A Review of Data Mining Techniques with their Merits & Demerits

Dr. Pardeep Mittal¹
Professor, Department of CSE
Guru Kashi University
Talwandi sabo, Punjab, India

Sukhpreet Singh², Amritpal Singh³, Priyanka⁴
M.Tech, Department of CSE
Guru Kashi University
Talwandi sabo, Punjab, India

Abstract: The basic principle of data mining is to analyze the data from different angle, categorize it and finally to summarize it. In today’s world data mining have increasingly become very interesting and popular in terms of all application. The need for data mining is that we have too much data, too much technology but don’t have useful information. There are different software’s of data mining with different techniques to analyze data. To extract hidden predictive information from large volume of data, data mining techniques are required. This paper introduce various data mining techniques such as decision tree, Neural network, Nearest neighbor and clustering, genetic algorithm, rule induction.

Keywords: Neurons, Extract, Pattern, Decision, Technique, Consideration, Genetic

I. INTRODUCTION

There are many different methods used to perform mining task. These techniques not only required specific type of data structure but also imply certain type of algorithm approach. The technologies for generating and collecting data have been advancing rapidly. At the current stage, lack of data is no longer a problem. The explosive growth in data and databases results in the need to develop new technology and tools to process data into useful information and knowledge intelligence and automatically Data Mining. Data mining is an active research area and research is on going to bring statically analysis and AI techniques to gather to address the issues.

II. TYPES OF TECHNIQUES

A. DECISION TREE

It is a classification schema and is used to find description of several pre-defined classes and classify a data item into one of them. Decision tree model is a computational model consisting of three parts:
1. A decision tree is a tree where the root and each internal node are labeled with a question.
2. An algorithm to create the tree.
3. An algorithm that applies the tree to data and solve the problem under consideration.

1) EXAMPLE: Suppose that students in a university are to be classified short, tall, medium based on their height. Assume the data base schema is {Name, Address, Gender, Height, Age , Year}. To construct a Decision tree we must identify the attributes that are important for classification problem.

Suppose height and age and gender are chosen. The female of height 1.95 is considered tall but male of same height not considered tall. For the students of university or classification we have only gender and height because most of the student over 17year of age and can’t be considered as child. This sample form’s the bases of how the tree is constructed, one possible resulting tree is shown below:

2) Merits of Decision tree:
1. The Decision Tree algorithm produces accurate and interpretable models with relatively little user intervention.
2. The algorithm is fast, both at build time and apply time.
3. Decision trees are self–explanatory and when compacted they are also easy to follow.
4. Decision trees can handle both nominal and numeric input attributes.
DEMERITS OF DECISION TREE:
1. A sub-tree can be replicated several times.
2. Error-prone with too many classes.
3. Decision boundaries are rectilinear.

APPLICATIONS: Decision trees are successfully used to solve real-world problems in the following fields such as Banking, Industry, Medicine, Molecular biology.

B. NEURAL NETWORKS
It is a class of learning algorithms consisting of multiple nodes that communicates through their connecting synapses. Neural network is a computational model and a directed graph. A neural network is a directed graph. F=(V, A) with vertices V= \{1,2,...,n\} and A={<i , j>| 1<=i, j<=n}, with the following restrictions:
1. V is partitioned into a set of input nodes, V_i, hidden nodes, V_h, and output nodes V_o.
2. The vertices are also partitioned into layers \{1,........,k\} with all input nodes in layer 1 and output nodes in layer k. All hidden nodes are in layers 2 to k-1 which are called hidden layers.
3. Any arc <i, j> must have node i in layer h-1 and node j in layer h.
4. Arc [i, j] is labeled with a numeric value W_{ij}.
5. Node i is labeled with function f_i.

Neural network consists of multiple, interconnected cells whose behavior is based on the neurons that control the behavior of human and animals. Each neuron receives signals from system inputs or from the other neurons. Based on the signals it receives, it generates an output, which it sends to system outputs or to other neurons. Fig. shows structure of neural network.

EXAMPLE: For example, handwritten character reorganization for training a computer to pronounce English text and many real-world business problems and have already been successfully applied in many industries. Neural network are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

![Neural Network Diagram](image)

**Fig.2 NEURAL NETWORK**

1) Merits of Neural Network:
1. Neural networks can be implemented in parallel hardware.
2. Neural networks can be updated with fresh data, making them useful for dynamic environments.
3. Neural networks are very flexible with respect to incomplete, missing and noisy data.
4. Neural networks are able to approximate complex non-linear mappings.

2) Demerits of Neural Network:
1. There is no explicit set of rules to select the most suitable neural network algorithm.
2. Neural networks are totally dependent on the quality and amount of data available.
3. Poor Transparency: Neural networks operate as “black boxes”.
4. Large complexity of the network structure.

3) Applications:
Prediction – weather, stocks, disease
Classification – financial risk assessment, image processing
Data association – Text Recognition (OCR)
C. NEAREST – NEIGHBOR AND CLUSTERING

These are the oldest techniques used in data mining. Clustering is namely that records are group or cluster together and put into the same group. Neural neighbor is a prediction technique that is similar to clustering. It simply stated that objects that are near to each other will have similar prediction values as well. Thus if you know the prediction value one of them you know predict from nearest neighbor.

1) EXAMPLE: Our simple two dimensional example is given here we subsequently associate the concept of mathematical distance with physical distance on the page and do not even think about this as a possible problem. When points are defined by multiple measures whose numerical scales cannot easily be compared with each other, though defining distance appropriate may take considerable thought and experimentation.

![MBA Graduation Data Divided into Neighborhoods](image)

Fig.3 NEAREST-NEIGHBOR

2) Differences between Nearest Neighbor and clustering:
   1. Nearest neighbor technique used for prediction as well as consolidation and clustering is used mostly for consolidating data into a high level view.
   2. Nearest neighbor technique generally only uses distances metrics to determine nearness and clustering can use other metrics besides distance to determine nearness of two records for example linking points together.

3) Merits of Nearest Neighbor:
   1. Nearest neighbor can be implemented by SQL and there is no training time required so user can begin to generate prediction immediately.
   2. This technique can be represented incremental addition to existing database infrastructure.

4) Demerits of Nearest Neighbor:
   1. As the user requirement increased to require higher accuracy rates or simplified modeling proto types and sentries can be calculated and used.
   2. It does not automatically perform over training detection and pruning.

5) Applications:
   Computer vision, Data compression, Internet marketing, Pattern recognition.

D. GENETIC ALGORITHM

It is a machine learning algorithm which is based on five parts:
   1. Starting set of individual
   2. Cross over technique
   3. Mutation algorithm
   4. Fitness function
   5. Algorithm that applies the crossover technique

1) Example: In genetic algorithm reproduction is defined by precise algorithms that indicate, how to combine the given set of individuals to produce new ones. These are called crossover and mutation algorithms. Given two individuals (parents) from the population the crossover techniques generates new individuals (children) by switching subsequences of the strings. Fig shows process of crossover:-
Fig. 4 GENETIC ALGORITHM

2) Merits of Genetic algorithm:
1. The major advantage to use the genetic algorithm is that they are easily parallelized.
2. Genetic algorithm is useful for clustering, prediction and even association rule.

3) Demerits of Genetic algorithm:
1. Determining the best fitness function is difficult.
2. Determining how to do crossover and mutation is difficult.

4) Applications:
- Airlines Revenue Management, Automated design, artificial creativity, Financial data analysis.

E. RULE INDUCTION

It is one of the major data mining technique and used system. Rule induction on a database can be a massive undertaken in which all patterns are systematically pulled out of the data and then accuracy and significance calculated, telling users how strong the pattern is and how likely it is to occur again.

1) Example: For example under the MBA analysis we can generate different rules like
- If deli product is purchased then they buy the bakery products 50% of the time and this pattern occur in 2% of all shopping baskets.
- If meat and bakery product are purchased and they also buy the deli products 90% of the time and this pattern occur in 0.5% of all shopping basket.

2) Merits of Rule Induction:
1. Rule induction technique is easy to understand.
2. It is easy to deploy and to display great utility for both discovery and prediction.

3) Demerits of Rule Induction:
1. For some high dimensional domains such as MBA, where small co relations can be extremely important, rule induction system out perform all others.

4) Applications:
- Risk analysis, Credit risk decisions, To select candidate from an list of candidate by few key question.

III. CONCLUSION

Data mining is a process of finding the hidden patterns, forecasting and discovery of knowledge from the data warehouse using the different data mining techniques. A number of data mining techniques are discussed in this paper like Decision Tree, Neural Network, Nearest Neighbor & Clustering, Genetic Algorithm and Rule Induction. Each technique has its own applications and advantages/disadvantages. After the study of all the techniques it become easy to decide a technique for data mining.

ACKNOWLEDGMENT

The authors wish to thank Mrs. Rachna Rajput for their co-operation. The authors are also thankful to HOD & faculty members of CSE dept. for their support to carry out this work.

REFERENCES
[6]. http://www.slideshare.net/huongcokho/data-mining-concepts

Books:
[7]. Arun K. Pujari, Data Mining Techniques
[8]. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques
[9]. BB Aggarwal,SP Tayal, Data Mining and Data Warehousing