CLUSTERING AND CLASSIFICATION ALGORITHMS IN DATA MINING

[1]Mrs.M.Dukitha, [2]Dr.A.Banumathi

[1]Assistant Professor, [2]Assistant Professor,

[1]Department of MCA, Er.Perumal Manimekalai College of Engineering, Hosur

[2]Department of Computer Science, Government Arts College, Karur.

ABSTRACT

Clustering and classification of data is a difficult problem that is related to various and applications. Challenge is greater as input space size become larger and feature scales are different from each other. The term "classification" is often used as an algorithm program for *all* data mining tasks. Instead, it is best to use the term to refer to the group of supervised learning algorithms used to search interesting data patterns. While classification algorithms become very hip and ubiquitous in data processing analysis, it is just but one of the many types of algorithms available to solve a specific type of data mining task. In this paper various clustering and classification algorithms are going to be address in detail. A detailed review on existing algorithms will be complete and the scalability of some of the existing classification algorithms will be examine.

KEYWORDS

Data mining, clustering, classification, supervised learning, scalability.

I. INTRODUCTION

There are so many methods for data classification. usually the selection of a particular method can depend on the application. The selection of a testing method for data classification may depend on the volume of data and the number of classes present in that data. further, the classification algorithms are designed in a custom manner for a specific purpose to solve a particular classification situation. The earlier separation between classification and clustering is that classification is used in supervised learning method where predefined labels are assigned to instance by property on the different, clustering is used in unsupervised learning where similar instances are grouped, based on their features. Data mining also called knowledge discovery in large data allow hard and organization decision by assemble, accumulate, Analyzing and accessing business data. It uses variety of tools like question and reporting tools, analytical processing tools.

II. CLUSTERING

clustering is the task of group a set of objects in such a way that objects in the same group called a cluster are more parallel in some sense to each other than to persons in other groups clusters.

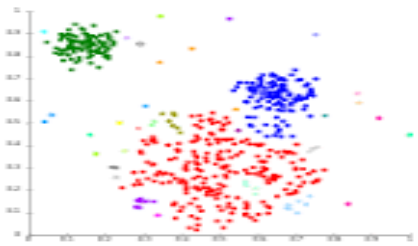


FIG1:Example of clustering

A. TYPES OF CLUSTERING

- **Hard Clustering**: In hard clustering, each data point either belongs to a cluster fully or not. For example, in the above example each client is put into one group out of the 10 groups.
- **Soft Clustering:** In clustering, instead of putting each data point into a separate cluster, a probability or probability of that data point to be in those clusters is assigned. For example, from the above scenario each custumer is assigned a

probability to be in either of 10 clusters of the retail store.

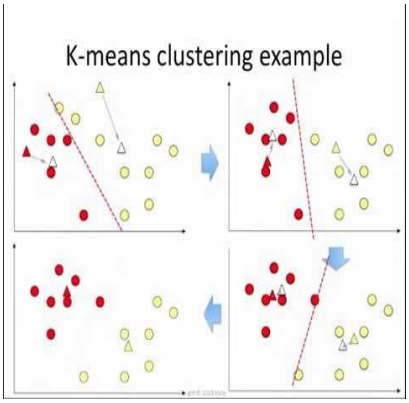
III. TYPES OF CLUSTERING ALGORITHM

- **Centroid models**: These are iterative clustering algorithms in which the idea of similar is resultant by the usefulness of a data point to the centric of the clusters. K-Means clustering algorithm is a trendy algorithm that fall into this group. In these models, the number of clusters required at the end has to be mention previous, which makes it central to have previous in order of the dataset. These models run iteratively to find the imperfect optima.
- **Distribution models:** These clustering models are based on the idea of how feasible is it that all data points in the cluster fit in to the same sharing. These model often experience from over fitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate standard distributions.
- **Density Models:** These models search the data freedom for areas of varied thickness of data points in the data space.
  It isolates numerous totally different density regions and assign the information points within these regions within the same cluster. examples of solidity models are DBSCAN and OPTICS.

A. K –MEANS ALGORITHM

K-means agglomeration could be a kind of unattended learning, that is employed once you have unlabeled knowledge (i.e., knowledge while not outlined class or groups). The goal of this algorithmic program is to seek out cluster within the data, with the quantity of teams symbolize by the variable K. The algorithmic program works iteratively to assign every datum to at least one of K teams supported the options that ar provided. knowledge points ar cluster supported feature distinction. The results of the K-means agglomeration algorithmic program are:

- ✓ The centroids of the K cluster, which may be use to label new knowledge
- ✓ label for the preparation data (each datum is assign to one cluster)



B. HIERARCHICAL CLUSTERING

Hierarchical bunch, conjointly called hierarchical cluster analysis, in these algorithmic rule that teams similar objects into teams referred to as clusters.
The end may be a set of cluster, wherever every cluster is distinct from one another cluster, and also the objects among every cluster ar loosely almost like one another .
It has two types:

- ❖ **Agglomerative:** This is a "bottom-up" approach: each inspection starts in its own cluster, and pair of clusters are merged as one moves up the hierarchy.
- ❖ **Divisive:** This is a "top-down" approach all observations start in one cluster, and splits are perform recursively as one moves down the ladder.

FIG2:Hierarchical clustering

IV. APPLICATIONS OF CLUSTERING

- ➢ Recommendation engines
- ➢ Market segmentation

- Social network analysis
- Search result grouping
- Medical imaging
- Image segmentation
- Anomaly detection

V. CLASSIFICATION

Classification construct the classification model by using training data set. Classification predict the value of classify attribute or class label.
For example: Classification of predict approval on the basis of customer data.
University gives class to the students based on marks.
If x >= sixty five, then top quality with distinction.

If 60<= x<= 65, then First class.
If 55<= x<=60, then Second class.
If 50<= x<= 55, then Pass class.

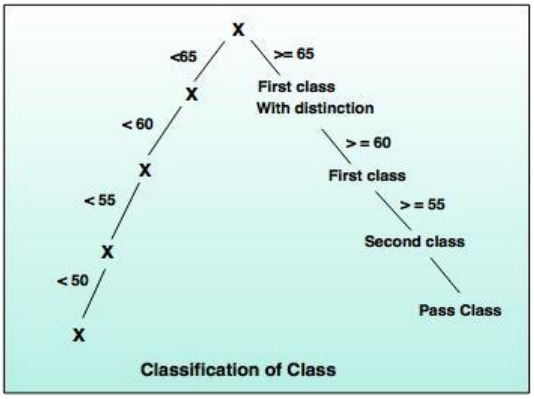


FIG3:Example of Classification

A. Decision tree:

- A decision tree performs the organization in the form of tree structure. It breaks down the dataset into small subsets and a
- choice tree can be designed simultaneously.
- The final result is a tree with decision node.

For example:

The following decision tree can be designed to declare a result, whether an applicant is eligible or not eligible to get the driving license.

Classification Requirements

- **The two important steps of classification are:**

1. Model construction

- A predefine class label is assigned to every sample tuple or object. These tuples or subset data are known as preparation data set.
- The constructed model, which is based on training set is represent as categorization rules, decision trees or mathematical formulae.

2. Model usage

- The constructed model is used to perform classification of unknown objects.
- A class label of test sample is compared with the resulting class label.
- Accuracy of model is compared by calculating the percentage of test set samples that are correctly confidential by the construct model.
- Test sample data and training data model are at all times different.

VI. DECISION TREE INDUCTIONMETHOD

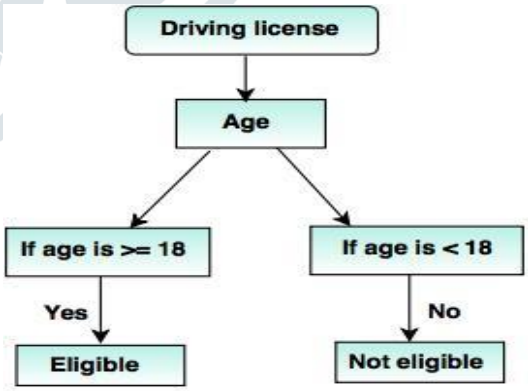


FIG4:Decision tree

A.Tree Pruning

- To avoid the over fitting trouble, it is necessary to prune the tree.
- Generally, there are two potential while constructing a decision tree. Some record may contain noisy data, which increase the

size of the decision tree. Another opportunity is, if the numbers of training examples are too small to produce a representative sample of the true target function.

- Pruning can be possible in a top down or bottom up fashion.

VII. CONCLUSION

This paper presents comprehensive description of data mining and its techniques and best methods and altos for techniques. Today all IT professionals engineers and researchers are operational on big data. Big data is term of concerning large volumes of complex data sets .The high performance computing paradigm is required to solve the problem of big data. Classification algorithms and significance of evolutionary computing (genetic programming) approach in scheming of efficient classification algorithms for data mining. Most of the earlier studies on data mining applications in various fields use the range of data types range from text to images and stores in mixture of databases and data structures. The different methods of data mining are used to take out the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important duty in this process and needs the knowledge of the field. Several attempts have been made to plan and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is compulsory. The domain experts shall be guide by the system to efficiently apply their information for the use of data mining systems to generate required knowledge.

The domain knowledgeable area unit needed to see the range of knowledge that ought to be collected within the specific drawback domain, choice of specific knowledge for data processing, cleanup and Transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge production. Most of the domain specific data mining applications

REFERENCE

[1] Holsheimer, M., Kersten, M., Mannila, H., Toivonen, H.
"A Perspective on Databases and Data Mining",
Proceedings KDD '95.

[2] Carpenter, G.A. & Grossberg, S. (2003), Adaptive
Resonance Theory, In M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, Second Edition.Cambridge, MA: MIT Press

[3]Ali,showkat and Kate A.Smith"On learningalgo selection for claasification"Applied soft completing 2006.

[4].https://docs.oracle.com/cd/B28359-01/datmine.111/b28129/process.htm

[5].https://en.wikipedia.org/wiki/data-mining.