Data Mining in Knowledge Discovery: A Comprehensive Study

Kamalpreet Bindra[1], Dr. Anuranjan Mishra[2], Suryakant[3]
[1]PHD Scholar NIU, Greater Noida
[2]HOD CSC, NIU , Greater Noida
[3]Asstt. Professor NIU, Greater Noida

Abstract

Data stored in colossal repositories are analyzed and extracted for meaningful and relevant knowledge. Data mining comprises the core algorithms that enable one to gain fundamental insights and knowledge from enormous pool of data. It is an interdisciplinary field which consists of areas such as database systems, statistics, machine learning and pattern recognition. Data mining is part of a larger process called KDD or knowledge discovery in databases, which is highly vast and comprises of some pre and post processing tasks. The steps comprising of KDD are highly iterative and interactive in nature. Upon the completion of these steps very specific and highly critical knowledge is generated. The research paper attempts to discuss context of data mining in the multidisciplinary and multi-dimensional KDD process as well as various data mining techniques such as exploratory data analysis, association, clustering and classification are discussed.

Keywords: KDD- knowledge discovery in data bases, data mining, exploratory data analysis, clusters

1 INTRODUCTION

In this age of real time technology based environment, data is goldmine. Mining of useful and critical data can result in predicting behaviours, future trends allowing business to make proactive knowledge driven decisions. Data mining can be used in many different sectors of business to both predict and discover trends, for ex, in past we were only able to answer and analyze what a company's clients had done but now we can predict what clientele will do. Knowledge discovery in databases has evolved and continues to evolve from the intersection of research fields like: machine learning, pattern recognition, knowledge acquisition in expert systems and AI. The ultimate goal is to extract high level knowledge from data sets to construct much larger data sets. As discussed data mining is a step within the entire KDD process[2]. There are two major data mining goals verification and discovery. Verification is verifying user's hypothesis about data while discovery is automatically finding novel patterns. The KDD not only addresses the extraction part of patterns through data mining, it performs some highly valuable pre-processing and post processing tasks. These tasks will be discussed in detail.

2 DATAMINING DEFINATION

Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data (Hanet .al). data mining is the process of exploration and analysis , by automatic or semi-automatic means ,of large quantities of data in order to discover meaningful patterns and rules (Berry and linof 2000). Data mining is an important component in the entire process of KDD, which helps the system to find interesting and novel patterns as well as descriptive, understandable and predictive models from large data sets[6].

Data mining consists of 5 major elements:

1. Extract, transform and load transaction into the data warehouse system.

2. Store and manage the data in a multidimensional database system

3. Provide access to business analyst and I.T. professionals

4. Analyze the data by application software

5. Present the data in a simpler form such as graph or a table.

Data mining has not only been a game changer in the world of statistical analysis and business but it has its uses in widely diverse areas also like:

- Retail marketing (eg identifying customer's buying patterns ,market basket analysis)

- Banking (e.g. detecting credit card fraud , identifying loyal customers)

- Biological data analysis

- Intrusion detection etc.

3 KDD: THE LARGER PICTURE

The KDD process is interactive and iterative, involving many steps with many decisions taken by the user. Brachman and Anand (1996)[2] gave a practical view of the KDD process, broadly some of the steps are:

1. Developing an understanding of the application domain

2. Creating a target data set.

3. Data cleaning and pre processing

4. Choosing the appropriate data mining task

5. Choosing the data mining algorithm

6. Data mining algorithm employment

7. Interpretation of mined knowledge

8. Using the discovered knowledge.

A brief discussion of each step is mandatory, in the very first step discovering and creating an understanding of the application domain is performed. Identification of the main goal of the KDD process is very important. In the second stage target data sets are created. it is like selecting data samples on which discovery is to be performed .the data may have an algebraic, geometric or probabilistic viewpoint which can play a key role in mining. During the next immediate step some data cleaning is performed like removal of noise accounting for mining fields. After this preparation, fourth step is for choosing the function of data mining which decides the purpose of the model derived by data mining algorithm (eg classification, clustering and summarization). Choosing the data mining algorithm is the next step and it involves selecting methods for deciding what all models and parameters would be appropriate and matching with the overall KDD criteria. After the mined knowledge is acquired it becomes very crucial to present it in a user understandable manner and using the discovered knowledge.

4 THE DATA MINING TECHNIQUES

There are various techniques used for accomplishing the mining task. These techniques that are used for data mining are exploratory data analysis; frequent pattern mining, clustering and classification .the research paper lays the basic foundation of these tasks.

4.1 Exploratory data analysis

The algebraic, geometric and probabilistic view point of data play a key role in data mining .given a dataset of n points in a d dimensional space , exploratory data analysis explores the numeric and categorical attributes of the data individually or jointly to extract key characteristics of the data sample. Exploratory data analysis or EDA often lays the foundation of data preparation in the KDD process. EDA typically consists of these following steps:

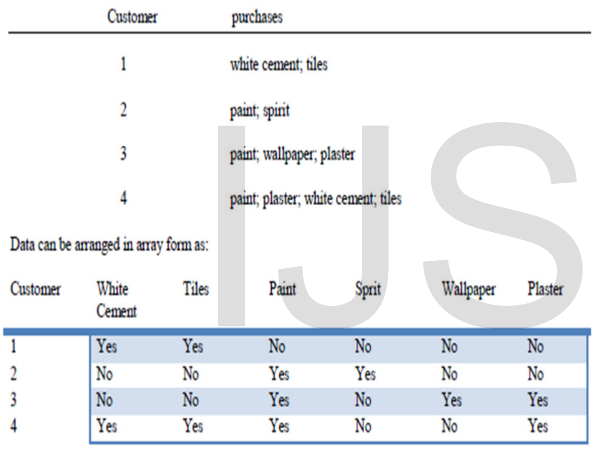
- Problem definition: the problem to be solved along with the projected deliverables should be clearly defined. Which means it should be clear how data mining project will address the problem.

- Data preparation :prior to starting any data analysis or data mining project, the data should be collected, characterized, cleaned, transformed and partitioned into an appropriate form for further processing.

- Implementation of the analysis: on the basis of the information from steps 1 and 2 appropriate analysis techniques are selected for summarizing the data. Summarization is a process where data is reduced for interpretation without deleting any important information.

4.2 Frequent mining data sets or association

Association among objects often leads to find how often two or more objects of interest to co-occur. An association rule has two parts ,an antecedent(X) and consequent(Y) like IF X THEN Y for example, 80% of those who buy music online also buy books online, 40% of those who buy mobile phones also buy power banks. In data mining, association rules are useful for analyzing and predicting customer behaviour .The prototypical application is market basket analysis. That is to mine the sets of items that are frequently bought together at a supermarket or any other database for instance as one of the reasons behind maintaining any database is to be able to find interesting patterns and trends in the data. Market baskets here play the role of a typical cart which helps to analyze customer shopping trends. Once we mine the frequent sets they enable us to find out association rules among the item sets .for example we can say that buyers who buy milk also tend to buy cereal. Association can be better understood by this example of a typical super market scenario where data consists of transaction records, each containing a set of items purchased by that customer. For example:

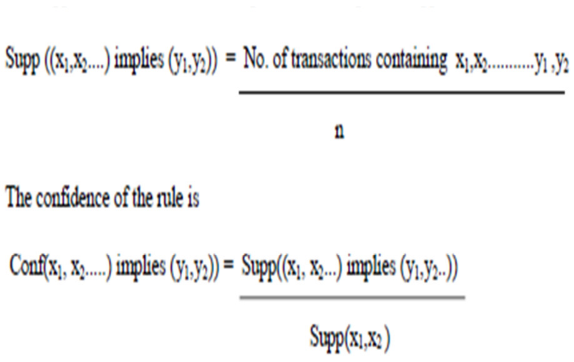


Customer	purchases
1	white cement; tiles
2	paint; spirit
3	paint; wallpaper; plaster
4	paint; plaster; white cement; tiles

Data can be arranged in array form as:

Customer	White Cement	Tiles	Paint	Sprit	Wallpaper	Plaster
1	Yes	Yes	No	No	No	No
2	No	No	Yes	Yes	No	No
3	No	No	Yes	No	Yes	Yes
4	Yes	Yes	Yes	No	No	Yes

Let P be the set of all purchases and let n be the number of transactions. Each transaction record is a subset of S. rules of the form “(x1,x2....xj)”implies “(y1,y2...yk)”are considered where x1,x2 ....y1,y2 are all elements of S.the collection (x1,x2.....xj) is called an *item set.* As stated earlier association rules are created frequent IF/THEN patterns, criteria like *confidence* and *support* are used to identify relationships. Confidence indicates the number of time the if/then statements are found to be true and support is an indication of how frequently items appear in the database. In this particular example the support of the rule is defined as:



$$\text{Supp}((x_1,x_2....)\text{ implies }(y_1,y_2)) = \frac{\text{No. of transactions containing } x_1,x_2..........y_1,y_2}{n}$$

The confidence of the rule is

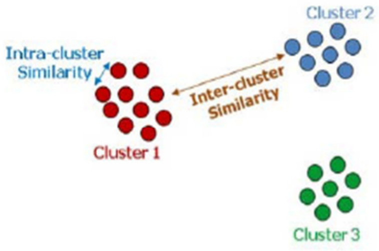
$$\text{Conf}(x_1, x_2.....)\text{ implies }(y_1,y_2)) = \frac{\text{Supp}((x_1, x_2...)\text{ implies }(y_1,y_2..))}{\text{Supp}(x_1,x_2)}$$

The Apriori algorithm*:* The Apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. Suppose there a total of m items in S. The number of subsets of S is 2m, thus to check every transaction record to see which sets it belong to require n2m checks. This is computationally infeasible when m is even of moderate size. This is an instance of “curse of dimensionality “. However if we restrict to sets with support greater than 0 the search becomes feasible. These are called frequent item sets. The first efficient algorithm for finding all sets with a given level of support was given by agarwal and srikant1994 [4], and was supposedly improved by these authors and others. Once all the sets with support greater than s0have been found and their supports recorded, it is then a straight forward matter to calculate the confidence.

5 CLUSTERING

Data clustering is the most complex yet crucial data mining technique. It is a method of grouping similar objects together. Here partitions are made called clusters .points or data with a cluster are as similar as possible where as points across clusters are as dissimilar as possible. Apart from being used in data mining and pattern recognition it can be used in much application such as:

Data compression, image analysis, bioinformatics, academics, search engines, intrusion detection etc. Data clustering is based on the similarity or dissimilarity (distance) measures between these points. Hence these measures make the cluster analysis meaningful. The high quality of clustering is to obtain high intra-cluster similarity and low inter cluster similarity as shown is the figure



that data mining algorithms should fulfil are scalability in terms of memory requirements and execution time, and consistent quality of the results as the input size grows. Especially the scalability requirement distinguishes data mining algorithms from the algorithms used in machine learning. Depending on the data and desired cluster characteristic there are different types of cluster paradigms such as:

1. Representative based algorithms
2. Hierarchical based algorithms
3. Density based algorithms
- Density based connectivity clustering
- Density functions clustering

4. Graph based
5. Spectral clustering

5.1 Taxonomy of clustering algorithms

A Rrepresentative –based clustering
Given a data set of n points in d dimensional space, D = {xi}ni=1,and given the desired number of clusters k, the main aim of representative based clustering is to partition the dataset into k groups or clusters c = (c1,c2,c3....).
One of the most popular representative algorithms are K-means and expectation maximization (EM) algorithms.

A.1 *K-means algorithm*:

K- means is a heuristic algorithm that partitions a data set into K clusters by minimizing the sum of squared distance in each cluster [BRL00].The algorithm is a greedy algorithm and it performs hard clustering which means each point is assigned to only one cluster. It has three main steps:
1) Initialization by setting seeding points.
2) Dividing all data points in K clusters based on k.
3) Updating K centroids based on newly formed clusters, it can be shown that the algorithm always converges after several iteration of the steps 1 and 2.
The strength of this algorithm lies in its computational efficiency. K-means is the most widely used and popular clustering algorithm.

A.2: *Expectation maximization Clustering*:

The K-means approach is an example of a hard assignment clustering where each point can belong to only one cluster , this approach can be generalized to consider soft assignment of points to clusters , so that each point has a probability of belonging to each cluster .

B .*Hierarchical clustering*

Given *n* points in a *d* dimensional space, the goal of hierarchical clustering is to create a sequence of nested partitions, which can be easily visualized via a tree or hierarchy of clusters also called the dendrogram. The clusters in the hierarchy can be at the lowest level of the tree, consisting of each point in its own cluster and at the highest point (the root) consisting of all points in one cluster. Both of these may be considered trivial clusterings.at some point intermediate meaningful clusters can be formed. There are two main approaches to mine hierarchical clusters, *agglomerative* and *divisive*. Agglomerative strategies work in a bottom-up manner. Which means starting with each of the n points in a separate cluster, they repeatedly merge the most similar pair of clusters until all points are members of the same cluster. Divisive strategies are just the opposite, working in a top down manner, starting with all the points in the same cluster, they recursively split the clusters until all points are in separate clusters.

C Density based clustering

The representative based clustering methods like k-means and EM are suitable for finding convex clusters. On the other hand for non convex shaped clusters, density based algorithms such as DBSCAN(density based spatial clustering of applications with noise) are popularly used. it was proposed by Martin Ester,Hans-PeterKrigel,JorgSander and Xiawoei in 1996.

D Graph based and spectral clustering

The goal of graph based clustering is to cluster over graphs .given a graph, the notes are clustered by using edges and their weights, which represent the similarity between the incident nodes. This type of clustering is related to divisive hierarchical clustering, as many methods partition the set of nodes to obtain the final clusters using pair wise similarity matrix between nodes. Spectral clustering on the other hand treats clustering as a graph partitioning problem without making specific assumptions on the form of clusters. Cluster points using eigenvectors of matrices are derived from data, data are mapped to a low dimensional space that are separated and can be clustered.

6 CONCLUSIONS

Data mining has the most important and promising features of interdisciplinary developments in information technology. . The paper is an attempt to establish that data mining is so much more than just data analysis or simple extraction. Data mining techniques can answer business questions that were too time consuming to answer previously. The future in data mining is full of developing algorithms that enable user to cluster patterns more accurately resulting in more specific and user desired trends. Data mining's application will enrich human life in various fields such as business, education, medical field, scientific field and politics.

REFERENCES

1.Mrs.Bharti M. Ramageri , "Data mining Techniques and Applications ", Indian journal of computer science and engineering ,Vol 1. No.4 , pp 301-305
2. Usama Fayyad ,Gregory Piatetsky – Shapiro ,"From data mining to knowledge discovery in data bases"
3.R .agarwal and R.srikant 1994 .fast algorithms for mining association rules in large databases .proceedings of the 20th international conference on very large databases , pages 487-499
4.Mohammad J. Zaki , Wagner Meira jr. , datamining and analysis:fundamental concepts and algorithms.
5.Sangeeta goele, Nisha chanana , " data mining trends in past current and future ", International journal of computing and business research , in Proc -1 society 2012.
6. Hemlata Sahu ,Shalini Sharma ,Seema gondhalkar ,"A brief overview on data mining survey", International journal of computer technology and electronics engineering , vol1. Issue 3.
7.D T pham ,S S dimov "Selection of K in K-means clustering ",manufacturing engineering centre ,Cardiff university ,UK
8.Xindong Wu,Vipin kumar... "Top 10 algorithms in data mining",Springer –verlag 08.