International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

An Effective Unsupervised Machine Learning Technique and Research Challenges

Vishal S. Thakare[1], Jayshri S. Sonawane[2]

[1]R. C. Patel Institute of Technology, Shirpur

[2]R. C. Patel Institute of Technology, Shirpur

Abstract: ***Clustering is to categorize data items with similar structures or patterns into the same group for reducing data complexity and facilitating interpretation. It is common technique for statistical data, machine learning and computer science analysis. Clustering is a kind of unsupervised learning. 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: Cluster, K-means, Mean-Shift Clustering, Euclidean Distance, Tanimoto Distance, Agglomerative Hierarchical Clustering

1. Introduction

The 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.

A good clustering method will produce high quality clusters with high intra-class similarity low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. In general, various existing approaches used for clustering as to define a clustering quality function Qn, and then construct an algorithm which is able to minimize or maximize Qn. There exists a huge variety of clustering objective functions: the K-means objective function based on the distance of the data points to the cluster centers, graph cut based objective functions such as ratio cut or normalized cut, or various criteria based on some function of the within- and between-cluster similarities. Once a particular clustering quality function Qn has been selected, the objective of clustering is stated as a discrete optimization problem. Given a data set Dn ={D1, . . . , Dn} and a clustering quality function Qn, the ideal clustering algorithm should take into account all possible partitions of the data set and output the one that minimizes Qn.

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.

2. 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

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k[th] attributes (components) or data objects p and q.

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

2) Manhattan Distance is useful in some urban environments with orthogonal road networks.

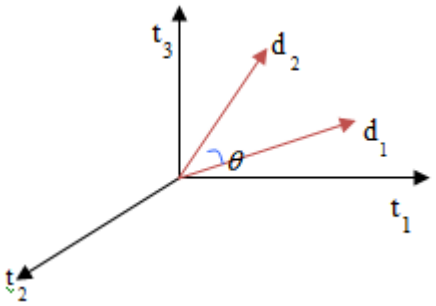
Movement is limited to city streets:

$$d_m = |x_1 - x_2| + |y_1 - y_2|$$

a reminder – the | symbols denote absolute value.

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.

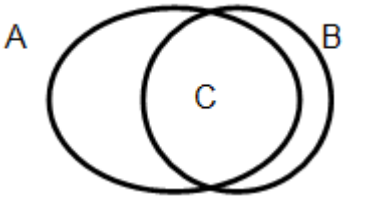


$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

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

4) Tanimoto Distance
Definition:

- Value range: [0,1]
- Tc is also known as Jaccard coefficient
- Tc is the most popular similarity coefficient



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

Role of Clustering in Machine Learning

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

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. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data.
2) **Mean-Shift Clustering** - It is a sliding-window-based algorithm that attempts to find dense areas of data points. This algorithm is based on centroid, That is the goal 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 Spatial clustering of Applications with Noise (DBSCAN) –** It is an advancement of earlier explained clustering. It begins with an arbitrary starting data point that has not been visited. The neighborhood of this point is extracted using a distance epsilon ε. If there are a sufficient number of points within this neighborhood then the clustering process starts and the current data point becomes the first point in the new cluster for this first point in the new cluster, the points within its ε distance neighborhood also become part of the same cluster. This procedure of making all points in the ε neighborhood belong to the same cluster is then repeated for all points in the ε neighborhood belong to the same cluster is then repeated for all of the new points that have been just added to the cluster group.
4) **Expectation-Maximization (EM) using Gaussian Mixture Models (GMM) -** GMM gives flexibility than K-means. With GMM 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 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) **Agglomerative Hierarchical Clustering –** 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. This hierarchy of cluster is represented as a tree (or dendrogram).

3. 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 dominates 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
Machine learning algorithms struggle with accuracy because of the unequal distribution for dependent variable.

- The accuracy of clustering must be increase.
- Machine Learning algorithms should identify that data set are balanced or imbalanced for clustering.
- 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.

4. Conclusion

Clustering is an important aspect of machine learning from the performance point of view. Clustering performs keen role in machine learning to form 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," Information Systems, 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 Wu, 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] Mapping, Learning, Visualization, Classification, and Understanding of fMRI Data in the NeuCube Evolving Spatiotemporal Data Machine of Spiking Neural Networks, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 28, NO. 4, APRIL 2017