Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.879 – 883

RESEARCH ARTICLE

Data Mining Techniques

Sayyed Muzammil Ali[1], Prof. Ms. R.R Tuteja[2]

[1]Department Computer Science & Engineering & S.G.B.A.U, India

[2]Department Computer Science & Engineering & S.G.B.A.U, India

[1] smali.muzammil@gmail.com; [2] ranu.tuteja@gmail.com

Abstract— Knowledge discovery in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. In this paper, we provide an overview of common knowledge discovery tasks and approaches to solve these tasks. The concept of data mining was summarized and its significance towards its methodologies was illustrated. The data mining based on Neural Network and Genetic Algorithm is researched in detail and the key technology and ways to achieve the data mining on Neural Network and Genetic Algorithm are also surveyed.

Keywords— Data Mining, Data mining task, Data mining life cycle, Data mining Methods

I. INTRODUCTION

The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining". It is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in database.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information –information that can be used to increase revenue, cuts costs, or both. Data mining software is allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data .As and when the customer will require the data should be retrieved from the database and make the better decision. The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of "Data mining" is due to the perception of *"we are data rich but information poor"*. There is huge volume of data but we hardly able to turn them in to useful information and knowledge for managerial decision making in business. To generate information it requires massive collection of data. It may be different formats like audio/video, numbers, text, figures, Hypertext formats .To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining'. Data mining is

© 2014, IJCSMC All Rights Reserved

the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [1,2]. This paper mainly contains 4 sections .Section 1 is completely introduction to the data mining concept. Section 2 describes the data mining task which describes that how the data will be store, how to retrieve, how to analyze the data and focuses the data mining classification tasks .section 3 provides the data mining life cycles. Section 4 describes shortly, some of the popular data mining methods.

II. DATA MINING TASK

Data mining deals with what kind of patterns can mined. On the basis of kind of data to be mined there are two kinds of functions involved in data mining, that are listed below.

A. *Descriptive*

Descriptive mining tasks characterize the general properties of data in database. Predictive mining tasks perform inference on the current data in order to make predictions. Below is list of descriptive functions:

- Class/ concept description
- Mining of frequent pattern
- Mining of associations
- Mining of correlations
- Mining of clusters

B. Classification and prediction

Predictive data mining tasks perform inference of the current data in order to make prediction. Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation. The goal of a predictive data mining model is to predict the future outcomes based on passed records with known answers.

Classification requires the data mining algorithm to partition the input space in such a way as to separate the examples based on their class.

III. DATA MINING LIFE

The life cycle of a data mining project consists of six phases [2, 3]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

A. Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary goal designed to achieve the objectives.

B. Data mining understanding

The data understanding phase starts with data collection and proceeds with activities to become familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

C. Data preparation

This phase covers all activities to construct the final dataset .Data preparation tasks are likely to be performed more times, and not in any prescribed order. Tasks include table, record, attribute selection as well as transformation and cleaning of data for modeling tools.

D. Modeling

In modeling phase, various modeling techniques are selected and applied their parameters are calibrated to optimal values. There are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

E. Evaluation

At this stage in the project the model (or models) built appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model more thoroughly, and review the steps executed to construct the model, to be certain it properly achieves the business

© 2014, IJCSMC All Rights Reserved

objectives. A key objective is to determine if there is some important business issue that has not been considered sufficiently. At the end of this phase, a decision on the use of the data mining results should be reached.

F. Deployment

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

IV. METHODOLOGIES OF DATA MINING

A DECISION TREE CLASSIFIERS

Decision Trees (DT) are like those used in decision analysis where each non-terminal node represents a test or decision on the data item considered. Depending on the outcome of the test, one chooses a certain branch.As in data mining applications, very large training sets with several million examples are common; a decision tree classifier scales well and can handle training data of this magnitude. So, for classifying large data sets, our focus mainly on decision tree classifiers. Tree-shaped structure that represent sets of decisions. In decision tree node represent a test on an attribute value, branch represents an outcome of the test and tree leaves represent classes or class distribution. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented the sequence of tests. Interior node in the decision tree tests the value of some input variable, and the branches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a input instance is start at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached [4].Decision tree is represented in figure1.

A disadvantage of DT is that trees use up data very rapidly in the training process. They should never be used with small data sets. They are also highly sensitive to noise in the data, and they try to fit the data exactly, which is referred to as "overfitting".

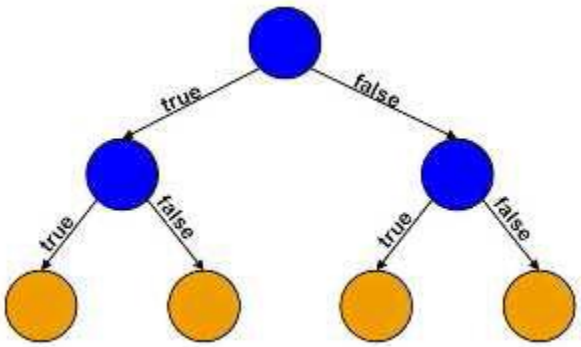

Fig 1 Decision tree

B . 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 anipulation 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. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications [5].

© 2014, IJCSMC All Rights Reserved

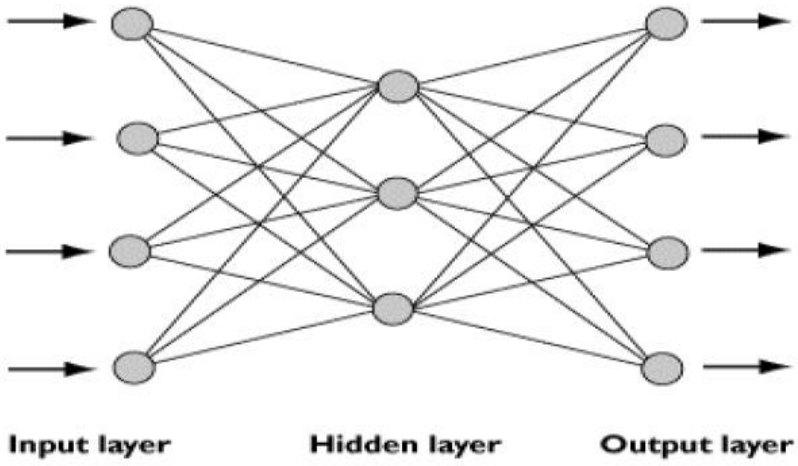

Fig: 2 Neural Network with hidden layers

It is shown in fig.2. Artificial neural network have become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technology. ANN is an adaptive, non linear system that learns to perform a function from data and that adaptive phase is normally training phase where system parameter is change during operations. After the training is complete the parameter are fixed. If there are lots of data and problem is poorly understandable then using ANN model is accurate, the non linear characteristics of ANN provide it lots of flexibility to achieve input output map.

C .Genetic Algorithm

Genetic Algorithm attempt to incorporate ideas of natural evaluation The general idea behind GAs is that we can build a better solution if we somehow combine the "good" parts of other solutions (schemata theory), just like nature does by combining the DNA of living beings [6].

The Genetic Algorithm was developed by John Holland in 1970. They are based on the genetic processes of biological organisms. Over many generations, natural populations evolve according to the principles of natural selection and "survival of the fittest", first clearly stated by Charles Darwin in the Origin of Species. GAs are adaptive method which may be used to solve search and optimization problems. After a number of new generations built with the help of the described mechanisms one obtains a solution that cannot be improved any further. This solution is taken as a final one [7]. When GAs are used for problem solving, the solution has three distinct stages.

1) Selection: The selection function chooses the parents using roulette wheel and uniform sampling, based on expectation and number of parents.

2) Crossover: The crossover function is position independent. This crossover function creates the crossover children of the given population using the available parent.

3) Mutation: Produce new gene individuals by recombining features of their parents.

D. Rule Extraction

Andrews [8] identifies three categories for rule extraction procedures: decompositional, pedagogical, and eclectic. Decompositional rule extraction involves the extraction of rules from a network in a neuron-by-neuron series of steps [9].The drawbacks of decompositional extractions time and computational limitations. The advantages of decompositional techniques are that they do seem to offer the prospect of generating a complete set of rules for the neural network.

Pedagogical rule extraction [10] treats the entire network as a black box. In this approach, inputs and outputs are matched to each other. The decompositional approaches can produce intermediary rules that are defined for internal connections of a network, possibly between the input layer and the first hidden layer. The eclectic approach is merely the use of those techniques that incorporate some of a decompositional approach with some of a pedagogical approach.

There are several main rule formats. Rule extraction algorithms will generate rules of either conjunctive form or subset selection form, commonly referred to as M-of-N rules named for the primary rule extraction that makes use of the form. All rules follow the natural language syntactical if-then prepositional form.

© 2014, IJCSMC All Rights Reserved

Conjunctive rules follow the format:
IF condition1 AND condition2 AND condition3 THEN RESULT
Generally a rule consists of two values. A left hand antecedent and a right hand consequent. An antecedent can have one or multiple conditions which must be true in order for the consequent to be true for a given accuracy whereas a consequent is just a single condition.

Craven and Shavlik in there paper [11] listed criteria for rule extraction, which are as follows:

A. ***Comprehensibility****: The extent to which extracted representations are humanly comprehensible.*

B. ***Fidelity:*** *The extent to which extracted representations accurately model the networks from which they were extracted.*

C. ***Accuracy:*** *The ability of extracted representations to make accurate predictions on previously unseen cases.*

D. ***Scalability:*** *The ability of the method to scale to networks with large input spaces and large numbers of weighted connections.*

E. ***Generality:*** *The extent to which the method requires special training.*

CONCLUSIONS

In this paper we briefly reviewed the various data mining applications. data mining has begun to mature as a discipline, its methods and techniques have not only proven to be useful, but have begun to be accepted by the wider community of data analysts. At present data mining is a new and important area of research and ANN itself is a very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The commercial, educational and scientific applications are increasingly dependent on these methodologies.

REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.
[3] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R... "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark) , DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".
[4] Lior Rokach and Oded Maimon, "Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and Artificial Intelligence)", ISBN: 981-2771-719, World Scientific Publishing Company, , 2008.
[5] R. Andrews, J. Diederich, A. B. Tickle," A survey and critique of techniques for extracting rules from trained artificial neural networks", Knowledge-Based Systems, vol.- 8,no.-6, pp.-378-389,1995.
[6] Ankita Agarwal, "Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012.
[7] Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. MIT Press.
[8] Andrews, Robert; J. Diederich; and A. B. Tickle. 1995. A Survey and Critique Of Techniques For Extracting Rules From Trained Artificial Neural Networks. Knowledge Based Systems 8:373-389.
[9] Darrah, Marjorie, Brian Taylor and Michael Webb. A Geometric Rule Extraction Approach used for Verification and Validation of a Safety Critical Application. 2005 Florida Artificial Intelligence Research Society Conference, Clear Water Beach, FL, May 16-18, 2005.
[10] Nayak R., R. Hayward and J. Diederich. 1997. "Connectionist Knowledge Base Representation by Generic Rules from Trained Feed forward Neural Networks", Proceeding of Connectionist Systems for Knowledge Representation and Deduction Workshop, Townsville, Australia, July 13-19.
[11] M. Craven and J. Shavlik, "Learning rules using ANN ", Proceeding of 10th International Conference on Machine Learning, pp.-73-80, July 1993.

© 2014, IJCSMC All Rights Reserved