ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

Data Mining Technique, Method and Algorithms

Deepti
Dept. of CSE, Guru Nanak Dev University, Amritsar, Punjab, India

Abstract

Data mining is a technique of finding and processing useful information from large amount of data. The paper covers all data mining techniques , algorithms and some organisations which have adopted data mining technology to have better information about business patterns.

Keywords

Data Mining, Clustering, Data Mining Applications, Algorithms

I. Introduction

Data mining is a process of extraction of useful information and patterns from huge data. Also known as knowledge discovery process.in this we primarily concentrate on cleansing of data. it is a powerful concept for analysing data. Process of analysing interesting patterns from bulk of data. It is a logical process by which one can search through large amount of data in order to find large amount of info.

There are three step involved in data mining:

- Exploration
- Pattern Identification
- Deployment

A. Exploration

In this step data is cleaned and then it is transformed into the required form.

B. Pattern

Once the data is explored ,refined and defined in pattern identification we form identify and choose the pattern to make best predictions. By using pattern recoganisation technologies and statistical and mathematical techniques to shift through warehouse information, it helps recognise significant facts, relationships, trends, patterns, exceptions and anomalies that are otherwise go unnoticed

C. Deployment

Patterns are deployed for desired outcome In the present framework there is bulk of data which is being collected and stored in databases everywhere across the world. As there is large amount of data is presents so it becomes very complex process to extract the important and useful information from it.

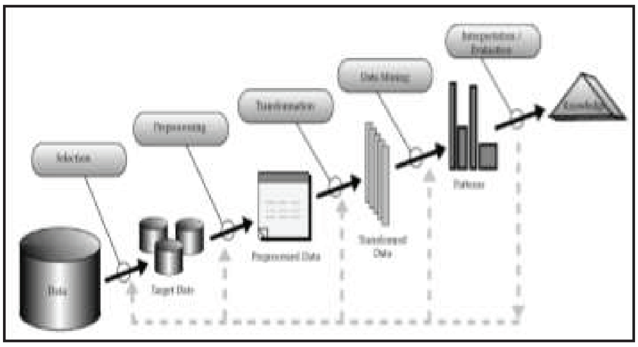

Fig. 1: Knowledge Discovery Process

II. Data Mining Techniques

Techniques that are used in data mining and data recovery and examples for how to use various tools to form data mining.

A. Association Rule

This rule refers to find correlation between larget set of data that has been frequently used. This rule also helps businesses to make certain decisions such as designing catalogues, marketing and customers shopping behaviour analysis.

Types of Association Rules :

- Multilevel association rules
- Multi-dimensional association rules
- Quantities association rules

B. Classification

In this technique we employ a pre-classified to create a design that can define the population of records. Fraud detection and credit-risk are application of this type of analysis.

3. Clustering

By analysing one or more attributes we can cluster individual block of information together to form a structure opinion. Clustering is useful to check different data as it correlates with other data so we can see the similarity and the ranges agree.

4. Prediction

Regression technique can be used for prediction. This analysis can be used to form the connection between one or more independent and dependent variables. In data mining there are two kinds of variables – independent variables and response variables. Independent variables are those in which attributes is already known and in response variables we predict the attributes.

5. Sequential Patterns

These are useful for similar fashion or regular occurrence of alike events.

6. Decision Trees

A decision tree is a structure that includes root node, branches and leaf node each internal node denotes a test on an attribute and each branch denotes the outcome of a test and each leaf node holds a class label . the topmost node in the tree is the root node . decision tree is often used with classification system to attribute type information with predictive system.

7. Machine Learning

Machine learning produces comparable predictive accuracy. its high performance than statistical methods can be attributed to the fact that it is free from parametric and structure assumptions.

8. Neural Networks

Tt is an information processing paradigm that is inspired by the way biological nervous system works. It is composed of large number of highly interconnected processing elements working on unison to solve specific problems. It is configured for specific applications like pattern recognition or data classification.

9. Genetic Algorithm

A genetic algorithm is a method of solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The algorithm repeatedly modifies a population of individual solutions at each step. The genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the children for next generation. Genetic algorithm are used to solve the problem that are not well suited for standard optimization algorithms including problems in which the objective function is discontinues non diffentiable or highly nonlinear.

A. Fuzzy Logic

Fuzzy logic is an extended version of classical logical system which provides an efficient conceptual model for handling problem of knowledge representation in an atmosphere of uncertainty and imprecision.

Chractestics of fuzzy logic:

1. Absolute reasoning is viewed as limited case of approximate reasoning
2. In fuzzy logic everything is a matter of degree.
3. Any logical system can be fuzzyfied
4. In fuzzy logic , knowledge is interpreted as a collection of elastic or equality, fuzzy constrained on a collection of variables.

III. Data Normalization

There are several approaches for normalization. the most popular methods are Z-score data Normalization, data normalization by Decimal Scaling and Min-Max Normalization. The Min-Max Normalization performed a linear transformation on the original data.

A. Min-max Normalization

Suppose that minx and maxx are the minimum and maximum of feature X. We would like to map interval [minX, maxX]. We would like to map interval [minX,maxX] into a new interval [new_minX, new_maxX]. Consequently, every value v from the originalinterval will be mapped into value new_v using the following formula:

new−v = v−minx mixx−minx

(new−maxx −new−minx)+new−minx

B. Z-score Normalization

Z-score normalization also called zero-mean normalization. The values of attribute X are normalized using the mean and standard deviation of X. A new value new_v is obtained using the following expression:

new −v =

v−ux σx C. Normalization by decimal scaling:

Normalizes by moving the decimal point of values of feature X. The number of decimal points moved depends on the maximum absolute value of X. A modified value new_v corresponding to v is obtained using new −v = v 10c where c is the smallest integer such that max (|new_v|) < 1. The area of data mining are used various types clustering approaches. But every clustering technique has some advantage and disadvantage. Every clustering technique is not appropriate for all the condition.

IV. Enabling Technologies

The actual implementation of Clustering Algorithms requires certain technologies. Some of the available technologies enabling Clustering Algorithms are:

A. K-means Clustering Technique

K-mean clustering is a simple partitioning algorithm. It partitioned 'n' data objects into K sets of clusters for resulting the low inter cluster similarity and high intra cluster similarity. Cluster similarity is measured by the mean value of the objects in a cluster, which can be called as the cluster's centroid.

Algorithm:
Input: the number of cluster k and a database containing n objects.
Output: A set of k clusters that minimizes the squared error criterion.
Procedure:
Step 1. Set centre of the clusters.
Step 2. Attribute the nearest group to every data
Step 3. Set the place of every group to the mean value of all data-points which fit in to thatgroup
Step 4. Recap steps 2-3 until all object is not classify

Disadvantages: The disadvantage of K Means clustering technique.

- The learning algorithm provides the local optima of the squared error function.
- Applicable only when mean is defined.
- Unable to handle noisy data and outliers.
- Algorithm fails for non-linear data set.

B. K- Medoids Clustering Method

The k- medoids algorithm is a clustering algorithm related to the k-means algorithm. Difference between K-Means clustering and K-Medoids Clustering:

K-means Compute group centre but in Kmedoids clustering each group's centroid is denoted by a point with in the groups. Kmeans is less strong than K-medoids in existence of noise because a medoids are less effected by noisy values. So both clustering algorithms are not gives good performance for noisy data.

C. Hierarchical Clustering Technique

In general, there are two types of hierarchical clustering technique:

1. Agglomerative Hierarchical clustering
2. Divisive Hierarchical clustering

1. Agglomerative Hierarchical clustering:

- Bottom-up strategy
- Each cluster starts with only one object
- Clusters are merged into larger and larger clusters until All the objects are in a single cluster
- Certain termination conditions are satisfied

Algorithm:

Step 1. Find the two closest objects and merge them into cluster
Step 2. Find and merage the next two closest points, where a point

is either an individual object or a cluster of objects.
Step 3. If more then one cluster remains, return to step 2.

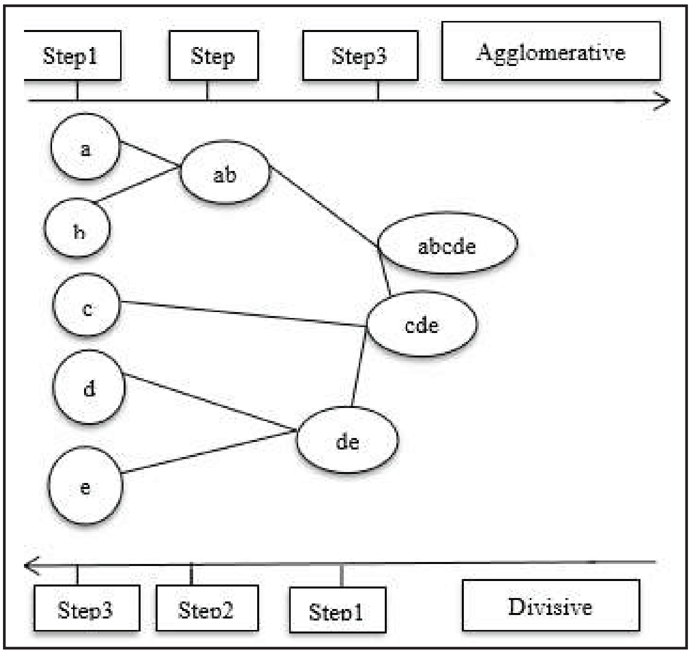

Fig. 2:

2. Divisive Hierarchical Clustering

- This top-down approach of Hierarchical clustering.
- Start with all objects in one cluster
- Clusters are subdivided into smaller and smaller clusters until
- Each object forms a cluster on its own
- Certain termination conditions are satisfie

D. Grid-based Algorithm

Grid Based methods quantize the object space into a finite number of cells that form a grid structure. All of the cluster operations are performed on the grid structure the main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.

There are two approaches:
STING: statistical information Grid approach.
CLIQUE: clustering high-dimensional space
STING: statistical information Grid approach: The statistical parameter can be used in a top-down, gridbased. Methods as follows. First a layer within the hierarchical structure is determined from which the queryanswering process is to start. This layer typically contains a small number of cells. for each cell in the current layer. This process is repeated until the bottom layer is reached.

Advantage

- The grid-based computation is queryindependent
- The grid structure facilitates parallel processing incremental updating
- Sting goes through the database once to compute the statistical parameters of the cells .
- The time complexity of generating clusters is o(n),where n is the total number of objects.

CLIQUE: clustering high-dimensional space: The CLIQUE (CLustering In QUEst) clustering algorithm integrates density-based and grid-based clustering.

- Given a large set of multidimensional data points ,the data space is usually not uniformly occupied by the data points.
- A units is dense if the fraction of total data points contained in it exceeds an input model parameter.

E. Model-based methods:

Model-based clustering methods attempt to optimize the fit between the given data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. Model-based clustering methods follow two major approach:

1. Statistical approach
2. Neural network approach

1. Statistical Approach

Conceptual clustering is a form of clustering in machine learning that given a set of unlabelled object ,produces a classification scheme over the objects. unlike conventional clustering, which primarily identifies groups of like objects, conceptual clustering goes one step further by also finding characteristic descriptions for each group. clustering is a twostep process First, clustering is performed, followed by characterization. Here, clustering quality is not solely a function of the individual object. Rather, it incorporates factors such as the generality and simplicity of the derived concept descriptions.

2. Neural Network Approach

The neural network approach to clustering to clustering tents to represent each cluster as an exemplar. New objects can be distributed to the cluster whose exemplar is the most similar, based on some distance measure. The attributes of an object assigned to a cluster can be predicted form the attributes of the cluster exemplar.

F. Density-based Algorithm

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance. This is called distance-based clustering.

Issues

Data transformation issues:

- What measure of similarity and dissimilarity should be used?
- Should the data be standardized?
- How should non equivalence of metric among variables be addressed?
- How should interdependencies in the data be addressed?

Solution Issues

- How many clusters should be obtained?
- What clustering algorithm should be used?
- Should all cases be included in a cluster analysis or should some subset be ignored?

Validity Issues

- Is the cluster solution different from what might be expected by chance?
- Is the cluster solution reliable or stable across samples?

- Are the cluster related to variables other than those used to derive them? Are the clusters useful?

Variable Selection Issue

- What is the best set of variables for generating a cluster analytic solution?

V. Major Challenges

Visualization is the critical challenge of cluster visualization.

- Visualizing large and multidimensional datasets
- Providing a clear overview and detailed insight of cluster structure
- Having linear time complexity on data mapping from higher dimensional to lower dimensional space
- Supporting interactive cluster visual representation dynamically

VI. Future Scope

- Large scale data sets are usually distributed as: Physical Logical
- Most of the algorithms are scan data many times, so they are not scalable.
- Present algorithms used only a static load balancing based on the initial data decomposition, and they assumed a homogeneous dedicated environment.
- Most of the methods partition the data base horizontally in equal size – blocks.

VII. Conclusion

This paper deals with study of different types of clustering algorithms. It first defines the data mining process which is the method of finding usable information from a huge amount of databases. Then it defines the clustering process which is the procedure of the objects in groups whose members contain some kind of similarity. After that a detailed study of clustering algorithms and their comparison in different views are examined. I am giving issues and challenges which is helpful in future researchers to carry on their work.

References

[1] KM Archana Patel and Prateek Thakral,"The Best Clustering Algorithms in Data Mining", International Conference on Communication and Signal Processing, April 6-8, 2016, India.

[2] Md Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar,"Improvement of K-means Clustering algorithm with better initial contriods based on weighted averagr", 7th International Conference on Electrical and Computer Engineering, 2012, pp. 647-650.

[3] Jiawei Han, MichelineKamber,"Data Mining: Concepts and Techniques", Second Edition

[4] Rama. B et. Al,"A Survey on clustering Current status and challenging issues", International Journal on Computer Computer Science and Engineering, Vol. 02, No. 09, pp. 2976-2980, 2010.