A Summarized Report on Data Mining and It's Essential Algorithms

K Murali Gopal,
Assoc. Prof., Dept. Of CSE,
GIET, GUNUPUR, Odisha, India,

Ranjit Patnaik
Asst. Prof., Dept. Of CSE,
GIET, GUNUPUR, Odisha, India,

***Abstract* - In this today's generation enormous amount of data stored in databases and data warehouses, for analysis the stored data for business intelligence to decision making, becomes difficult. Data mining is a process of deriving knowledge from such a huge data. In this article a summarized report on the data mining and its essential algorithms are categorized.**

Keywords—Data Mining; Classification, Apriori Algorithm,K Means Algorithm, C4.5 Algorithm, Page Rank, k-NN Classification, NAÏVE BAYES

I. INTRODUCTION

Data Mining (also known as Knowledge Discovery) is a process of analyzing large pre-existing data from different perspective and summarizing it into useful information that can be used to increase revenue, cost cutting or both [1]. It also can be considered as a computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [2]. It becomes important for today's business analysis and decision making. The use of distributed information system collects Terabytes or Petabytes of data which contain much information for growth of information, but manual finding of such information is very difficult so Data Mining helps to find information automatically.

II. CLASSIFICATION OF DATA MINING TECHNIQUES

Broadly the data Mining problems can be classified into two models. Namely Predictive and Descriptive[3,4].

Predictive model is analyzing current and historical data to make predictions about future, or other unknown events like identify the risk and opportunity. For example Cross Selling, Fraud Detection.

Descriptive model find the relationships among the data in way they exist to provide a human-interpretable patterns that describe the data. For Example Retail Rack Management.

Predictive model focus on a single customer behavior where as descriptive model identify different relationship between products and customers.

- The further division of these model as follows:
- Predictive Model
- Classification
- Regression
- Outlier Detection
- Descriptive Model
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery

Classification is the problem of finding to which subgroup or category a new item belongs. On the basis of training data the subgroup attributes are decided and the new item based on its attributes assigned to one of such subgroups.

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing the relationship between a dependent variable and one or more independent variables.

Outlier Detection is the observation of such items or events which does not confront to the excepted pattern in the dataset.

Clustering is an approach of partitioning a group of elements into more than one subgroup / cluster where elements of each subgroup are similar in characteristics based upon their inter-cluster or intra-cluster distance measure.

Association Rule Discovery is a method for discovering interesting relationship between the variables in a large dataset. It doesn't consider the order of variables either in transaction or across transaction.

Sequential Pattern Discovery is concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence.

III. ESSENTIAL ALGORITHMS

A. Apriori Algorithm

Apriori algorithm is one of the popular algorithm for association rule discovery. Association rule discovery is a technique of uncovering the relationship between the variables in a huge database.

Apriori uses an iterative approach known as a level wise search, where k-itemsets are used to explore (k+1) itemsets. The algorithm as follows:

Step 1: find 1-itemset (k=1) from the dataset by scanning and counting the appearance of the item.

Step 2: Filter the itemset which does not satisfy the support. (Prune Step)

Step 3: From the result find the next itemset (k+1) by counting.

Step 4: continue the steps 2 and 3 until no more itemset found.

The association rule can be generated by using confidence factor, which can be calculated using following formula:

confidence(A→B)=P(B/A)=support_count(AUB)/support_count(A)

where

support_count(AUB) : count of frequency of A and B in Dataset

support_count(A): count of frequency of A in Dataset.

Easy implementation and easy parallelism are the advantage but the number of scanning to a large memory resident data set is the disadvantage.

Few Advance association rule discovery algorithms are FP growth algorithm, Equivalence Class Transformation algorithms.

B. K Means Algorithm

k-Means algorithm is the most popularly used centroid based clustering algorithm. Input for this algorithm is no of cluster (K) and the data set(D) containing n items. If the K vales is undefined that this leads to NP hard problem else it will generate k clusters.[8]

In this partitioning algorithm is as follows:

Step 1: Randomly select k objects as initial cluster center.

Step 2: Find each object distance to each cluster center and assign to nearest cluster.

Step 3: Update the cluster means

Step 4: Repeat step 2 and 3 till no change in Cluster center.

The distance can be measured as follows:

Euclidean Distances:

L2 norm : d(x,y) = square root of the sum of the squares of the differences between x and y in each dimension. The most common notion of "distance."

L1 norm : sum of the differences in each dimension. (Manhattan distance).

L∞ norm: d(x,y) = the maximum of the differences between x and y in any dimension

Non-Euclidean Distances:

Jaccard distance for sets : 1 minus ratio of sizes of intersection and union.

Cosine distance: angle between vectors from the origin to the points in question.

Edit distance: number of inserts and deletes to change one string into another.

In large globular data set K means provides a faster cluster then hierarchical clustering. But disadvantage of K means algorithm is that it does not work well in case of non-globular clusters and predicating initial value of K is difficult.

Few advance k-means algorithms are K++ clustering and moving k-means algorithm.

C. C4.5 Algorithm

C4.5 is a classifier. It generate decision tree using a set of training data set using information entropy. The training data set a set of S=s1,s2,…already classified samples. Each sample si consists of p dimension vector (x1,i,x2,i,…xp,i) where xj represented attribute or feature of that class to which si belongs.

The algorithm as follows:

Step 1: If S contain one or more sample all belongs to the same class C, Then decision tree is having a leaf C.

Step 2: if S contain no sample then the decision tree contain a leaf node belongs to the class C i.e. most frequent class of the parent node.

Step 3: if T contain mixture of classes then partition T into subsets of samples heading towards a single class based on information entropy.

Find information gain for each X

Search for the best information gain.

Divide the test set into multiple subsets based on attribute X.

Put each subset as a leaf node.

Step 4: Repeat on leaf subset from step 1,2 and 3 until no more division.

After the decision tree is ready, Test with the test sample for post pruning process. Any new sample can be used to travels through the if and else rules generated by C4.5 algorithm to decide the class to which the new sample belong to.

C4.5 builds models that can be easily interpreted and implemented. It is good for small variant of data set but difficult for large variant or closed related attributed data set.

Advance algorithms of this kind are C5.0.

D. Page Rank

PageRank is a link analysis algorithm based on WebGraph. It find the importance of a page by counting the web page linked with it.

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites[Google].

The page rank can be calculated by using the following formula:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

i.e. the PageRank value for a page u is dependent on the PageRank values for each page v contained in the set Bu (the set containing all pages linking to page u), divided by the number L(v) of links from page v.

E. k-NN Classification

k-NN classification is a non-parametric, instance based learning or lazy learning. The training set consists of vectors in a multidimensional feature space each with a class label.

A new object is classified based by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. A commonly used distance metric for continuous variables is Euclidean distance.

The main advantage of k-NN classifier is simplicity. As the learning time is small but in a large data set nearest neighbor finding will take more time.

F. CART

CART (Classification And Regression Tree) is a non-parametric decision tree learning techniques for binary splitter based on one variable [5].

In this algorithm each node is split into two nodes based on impurity. The impurity can be measured by finding Misclassification Rate, Entropy, Gini Index. The splitting process can be terminated when no more splitting occur

The algorithm is simple and robust with outliers. It uses conditional variable effectively. But the tree structure is unstable. The optimality of tree may not be global optimal.

G. NAÏVE BAYES

Naive Bayes classifiers is a popular text categorization method based on applying Bayes' theorem with strong independence assumptions between the features.[6]

D is the data set and h is the Hypothesis. Then

P(h/D)= (P(D/h) P(h)) / P(D)

P(h) : Prior probability of hypothesis h

P(D) : Prior probability of training data D

P(h/D) : Probability of h given D

P(D/h) : Probability of D given h

IV. CONCLUSION

In this article the brief discussion regarding the data mining algorithms are explained. Data mining is a tool having different algorithm to find the relation among the data in a huge data. These are few important and popular algorithms.

V. REFERENCES

1. Data Mining : What is Data Mining? [http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm]
2. “Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28
3. Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan:” A SURVEY ON DATA MINING TECHNIQUES” IJARCCE Vol. 3, Issue 1, January 2014 pg 5002-5003
4. Raj Kumar, Dr. Rajesh Verma : “Classification Algorithms for Data Mining: A Survey”: IJICT Vol. 1 Issue 2 August 2012 pg: 07 - 14
5. Wei-Yin Loh :“Classification and Regression Tree Method” : 315–323, Wiley, 2008)
6. Chai, K.; H. T. Hn, H. L. Chieu; “Bayesian Online Classifiers for Text Classification and Filtering”, Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, August 2002, pp 97-104.
7. A Systematic Overview of Data Mining Algorithms by Sargur Srihari of University at Buffalo, The State University of New York.
8. Shailendra Singh Raghuwanshi & PremNarayan Arya :“Comparison of K-means and Modified K-mean algorithms for Large Data-set “ : IJCCN Volume 1, No.3, November – December 2012 pg: 106 -110.
9. Rakesh Agrawal & Ramakrishnan Srikant: “_Fast Algorithms for Mining Association Rules” .