International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 1, Jan-Feb 2015

RESEARCH ARTICLE OPEN ACCESS

A Study on Algorithmic Approaches and Mining Methodologies In Data Mining

S. Padmapriya
Assistant Professor
Department of Computer Science
Srimad Andavan Arts and Science College
Tamil Nadu – India

ABSTRACT

Data mining finds valuable information hidden in large volumes of data that need to be turned into useful information**.** *It is considered to deal with huge amounts of data which are kept in the database***.**Data mining is the analysis of data and the use of software techniques for finding hidden patterns and regularities in sets of data Knowledge discovery from *the large data set becomes difficult. The increase in demand of finding pattern from huge data is improved by means of data mining algorithms* and techniques**.**Researchers presented a lot of approaches and algorithms for determining patterns. This paper presented various data mining algorithms and mining methods to discover valuable patterns from the hidden information.

***Keywords*:-**Data mining, Knowledge Discovery**.**

I. INTRODUCTION

Data mining is an emerging trends.The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this datam is negligible if meaningful information or Knowledge cannot be extracted from it [4].Data mining, otherwise known as knowledge discovery, attempts to answer this need. In contrast to standard Statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses. Now a day, advances in hardware technology have lead to an increase in the capability to store and record personal data.

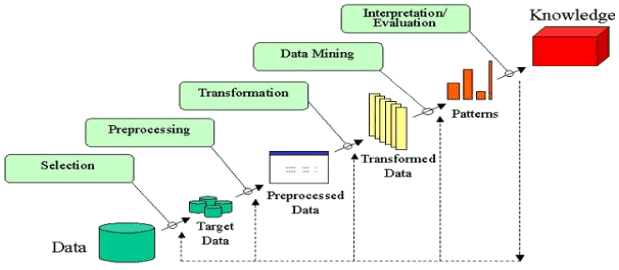

Figure:1 Data Minig Process

II. BASIC TERMINOLOGY

Data:

Data is information typically the results of measurement (numerical) or counting (categorical). Variables serve as placeholders for data. There are two types of variables, numerical and categorical.A numerical or continuous variable is one that can accept any value within a finite or infinite interval. There are two types of numerical data, interval and ratio. Data on an interval scale can be added and subtracted but cannot be meaningfully multiplied or divided because there is no true zero. For example, we cannot say that one day is twice as hot as another day. On the other hand, data on a ratio scale has true zero and can be added, subtracted, multiplied or divided (e.g., weight).A categorical or discrete variable is one that can accept two or more values (categories). There are two types of categorical data, nominal and ordinal. Nominal data does not have an intrinsic ordering in the categories. For example, "gender" with two categories, male and female. In contrast, ordinal data does have an intrinsic ordering in the categories. For example, "level of energy" with three orderly categories (low, medium and high).

Data Preparation:

Data preparation is about constructing a dataset from one or more data sources to be used for exploration and modelling. It is a solid practice to start with an initial dataset to get familiar with the data, to discover first insights into the

data and have a good understanding of any possible data quality issues. Data preparation is often a time consuming process and heavily prone to errors. The old saying "garbage-in-garbage-out" is particularly applicable to those data mining projects where data gathered with many invalid, out-of-range and missing values. Analyzing data that has not been carefully screened for such problems can produce highly misleading results. Then, the success of data mining projects heavily depends on the quality of the prepared data.

Dataset:

Dataset is a collection of data, usually presented in a tabular form. Each column represents a particular variable, and each row corresponds to a given member of the data.Data sets are classified into two types test data and training data.

III. MINING METHODOLOGY

a.Classification

Classification problems aim to identify the characteristics that indicate the group to which each casebelongs. This pattern can be used both to understand the existing data and to predict how new instances will behave[1].Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long- distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier.

b..Regression

Regression uses existing values to forecast what other values will be. In the simplest case, regression uses standard statistical techniques such as linear regression[1].Unfortunately, many real-world problems are not simply linear projections of previous values. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values.The same model types can often be used for both regression and classification. For example, the CART (Classification And Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables).Neural nets too can create both classification and regression models.

c. Logistic regression:

Logistic regression is a generalization of linear regression. It is used primarily for predicting binaryvariables (with values such as yes/no or 0/1) and occasionally multi-class variables[1]. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predict whether the event itself (the response variable) will occur, we build the model to predict the logarithm of the odds of its occurrence. This logarithm is called the log odds or the logit transformation.

The odds ratio: $\frac{\text{Probability of an event occurring}}{\text{Probability of the event not occurring}}$

d.Neural networks

Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions.(Actual biological neural networks are incomparably more complex.) Neural nets may be used in classification problems (where the output is a categorical variable) or for regressions (where the output variable is continuous). A neural network starts with an input layer , where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer . Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer . The output layer consists of one or more response variable.

Algorithm

There are different types of neural networks, but they are generally classified into feed-forward and feed-back networks.A feed-forward network is a non-recurrent network which contains inputs, outputs, and hidden layers; the signals can only travel in one direction. Input data is passed onto a layer of processing elements where it performs calculations. Each processing element makes its computation based upon a

weighted sum of its inputs. The new calculated values then become the new input values that feed the next layer. This process continues until it has gone through all the layers and determines the output. A threshold transfer function is sometimes used to quantify the output of a neuron in the output layer. Feed-forward networks include Perceptron (linear and non-linear) and Radial Basis Function networks. Feed-forward networks are often used in data mining.

A feed-back network has feed-back paths meaning they can have signals traveling in both directions using loops. All possible connections between neurons are allowed. Since loops are present in this type of network, it becomes a non-linear dynamic system which changes continuously until it reaches a state of equilibrium. Feed-back networks are often used in associative memories and optimization problems where the network looks for the best arrangement of interconnected factors.

f.Support Vector machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration[3]. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

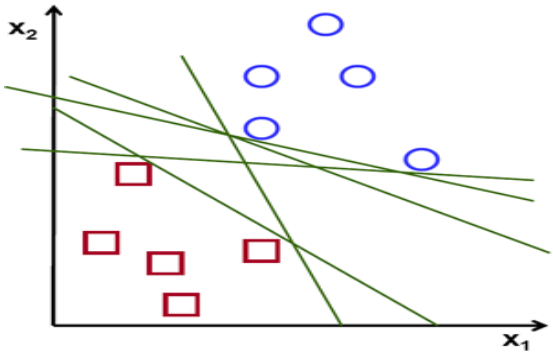

Figure2:Support Vector Method

IV. ALGORITHMIC APPROACHES

a.Genetic algorithms

Genetic algorithms are not used to find patterns per se, but rather to guide the learning process of data mining algorithms such as neural nets. Essentially, genetic algorithms act as a method for performing a guided search for good models in the solution space.They are called genetic algorithms because they loosely follow the pattern of biological evolution in which the members of one generation (of models) compete to pass on their characteristics to the next generation (of models), until the best (model) is found. The information to be passed on is contained in “chromosomes,” which contain the parameters for building the model.

b.Artificial Neural Networks

An ANN is comprised of a network of artificial neurons (also known as "nodes"). These nodes are connected to each other, and the strength of their connections to one another is assigned a value based on their strength: inhibition (maximum being -1.0) or excitation (maximum being +1.0). If the value of the connection is high, then it indicates that there is a strong connection. Within each node's design, a transfer function is built in. There are three types of neurons in an ANN, input nodes, hidden nodes, and output nodes.The input nodes take in information, in the form which can be numerically expressed. The information is presented as activation values, where each node is given a number, the higher the number, the greater the activation. This information is then passed throughout the network. Based on the connection strengths (weights), inhibition or excitation, and transfer functions, the activation value is passed from node to node. Each of the nodes sums the activation values it receives; it then modifies the value based on its transfer function. The

activation flows through the network, through hidden layers, until it reaches the output nodes. The output nodes then reflect the input in a meaningful way to the outside world.

c.The k-means Algorithm

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

1.The algorithm arbitrarily selects k points as the initial cluster centers ("means").

2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.

3.Each cluster center is recomputed as the average of the points in that cluster.

4.Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters[2].

b.Nearest Neighbour Algorithms

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors[2].k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.A shortcoming of the k-NN algorithm is that it is sensitive to the local structure of the data[2].

Choosing the Number of Clusters

One of the main disadvantages to k-means is the fact that you must specify the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance. For example, if you had a group of people that were easily clustered based upon gender, calling the k-means algorithm with k=3 would force the people into three clusters, when k=2 would provide a more natural fit. Similarly, if a group of individuals were easily clustered based upon home state and you called the k-means algorithm with k=20, the results might be too generalized to be effective.

d.K-nearest neighbor and memory-based reasoning (MBR)

When trying to solve new problems, people often look at solutions to similar problems that they have previously solved. K-nearest neighbor (k-NN) is a classification technique that uses a version of this same method. It decides in which class to place a new case by examining some number the "k" in k-nearest neighbor of the most similar cases or neighbors . It counts the number of cases for each class, and assigns the new case to the same class to which most of its neighbors belong.

e.Bayesian Algorithms

Bayesian approches are a fundamentally important DM technique. Given the probability distribution, Bayes classifier can provably achieve the optimal result.Bayesian method is based on the probability theory. Bayes Rule is applied to

calculate the posterior from the prior and the likelihood, because the later two is generally easier to be calculated from a probability model.One limitation that the Bayesian approaches can not cross is the need of the probability estimation from the training dataset.It is noticeable that in some situations, such as the decision is clearly based on certain criteria, or the dataset has high degree of randomality, the Bayesian approaches will not be a good choice. Bayes theorem plays a critical role in probabilistic learning and classification. Build a generative model that approximates how data is produced uses prior probability of each category given no information about an item. Categorization produces a posterior probability distribution over the possible categories given a description of an item

$P(C,D)=P(C|D)P(D)=P(D|C)P(C)$

$$P(C|D)=\frac{P(D|C)P(C)}{P(D)}$$

V. MODEL EVALUATION

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data mining because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data mining, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance.

Hold-Out:

In this method, the mostly large dataset is randomly divided to three subsets: Training set is a subset of the dataset used to build predictive models. Validation set is a subset of the dataset used to assess the performance of model built in the training phase. It provides a test platform for fine tuning model's parameters and selecting the best-performing model. Not all modelling algorithms need a validation set. Test set or unseen examples is a subset of the dataset to assess the likely future performance of a model. If a model fit to the training set much better than it fits the test set, over fitting is probably the cause.

Cross-Validation

When only a limited amount of data is available, to achieve an unbiased estimate of the model performance we use k-fold cross-validation. In k-fold cross-validation, we divide the data into k subsets of equal size. We build models k times, each time leaving out one of the subsets from training and use it as the test set. If k equals the sample size, this is called "leave-one-out".

Time series

Time series forecasting predicts unknown future values based on a time-varying series of predictors. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time, especially the hierarchy of periods (including such varied definitions as the five- or seven-day work week, the thirteen-“month” year, etc.), seasonality, and calendar effects such as holidays, date arithmetic, and special considerations such as how much of the past is relevant [1].

VI. CONCLUSION

Due to increase in the demanding need of valuable data and accuracy new methods and techniques are needed to be identified to improve the quality parameter. This paper describes different methodologies associated with different algorithms used to handle huge data and also it gives an overview of various techniques and algorithms used in big data sets.

REFERENCES

[1] http://www.twocrows.com/intro-dm.pdf

[2] http://en.wikipedia/wiki/k-nearest neighbour algorithm

[3] http://research.cs.queensu.ca/home/xiao/dm.html

[4] Mohammed Younus,Dr.Ahmad A.Alhamed,Khazi Mohammed Farooq,Fahmida Begum“Data Mining Modeling Techniques and Algorithm Approaches in Privacy Data”,ijarcsse volume4-2014.

[5] Rachna Somkunwar,“A study on Various Data Mining Approaches of Association Rules”,ijarcsse volume2-2012.