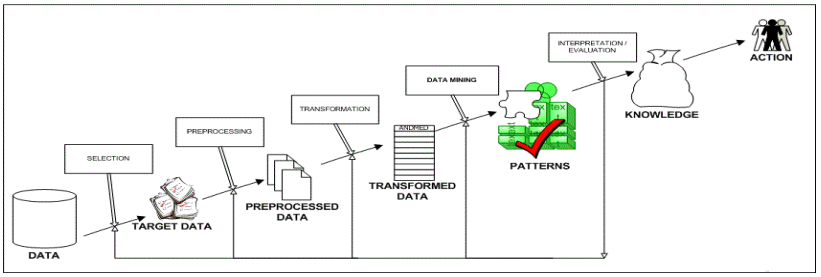


Fig3: Data Mining Process predicts data by a learning process**.**

5 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 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.

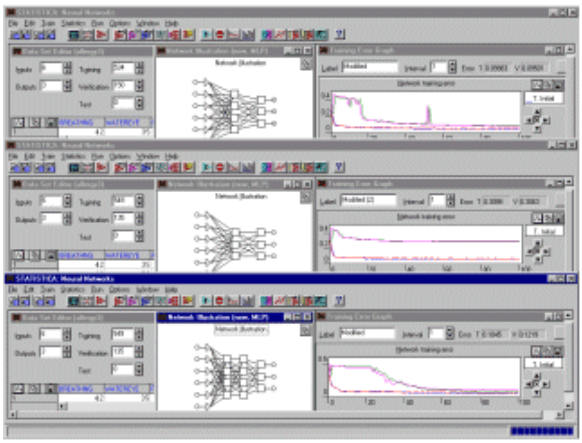
5.1. Feed forward Neural Network in Data Mining.

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.

5.2. 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.

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).

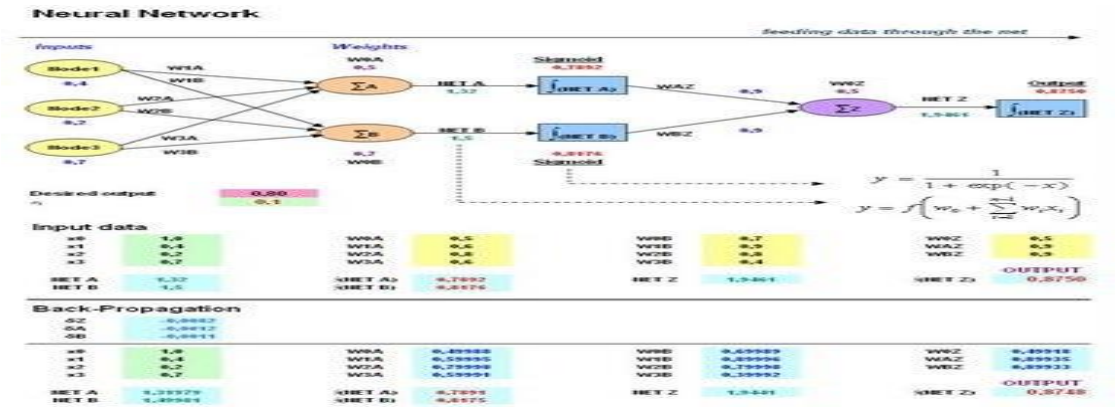
5. Assign "blame" for the local error to neurons at the previous level, giving greater responsibility to neurons connected by stronger weights.

6. Repeat the steps above on the neurons at the previous level, using each one's "blame" as its error.

6. Results & Conclusion.

6.1 Input Data set and outputs

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. As Software companies develop more sophisticated models with user-friendly interfaces the attraction to neural networks will continue to grow.

7. REFERENCES

[1] Heikki, Mannila. 1996. Data mining: machine learning, statistics, and databases, IEEE

[2] Fayadd, U., Piatesky -Shapiro, G., and Smyth, P. 1996.From Data Mining To Knowledge Discovery in Databases, AAAI Pres The MIT Press, Massachusetts Institute Of Technology. ISBN 0–26256097–6

[3] Fayap. Piatetsky-Shapiro, Gregory. 2000. The Data Mining Industry Coming of Age. IEEE Intelligent Systems.

[4] Salmin, Sultana et al. 2009. Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture, International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4, October, 2009.

[5] Hsu, J. 2002. Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century, The Proceedings of the 19th Annual Conference for Information SystemsEducators(ISECON2002),ISSN:15427382.AvailableOnline:http://colton.byuh.edu/isecon/2002/224b/Hsu.pdf.

[6] Shonali Krishnaswamy. 2005. Towards Situation- awareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application (2005), Proceedings of Conference on Intelligent Vehicles and Road Infrastructure 2005, pages-16,17.Available at http://www.csse.monash.edu.au/~mgaber/CameraReadyI

[7] Kotsiantis, S., Kanellopoulos, D., Pintelas, P. 2004.Multimedia mining. WSEAS Transactions on Systems,

[8] Abdulvahit,Torun. Ebnem, Düzgün. 2006. Using spatial data mining techniques to reveal vulnerability of people and places due to oil transportation and accidents: A case study of Istanbul strait, ISPRS Technical Commission II Symposium, Vienna. Addison Wesley,1st edition.

[9] T. M. Mitchell. 1982. Generalization as Search, Artificial Intelligence, 18(2), 1982, pp.203-226.

[10] R. Michalski., I. Mozetic., J. Hong., and N. Lavrac. 1986. The AQ15 Inductive Leaning System: An Overview and Experiments, Reports of Machine Leaning and Inference Laboratory, MLI-86-6, George Maseon University.

[11] J. R. Quinlan.1992.Programs for Machine Learning, Morgan Kaufmann.

[12] Z. K. Baker and V. K.Prasanna. 2005. Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs. In Submitted to the IEEE International Parallel and Distributed Processing Symposium (IPDPS '05).

[13] Jing He.2009. Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204

[14] Ali Meligy.2009. A Grid-Based Distributed SVM Data Mining Algorithm, European Journal of Scientific Research ISSN 1450-216X Vol.27 No.3. Pp.313-321 © Euro Journals Publishing, Inc. Available at : http://www.eurojournals.com/ejsr.htm

[15] S. Mitra, S. K. Pal, and P. Mitra. 2001. Data mining in soft computing framework: A survey, IEEE