ISSN: 0975-766X
CODEN: IJPTFI
Available Online through
Research Article
www.ijptonline.com

IMPROVED K-MEANS CLUSTERING ALGORITHM - WORKING WITH LABELED DATASETS

[1]Aniket Sen, [2]Prasun Jaiswal, [3]Vikram Das,[4]Prabadevi B
[1,2,3] MCA Student, [4]Assistant Professor
School of Information Technology and Engineering, VIT University, Vellore, India.
Email: prabadevi.b@vit.ac.in

Received on 25-10-2016 Accepted on 02-11-2016

Abstract:

Background and Objective: This paper proposes a better and improved K-Means clustering algorithm which deals with datasets comprising of a small amount of labeled data objects and the rest containing unlabeled data. **Materials and Methods:** The labels provide a fair idea about the number of clusters that the data might amount to. Since majority of the dataset contains unlabeled data, the proposed algorithm also finds the initial cluster points dynamically, by finding data objects that are most dissimilar to the existing cluster points. **Results and Conclusion:** Thus this version of the K-Means strives to function using less number of iterations due to the availability of labeled information on one hand and also results in a fairly optimal number of clusters present in the given dataset.

Keywords: Centroid, K-Means, data objects, labeled data

Introduction:

In today's ever-growing IT world, the amount of information generated is growing day by day. In order to utilize this vast information, useful patterns of meaningful data are extracted and stored for future references. This process of extracting useful data from a vast pool of information is known as Data Mining. This can be achieved using many predefined approaches like Association Rule Analysis, Clustering, Frequent Pattern Miningor Classification. Clustering is one of the many data mining techniques used to classify data which have some similar data attributes (into clusters). Clustering has many day to day applications in fields such as in the Health sector,in Banking, in Market analysis, and so on. Out of the many clustering algorithms, one of the easiest and widely used is the K-Means algorithm. It improves clusters by continuously moving data objects. It is simple, fast and highly efficient. However, this algorithm suffers from two main drawbacks. Firstly, selecting optimal centroid locations is the key to an effective clustering process. However, in the traditional K-Means, these are selected at random. This not only severely affect the algorithm's performance, but also, no two test runs on the same data set will yield the same result. Hence, results

cannot be matched or compared. Secondly, the number of initial clusters K needs to be given in advance as the input. The efficiency of the resulting clusters vary greatly on this value. If this value is too small, then clusters which should have been separated would be otherwise combined. And vice versa. Although this may be possible with smaller simulated datasets, assumption of any real time value of K for large real world datasets is nearly impossible for most people. However, in-order to overcome this problem, we can introduce the use of labels. If we add labels to a small part of the data set, it number of types of labels can be used to estimate an optimal value of K (viz. the number of clusters). Moreover, the position of these labels and be used to get a rough estimate regarding the initial centroid locations. However, it needs to be noted that if a large volume of the data contains labels, it can cause a lot of difficulties in the determining the initial number of clusters and initial centroid locations. So it is optimally expected that the actual data set be comprised of fewer labeled data objects rather than unlabeled ones. Hence, due to these disadvantages of the traditional K Means algorithm, this paper proposes a new modified K-Means algorithm where both the value of K and the initial centroid positions will be determined by the algorithm itself thus vastly improving the algorithm's efficiency.

Materials and Methods:

A. TheTraditional K-Means Algorithm: The traditional K-mean algorithm consists of the following steps:

1. Take an input value K (no. of clusters) and randomly select K clustering centers.
2. Assign a data object from the data set to the nearest cluster center forming an initial set clusters around the clustering centers (K-seeds)
3. Consider the data objects within each cluster and calculate the mean value. Then make it the new clustering center of the cluster.
4. Repeat steps 2 & 3 of add and update operations till two successive iterations amount to center values having insignificant difference between them.

Figure 1.1 below gives a graphical representation of the steps performed in the Traditional K-Means algorithm.

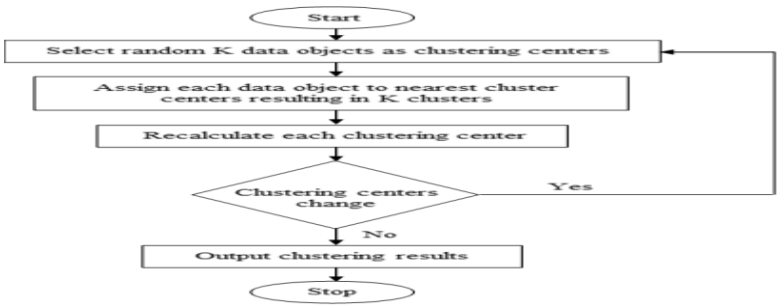

Figure 1.1Visual Representation of Traditonal K-Means.

B. Drawbacks

Few of the drawbacks of this algorithm are,

- The value of K (i.e. the number of clusters) has to be provided by the user, which mostly leads to inefficient outputs.
- The initial centroid locations are chosen at random leading to different results even on the same data set.
- Highly effected by surrounding noise.
- Does not work well with non-globular clusters.

C. Literature Survey

One of the most popularly used algorithms in Data Mining is the K-Means. However, due to its drawbacks, extensive research is being carried out to overcome its two major limitations- firstly, to efficiently select the initial centroid positions and secondly, to remove the need to enter the value of K as input.

Few of the studies that work in improving the K-Means algorithm and removing its various drawbacks have been discussed below:

Hanmin et al. [1] have proposed a way to overcome both major drawbacks of the K-Means with the use of labeled data objects. It is assumed that the given data set will contain a small minority of labeled data objects. The number of distinct labels will give K while the approximate positions of similar labels will be used to compute the initial centroid locations. It is observed that this method is preferable as this external factor improves the efficiency of the algorithm and speeds up the entire process. However, if the number of labeled data objects assumes majority of the data set, then the clustering becomes disintegrating.

Chadha et al. [2] have presented a way to use the K-Means algorithm without having to take the value of K as input. The algorithm starts by taking two clusters having data objects with the farthest distance between them. This dynamic estimation of the number of clusters (viz. K) overcomes a major drawback of the traditional K-Means and greatly improves the algorithm's efficiency. However, in this system, only numerical data objects can be clustered.

Gulati et al. [3] have given an overview on the various types of clustering algorithms and the comparison between them.The advantages, disadvantages, time complexities, space complexities and other such factors have been compared and the results of which have been shown.

Chaturbhuj et al. [4] have proposed a system where PSO (Particle Swarm Optimization) is used to find the optimum centroid positions, Gas algorithm is used to prevent the K-Means algorithm from converging to the local optimum and

Hadoop is used to allow parallel processing of large sets of data. This system overcomes the drawbacks of having to input the number of initial clusters beforehand and also prevents K-Means from converging to the local minima from the starting position.

Mehrotra et al. [5] have proposed a system to improve the speed of search results in websites. The system uses traditional K-Means with the Genetic algorithm to improve speed and accuracy of the clustering process. Moreover, labeled data objects are also used to increase clustering efficiency. Due to this system, the data is not confined to any particular input parameter and even outlying data objects are taken care of.

Swamy et al. [6] have proposed an improved K-Means approach using parallel processing to greatly reduce execution time. The proposed system firstly uses an initialization method to find the initial centroid locations. Moreover, parallel processing is also included to greatly reduce the execution time. This way, not only does the time of execution of the K-Means algorithm drastically reduce, but the use of the initialization method greatly increases the accuracy of the clustering.

Proposed System:

The idea behind the proposed system is enkindled by observing the concepts of the above mentioned journals and literary works. We observe two main concepts:

1. Use of labeled data sets makes the clustering procedure somewhat easy.
2. Dynamic resolution of clustering centers.

So we propose an algorithm which makes the use of data sets having a small amount of labeled and a considerable amount of unlabeled data objects.

Our algorithm follows these steps:

Let $X = \{x_1, x_2, \ldots\ldots, x_n\}$ be the data set containing N data objects and, $XL = \{x_{l1}, x_{l2}, \ldots\ldots, x_{lL}\}$ be the set of labeled data objects with L distinct labels.

1) We start a search by finding sets of two data objects farthest from each other by the dissimilarity formula

$$d(x_i, x_j) = \sqrt{((x_i - x_i)^2 + (x_i - x_i)^2 + \ldots + (x_i - x_i)^2)}$$

$$where\ x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\},$$

$$x_j = \{x_{j1}, x_{j2}, \ldots, x_{jm}\}\ and\ m\ is\ the$$

$$number\ of\ dimensions\ of\ a\ data\ object\ x_i\ or\ x_j$$

2) If the pair of data objects found belong the same label a cluster is formed by adding the two and the search is continued to find the next pair of data objects having most dissimilarity.

Otherwise if they are of different labels or xi,xj ∉ XL two clusters are formed with each data object being the clustering center.

3) Data objects are then assigned to the clusters either by their label value if the cluster for that label is already formed or by their least dissimilarity value to one of the existing clusters.

4) Cluster means are calculated and updated.

5) Again a search is made for a data object which is farthest from the existing clusters.

If the label of the object already has an existing cluster it is assigned to it,

Otherwise it is set as a new clustering center.

Steps 3, 4&5 are repeated until convergence.

Algorithm:

```
Max= -∞;
    fori = 1 to N
    {
        for j = 1 to N
        {
            D = d(x_i, x_j);
            if ( D > Max )
                Max = D;
                p1 = i, p2 = j;
        }
    }
    Cluster ( p1, p2 );
    do
    {
        fori = 1 to |C|
        {
            for j = 1 to N
            {
                D = d(x_i, x_j);
                if ( D > Max )
                    Max = D;
                    p1 = i, p2 = j;
            }
        }
        Cluster ( p1, p2 );
    }
    until converges;
}

Cluster ( i, j )
{
    if x_i ∈ XL AND x_j ∈ XL
        if x_i and x_j have same label '1'
            C = C + { x_li };
        else
            C = C + { x_li, x_li };
    else
        C = C + { x_i, x_j };

    fori = 1 to N
    {
        if x_i has a label which ∈ C
            Assign x_i to that cluster;
        else
            Assign x_i to closest cluster in C;

        Update clusters;
 }  }
```

The proposed algorithm targets at giving results faster than the previously known algorithms in this **domain as it can attain convergence earlier**. Moreover the dependency on the initial K value to be provided, random selection of initial clusters are overcome in this approach.

Conclusion:

In this paper, we have proposed an improved K-Means algorithm which works in removing both major limitations in the traditional algorithm, viz. taking the value of K as input and randomly selecting the initial centroid positions. According to the data provided by the fewer labeled data objects, the algorithm finds optimum initial centroid positions. In selecting the centroid locations from the unlabeled data, the data object that is farthest from any confirmed centroid location is selected to be the next centroid location. The result from the cluster numbers are then analyzed to confirm the final number of clusters and the initial centroid locations. Then the clusters are shown as output. Hence it can be seen that the efficiency of this proposed algorithm is much better as compared to the traditional K-Means.

References:

1. Hanmin, Ye, Lv Hao, and Sun Qianting. "An improved semi-supervised K-means clustering algorithm." Information Technology, Networking, Electronic and Automation Control Conference, IEEE. IEEE, 2016.
2. Chadha, Anupama, and Suresh Kumar. "An improved K-means clustering algorithm: a step forward for removal of dependency on K." Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on. IEEE, 2014.
3. Gulati, Hina, and P. K. Singh. "Clustering techniques in data mining: A comparison." Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on. IEEE, 2015.
4. Chaturbhuj, Kaustubh S., and Gauri Chaudhary. "Parallel clustering of large data set on Hadoop using data mining techniques." Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on. IEEE, 2016.
5. Mehrotra, Shashi, and Shruti Kohli. "Comparative analysis of K-Means with other clustering algorithms to improve search result." Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on. IEEE, 2015.
6. Swamy, Prateek, M. M. Raghuwanshi, and Ashish Gholghate. "An Improved Approach for k-Means Using Parallel Processing." Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on. IEEE, 2015.