A Survey of Clustering Mechanism and Research Challenges

Ritesh C. Sonawane[1], Dr. Hitendra D. Patil[2]

[1]*S.S.V.P.S B.S. Deore College of Engineering,Dhule*

[2] *Principal ,S.S.V.P.S B.S. Deore College of Engineering,Dhule*

[1]*ritesh.becomp@gmail.com,*
[2]*hitendradpatil@gmail.com*

Abstract

Clustering is find groups of data that are similar. Clustering is classification of objects into different groups. It is common technique for statistical data, machine learning and computer science analysis. Clustering is a kind of unsupervised learning. Clustering is organizing objects into groups. In this paper the various clustering techniques are discussed. Clustering techniques grouping the content of a website or product, segmenting customers or users, creating image segments to be used in image analysis application. Clustering is the technique segment the data to assign each training set. Clustering is the classification of objects into different group, or more precisely, the partitioning of a data set into subsets(cluster), so that the data in each subset(ideally) share some common trait- often according to some defined distance measure.

Keywords: Imbalanced-ratio(Imbr), Expectation-Maximization(EM),Gaussian Mixture Model(GMM)

I. INTRODUCTION

Clustering is process of grouping a set of objects into classes of similar objects. The purpose of clustering Segment the data to assign each training example to a segment. Classification and prediction are two very important forms of data analysis which are used to extract model describing important data classes or to predict future trends. The classification algorithms are likely to classify given inputs in some finite number of classes based on some attributes. These attributes are known as classifying attribute. The main difference between classification and clustering is classification classifying the data with the help of class labels and clustering is similar to classification but there are no class labels. Classification is supervised learning and clustering is unsupervised learning. Clustering methods are mainly suitable for the investigation of interrelationships between samples to make a preliminary assessment of the sample structure. It is required because it is very difficult for human to understand data in a high-dimensional space. There are many real world application of clustering such as In Biology it is needed for taxonomy of living things, In information retrieval it can be used for document or multimedia data clustering, It can also be used in market application. Typical clustering algorithms work nicely on relatively smaller data sets but huge databases may contain millions of data objects. Therefore we need highly scalable clustering algorithms for huge databases. Clustering technique ability to deal with different types of attributes and ability to deal with noisy data. Low-dimensional data handles the clustering algorithm and humans are good at judging the quality of clustering. Finding clusters in a high dimensional space is challenging. One of the most important task in clustering is to identify the types of data that often occur in cluster analysis. After identifying data how to preprocess them for this type of analysis.

It divided into two types of learning namely, supervised learning and unsupervised learning.

A) **Supervised learning -** Machine learning technique whereby a system uses a set of training examples to learn how to correctly perform a task.

B) **Unsupervised learning** – It is a class of problems in which one seeks to determine how the data are organized.

II Distance Measurement Method

Similarity can also be measured in terms of the placing of data points. By finding the distance between the data points, the distance/difference of the point to the cluster can be found.

1)Euclidean Distance

It is also known as 2-norm distance. One of the most common distance measure in published studies in that research area is the Euclidean distance. We can define the Euclidean distance between two points p and q as the length of the line segment pq. If p=(p1,….,.pn) and q=(q1,…….,qn) are two points. These two points in Euclidean n-space, then the distance from p to q in Cartesian coordinate is given as

$$d(p,q) = \sqrt{(p1-q1)^2 + (p2-q2)^2 + ... + (pn-qn)^2}$$

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

2)Manhattan Distance

It is also known as taxicab norm or 1-norm.In this method distance is calculated between two points is the addition of the differences of their coordinates. For points p and q where p=(p1,p2….,pn) and q=(q1,q2……..,qn)

$$d(p,q)= (|p1-q1|+ |p2-q2| +---------+ |pn-qn|)$$

3)Minkowski Distance

It is generalization of Manhattan distance and Euclidean distance. It is defined as-for points p and q where p=(p1….pn) and q=(q1……qn)

$$d(p,q)=(|p_1-q_1|^b +|p_2-q_2|^b +------+ |p_n-q_n|^b)\frac{1}{b}$$

Where b is positive integer. In some literature if b=1 it is considered as Manhattan distance and if b=2, it is considered as Euclidian distance.

3)Cosine Distance

Distance between vectors d_1 and d_2 captured by the cosine of the angle x between them.

Note – this is similarity, not distance.

No triangle inequality for similarity.

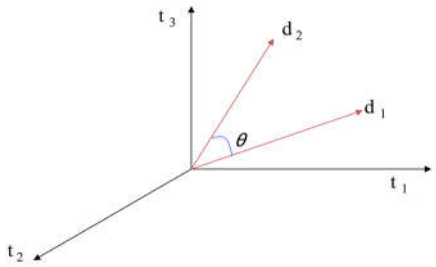

Fig 2.1 Cosine Distance

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

Cosine of angle between two vectors. The denominator involves the lengths of the vectors.

3)Tanimoto Distance

Definition:

1.value range: [0,1]

2.Tc is also known as Jaccard coefficient

3.Tc is the most popular similarity coefficient

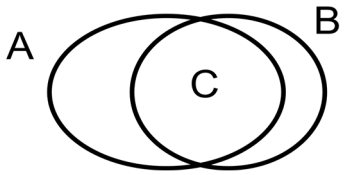

Fig.2.2 Tanimoto Distance

$$s(\mathbf{A}, \mathbf{B}) = \mathrm{Tc}(\mathbf{A}, \mathbf{B}) = \frac{c}{a + b - c}$$

III Clustering Algorithm

Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields. There are varieties of clustering algorithm.

Existing clustering algorithms:

1)K-means - It is an algorithm to classify or to group your objects based on attributes/features into K number of group or partition or cluster . K is positive integer number. The K mean algorithm is a centroid based portioning technique. The K means algorithm attempts to classify the given data sets or observations into k clusters. The K mean algorithm is iterative in nature. Let $x_{1,}$…….,x_n are data points and each data points will be assigned to one and only one cluster. Limitation of k-means clustering we need to know K in advance, It tends to go to local minima that are sensitive to the starting centroid. Another limitation of K-means it is disjoint and exhaustive and does not have notion of outliers. Outlier problem can be handled by k-medoid or neighborhood-based algorithm.

2) Mean-Shift Clustering - It is a sliding-window-based algorithm is a nonparametric clustering technique that attempts to find dense areas of data points. This algorithm is based on centroid, is to locate the center points of each group/class, which works by updating candidate for center points to be the mean of the points within the sliding-window.

3)**Density-based clustering (DBSCAN) –** Density based clustering algorithms have been developed to find clusters with arbitrary shape. The density based methods look upon clusters as a dense regions of objects in the data space which are separated by regions of low density. In density based method connectivity analysis is used to grow cluster. This method based on connected regions with sufficiently high density. To identify clusters algorithm looks at the density of points. Regions with high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of outliers. This algorithm is particularly suited to deal with large datasets, with noise, and is able to identify clusters with different sizes and shapes. DBSCAN is dependent on two important main concept density reach ability and density connect ability. These concept depend on two input parameter the size of epsilon neighborhood e and the minimum points in cluster m. The key idea of the DBSCAN algorithm is that the neighborhood of a given radius has to contain at least a minimum no. of points for each point of cluster. The density neighborhood has to exceed some predefined threshold.

4) Expectation-Maximization (EM) using Gaussian Mixture Models (GMM) – Gaussian Mixture Models gives flexibility than K-means. With Gaussian Mixture Models it is assumed that the data points are Gaussian distributed, this is a less restrictive assumption than saying they are circular by using the mean. It begin by selecting the number of clusters and after that randomly initializing the Gaussian distribution parameter for each cluster. Given these Gaussian distributions for each cluster, compute the probability that each data point belongs to a particular cluster. The cluster a point is to the Gaussian center. The more likely it belongs to that clusters.

5) Hierarchical Clustering – This clustering method consist of single cluster is series of partitions containing all objects to n clusters each containing a single object. It falls into two categories, top-down or bottom-up.

Bottom up algorithm treat each data point as a single cluster at the outset and then successively merge pairs of clusters until all clusters have been merged into a single cluster that contains all data points.

In agglomerative hierarchical clustering a series of partitions of the data, P_n , p_{n-1},…..,p_1 is produced. The first P_n is made up of n single objects cluster, and last p_1 is made up of single group containing all n cases. At each stage this method joins together the two clusters which are closest to each other. In first step it joins together the two objects that are closest to each other, since at the initial stage each cluster has only one object.

Divisive clustering separates n objects successively into better quality grouping. Dendogram is used to represent hierarchical clustering.

IV Clustering Challenges

Clustering in machine learning is a key for innovation and has a high potential for value creation. There are huge opportunities for example any small scale or large scale industry willing to provide their services through machine learning. There are also challenges like data collection, arrange the data in proper format, divide the data as per available category. Imbalanced learning occurs whenever some type of data distribution significantly dominate the instance space compared other data distribution. Data may be categorized depending on its Imbalance Ration (ImbR) which is defined as the relation between the majority class and minority class instances, by

ImbR = Negative instance / Positive instance

Where, Negative instance is the number of instances belonging to the majority class, and Positive instance is the number of instances belonging to the minority class. When Imbr value is greater than 1 that respective dataset is known as imbalanced.

Clustering Challenges

1] Machine learning algorithms struggle with accuracy because of the unequal distribution for dependent variable.

2] The accuracy of clustering must be increase.

3] Machine Learning algorithms should identify that data set are balanced or imbalanced for clustering.

4] Performance metrics such as precision, recall or F-score must be increase.

To increase the accuracy of the system by reducing instances which are belonging to the majority class.

CONCLUSION

Clustering is an important aspect of machine learning from the performance point of view. Clustering performs keen role in machine learning to formed a cluster as per the requirement. If the clustering not formed properly

then machine will not learned and it leads towards wrong output. The proposed system will overcome the limitation of existing clustering methodology.

References

[1] S.Guha, R. Rastogi and K.Shim, "Cure: an efficient clustering algorithm for large databases," vol. 26, no.1 , pp. 35-58, 2001.
[2] B. L. Milenova and M. M. Campos, "O-cluster: Scalable clustering of large high dimensional data sets," in IEEE International Conference on Data Mining (ICDM). IEEE, 2002, pp. 290–297.
[3] E. J. Otoo, A. Shoshani, and S.-w. Hwang, "Clustering high dimensional massive scientific datasets," Journal of Intelligent Information Systems, vol. 17, no. 2-3, pp. 147–168, 2001.
[4] Jian-Sheng ,Wei-Shi Zheng, "Euler Clustering on Large-scale Dataset", IEEE transaction on big data, vol no.14, 2017.
[5] Zheng Zhang, Li Liu, "Binary Multi-View Clustering", IEEE transaction on Pattern Analysis and Machine Intelligence,2018
[6]Yu-Jung Huang, "Machine-Learning Approach in detection and classification for defects in TSV-Based 3-D IC,IEEE TRANSACTIONS ON COMPONENTS, PACKAGING AND MANUFACTURING TECHNOLOGY, VOL. 8, NO. 4, APRIL 2018
[7] Dao Lam , "Unsupervised Feature Learning Classification With Radial Basis Function extreme Learning Machine Using Graphic Processors", IEEE TRANSACTIONS ON CYBERNETICS, VOL. 47, NO. 1, JANUARY 2017.
[8] Bartosz Krawczyk , "Learning from imbalanced data open challenges and future directions", April 2016.