CLASSIFY DATA THROUGH CLUSTERING ALGORITHM USING ARTIFICIAL INTELLIGENCE TECHNIQUE

[1]Gulafsa Parween, [2]Anshu Tiwari, [3]Manish Saxena

[1]Student, [2] Asst.Prof., [3] Asst.Prof.

[1]Computer Science Engineering , BIST, Bhopal. M.P.

Abstract: An optimal multi-objective clustering is one of the most popular, and, at the same time, curious supervised machine learning problems, that occurs in many fields of computer science such as data and knowledge mining, data compression, vector quantization, patterns detection and classification, Voronoi diagrams, recommender engines (RE), etc. The process of clustering analysis itself allows us to reveal various of trends and insights exhibited on the input dataset. The main idea of the k-means++ algorithm is that the entire process of centroids computation basically relies on the measure of optimal distance between those centroids being selected. It's highly recommended that the actual distance between centroids must be as largest as possible. Each cluster will contain the most similar data points, especially while performing the exclusive clustering operations, during which, each of the points is assigned to a particular cluster and cannot be included in more than one cluster simultaneously.

Index Terms **- cluster analysis, K-Means and k-means++ Clustering,**

I. INTRODUCTION

The cluster analysis (CA) process allows us to determine the similarities and differences between specific data, partitioning the data in such as a way that the similar data normally belongs to a specific group or cluster. For example, we can perform the clustering analysis of the data on a credit card customer to reveal what special offers should be given to a specific customer, based on the balance and loan amount criteria. In this case, all that we have to do is to partition all customers data into the number of clusters, and, then give the same offer to the similar customers. This is typically done by performing the multi-variate numerical data the multi-variate numerical data clustering analysis.

The main goal of performing the actual clustering is to arrange a set of data items having an associated numeric n-dimensional vector of features into the number of homogeneous groups, called - "clusters". In general, the entire clustering problem can be formulated as a certain objective similarity function minimization problem.

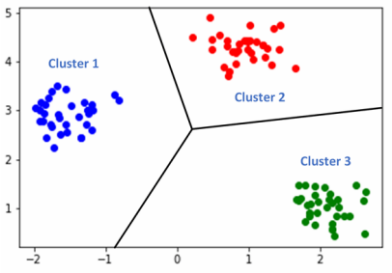

Figure. 1 Data Clustering

Among the known clustering algorithms, that are based on minimizing a similarity objective function, k-means algorithm is most widely used. Given a set of t data points in real n-dimensional space, and an integer k, the problem is to determine a set of k points in the Euclidean space, called centers, as well as to minimize the mean squared distance from each data point to its nearest centroid. This measure is often called the squared-error distortion and this type of clustering belongs to the variance-based clustering category.

Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-medians (or the multisource Weber problem) in which the objective is to minimize the sum of distances to the nearest center and the geometric k-center problem in which the objective is to minimize the maximum distance from every point to its closest center.

One of the most popular heuristics for solving the clustering problem based on using the k-means algorithm basically relies on a simple iterative method for finding a set of clusters. There is a number of variations of the k-means algorithm. Specifically, in this article, we will spotlight our discussion on the algorithm, proposed by Lloyd

II. METHODOLOGY

K-Means Clustering Algorithm

Initialization

The initialization phase of the k-means algorithm is rather intuitively simple. During the initialization phase, we normally select k – random points in the Euclidean space as a set of initial centroids prior to performing the actual clustering.

However, the following method has the number of disadvantages, which leads to the sub-optimal clustering problem occurrence. Specifically, the distance between randomly selected centroids, in the most cases, is not optimal (e.g. the distance between particular centroids is very small). The following normally has two main outcomes in either a poor quality of clustering or sufficient performance degrades.

K-Means++ Initialization

As a workaround to the sub-optimal clustering problem, we will formulate the k-means++ initialization algorithm that allows us to drastically improve the quality of clustering provided by the classical k-means procedure, as well as to increase its performance and lower the computational complexity. Normally we use the k-means++ alternative to achieve two main goals:

- Improve the convergence and quality of clustering provided by the classical k-means algorithm, over its known shortcomings and limitations;
- Provide a better performance of the NP-hard k-means clustering procedure, optimizing the iterations, and, thus, reducing its computational complexity to only

p=O(log(k))p=O(log(k))

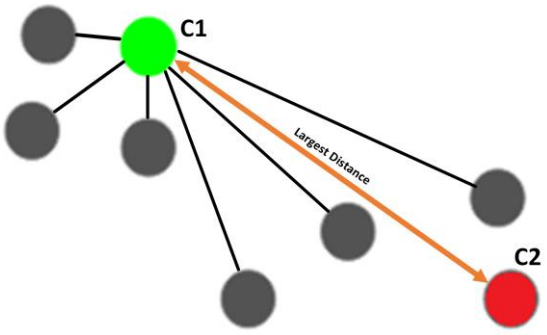

Figure2. Selecting the first and second centroids C1 and C2

III. EVOLUTION

There's the number of evaluation metrics, such as either the inertia magnitude or Dunn index, that allow us to measure a quality of clustering. By evaluating the inertia, we actually compute the sum of distances of all points to the centroid of each cluster, and, then add up each of these sums together. Actually, by computing the inertia we evaluate the sum of intra-cluster distances. Finally, we consider that as the lesser the inertia value as the quality of clustering is higher. The following formula illustrates the inertia metric computation:

$$I=\sum_{i=1}^{k}\sum_{j=1}^{c_i}|x_j-c_i|^2.$$

Dunn index is another metric by using which we can effectively evaluate the quality of clustering. The following metric basically represents a ratio between the minimum of inter-cluster distances and the maximum of intra-cluster distances. Unlike the inertia value, the Dunn index magnitude largely depends on the distance between clusters. In the most cases, to provide a high-quality clustering we want to maximize the Dunn index value. The formula below illustrates the Dunn index computation:

$$Dunnindex=\min\{inter-cluster\ distances\}\max\{intra-cluster\ distances\}$$

IV. PROPOSED WORK

After we've work about the initialization phase, let's now spend a moment and take a short glance at the k-means clustering procedure. According to the main idea of the k-means algorithm we need to arrange points with the smallest distance into multiple of groups called "clusters", and, since the newly built clusters have been produced, re-compute a centroid for each of these clusters, based on the process of finding the "center of the mass" magnitude, using the method discussed in the next paragraphs.

The k-means algorithm can be used for clustering the multidimensional data, in which each item is represented by a vector in n-dimensional Euclidean space, which components is a set of numeric features.

After brief introduction, let's formulate the classical k-means algorithm, which is very simple, Re-compute the centroids of each of the newly built clusters based on the "center-of-the-mass" magnitude, appending each new centroid being computed to the set of centroids

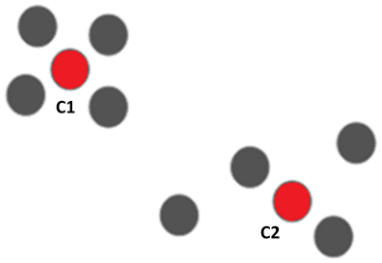

Figure.3 Recompute the new cluster centroide

1. Perform a check if the centroid of each already existing cluster has not changed and the points selected are not within the same clusters (e.g. we're not producing the similar clusters, containing duplicate points, which have already occurred in the previously built clusters);
2. If not, proceed with steps 1,2 and 3, otherwise terminate the k-means clustering process;

V. EXPERIMENTAL & RESULT ANALYSIS

For each value it performs a check if its a floating-point value. If so, it appends the following value to the list of features, building a multi-dimensional vector, which is then added to the list of items along with its label (e.g. the last comma-separated value in the current line). At the end of execution the following function returns a list of item tuples consisting of a vector of features and its label.

In fact, the k-means++ initialization procedure is the only possible algorithm addressing an optimal initial centroids selection. The following procedure is solely based on a technique of selecting centroids having an optimized distance between them, and, thus, ensuring that k-means algorithm itself will successfully converge at the very first steps of the

clustering process, significantly reducing the number of operations performed. Also, there's the number of k-means++ algorithm variations such as either distance- or probability-based. However, probability distribution-based k-means++ procedure, similar to the generic k-means initialization, sometimes tends to provide poor results of the initial centroids selection, due to the several randomization issues. As the result of the following function execution, we get a floating-point value of the Euclidean distance between two n-dimensional vectors of features.

Finally, in the chart below we will compare the results of the initial centroids selection by using the either k-means++ procedure or the generic k-means random initialization:

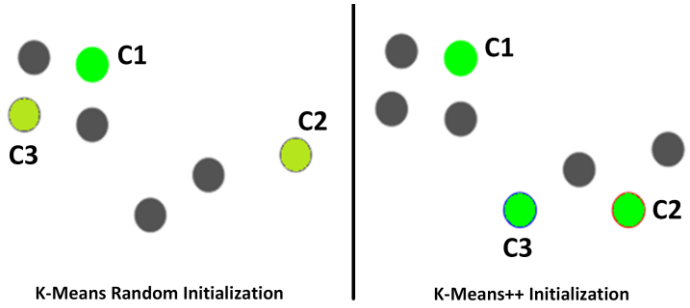

Figure 4. Result of centroids selection

VI. CONCLUSION

A comprehensive overview of machine learning algorithms for intelligent data analysis and applicationsThe following function listed above, accepts three arguments of either an input list of items, list of centroids indices or a list of points representing the new clusters centers. For each item it computes the Euclidean distance to each centroid or point in the list, performing a search to find a centroid for which this distance is the smallest. Then it assigns the following item to a cluster with a centroid having the smallest distance. The following function actually performs the entire dataset partitioning. Finally, the following function returns a list of tuples, each one is consisting of a centroid index and a list of item indices. Hence it's performing a check if the k-means procedure has converged and we're not producing the same clusters. If not, it proceeds with the next iteration of the k-means clustering process. the performance of sequential code fragments, performing the actual k-means clustering can be easily transformed into parallel using the various of existing high-performance c onsidering the correct algorithm, in return, can save time and efforts and assist in obtaining more accurate results. omputing (HPC) libraries and frameworks, providing even more performance of the clustering process itself. We also discussed several popular application areas based on machine learning techniques to highlight their applicability in various real-world issues. Finally, we have summarized the challenges faced and the potential research opportunities and future directions in the area. Therefore, the challenges that are identified create promising research opportunities in the field which must be addressed with effective solutions in various application areas.

REFERENCES

[1] Maciej Piernik1 · Tadeusz Morzy1, A study on using data clustering for feature extraction to improve the quality of classification,springer 2021

[2] MD REZAUL, Deep learning-based clustering approaches for bioinformatics, PP 393–415,2021

[3] Alapati YK, Sindhu K (2016) Combining clustering with classification a technique to improve classification accuracy bibtex. Int J Comput Sci Eng 5(6):336–338 .

[4] Guoqin Li,Youwei Sun,Yongping Chai. Congestion control method based on K-means clustering in VANET[J]. Computer Engineering and Design, 2020, v.41;No.398(02):42-46.

[5] S. Lu , "Clustering Method of Raw Meal Composition Based on PCA and k-means," 2018 37th Chinese Control Conference (CCC), Wuhan, 2018, pp. 9007-9010

[6] Agarwal J, Nagpal R, Sehgal R.Crime Analysis using K-Means Clustering[J]. International Journal of Computer Applications, 2018, 83(4):1-4

[7] Sujie Zhang. Algorithm research of optimal cluster number and initial cluster center[J].Application Research of Computers,2017(06)

[8] Adnan N, Nordin Shahrina Md, Rahman I, Noor A. The effects of knowledge transfer on farmers decision making toward sustainable agriculture practices. World J Sci Technol Sustain Dev. 2018.

[9] Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Algorithms. 2020;13(10):249.