Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Efficient Framework for Clustering Data Based on Dbscan and K-Means Algorithms

Romana Riyaz, Mohd Arif Wani

Dept. of Computer Sciences
University of Kashmir
Srinagar, j&k, India

Abstarct— In this paper we have presented a proposed a three step framework,starting with finding the initial clusters and then initalizing initial cluster centers and finially partitioning data into most optimal clusters.we have employed some the most effiecient algorithms like Dbscan and K-Means(XK-Means) and we have tested our approach on iris data set.

Keywords— exploratory vector;centroids;xk-means;euclidean distance;density rechable;Dbscan

I. INTRODUCTION

Clustering is one of the widely used knowledge discovery techniques to reveal structures in a data set that can be extremely useful to the analyst [1]. As clustering does not make any statistical assumptions to data, it is referred to as unsupervised learning algorithm. The goal of clustering is to group data into clusters such that the similarities among data members within the same cluster are maximal while similarities among data members from different clusters are minimal. Clustering algorithms can be broadly classified into hierarchical and non-hierarchical clustering algorithms.k-Means algorithm is the most well known and fast method in non-hierarchical clustering algorithms. Because of the simplicity of K-means algorithm, it is used in various fields. K-Means is a partitioning clustering method that separates data into k mutually exclusives groups. Through such the iterative partitioning, k-means algorithm minimizes the sum of distance from each data to its clusters. K-Means algorithm is very popular because of its ability to cluster a kind of huge data, and also outliers, quickly and efficiently. However, K-Means algorithm is very sensitive to the designated initial starting points as cluster centers. K-Means does not guarantee unique clustering because we get different results with randomly chosen initial clusters. The final cluster centroid may not be optimal ones as algorithm can converge into local optimal solutions. Therefore it is very important for K-Means algorithm to have good initial clusters.

II. LITERATURE SURVEY

Several methods have been proposed to solve the cluster initialization for K-Means algorithm. Some of the contributions have been discussed as under:

Reference [1] proposed the refinement algorithm that builds a set of small random sub-samples of the data, then clusters data in each sub-samples by K-Means. All centroids of all subsamples are the clustered together by K-Means using the k-centroids of each sub-sample as initial centers. The centers of the final clusters that give minimum clustering error are to be used as initial clusters for clustering the original set of data using K-Means algorithm.Reference [2] proposed two different methods. The firstone, which is the default option in the Quick Cluster procedure of IBM SPSS Statistics [3], takes the first K points in X as the centers. An obvious drawback of this method is its sensitivity to data ordering. The second method chooses the centers randomly from the data points. The rationale behind this method is that random selection is likely to pick points from dense regions,i.e. points that are good candidates to be centers. However, there is no mechanism to avoid choosing outliers or points that are too close to each other. Multiple runs of this method are the standard way of initializing k-means [1]. It should be noted that this second method is often mistakenly attributed to [10].Variance-based method [14] first sorts the points on the attribute with the greatest variance and then partitions them into K groups along the same dimension. The centers are then chosen to be the points that correspond to the medians of these groups. Note that this method disregards all attributes but one and therefore is likely to be effective only for data sets in which the variability is mostly on one dimension. The ROBIN (RobustInitialization) method [15] uses a local outlier factor (LOF) [16] to avoid selecting outlier points as centers. In iteration I (i ε {1, 2,...,K}), the method first sorts the data points in decreasing order of their minimum-distance to the previously selected centers. It then traverses the points in sorted order and selects the first point that has an LOF value close to 1 as the ith center. The computational cost of this method is dominated by the complexity of sorting, which is O(N log N).Reference[11]proposed the global K-Means algorithm which is incremental approach to clustering which dynamically adds one cluster center at a time through a deterministic global search procedure consisting of N(with N being the size of the dataset).Reference [3] proposed cluster center initialization algorithm (CCIA) to solve cluster initialization problem.CCIA is based on two observations, which some patterns are very similar to each other. It initiates with calculating mean and standard deviation for data attributes, and then separate the data with normal curve into certain partition.CCIA uses K-Means and density based multi scale data condensation to observe the similarity of data patterns before finding out the final initial clusters. The experiment results of CCIA performed the effectiveness and robustness this method to solve the several clustering problems. The directed-search binary-splitting method[17] is an improvement over the binary-splitting method in that it determines the value of using PCA. However, it has even higher computational requirements due to the calculation of the principal eigenvector in each iteration. Reference [4] uses a divisive hierarchical approach based on PCA (Principal Component Analysis). Starting from an initial cluster that contains the entire data set, the method iteratively selects the cluster with the greatest SSE and divides it into two sub clusters using a hyper plane that passes through the cluster centroid and is orthogonal to the direction of the principal eigenvector of the covariance matrix. This procedure is repeated until K clusters are obtained. The centers are then given by the centroids of these clusters. The Var-Part method is an approximation to PCA-Part, where the covariance matrix of the cluster to be split is assumed to be diagonal. In this case, the direction of the splitting hyper plane is orthogonal to the coordinate axis with the greatest variance. Reference [5] uses a two-phase pyramidal approach. The attributes of each point are first encoded as integers using 2 Q -level quantization, where Q is a resolution parameter. These integer points are considered to be at level 0 of the pyramid. In the bottom-up phase, starting from level 0, neighboring data points at level k (k ε 2 {0,1,. . .}) are averaged to obtain weighted points at level k + 1 until at least 20 K points are obtained. Data points at the highest level are refined using k-means initialized with the K points with the largest weights. In the top-down phase, starting from the highest level, centers at level k + 1 are projected onto level k and then used to initialize the k-th level clustering. The top-down phase terminates when level 0 is reached. The centers at this level are then inverse quantized to obtain thefinal centers. The performance of this method degrades with increasing dimensionality. Reference [6] first calculates K Independent Components (ICs) of X and then chooses the i-th (iε 2 {1, 2, . . . , K}) center as the point that has the least cosine distance from the i-th IC.

III. Proposed approach

In proposed approach we have used density based clustering method. The density based clustering algorithms are designed to discover clusters of arbitrary shape in databases with noise. Dbscan is typical density based clustering algorithm. It does clustering through growing high density area, and it can find any shape of clustering. In this paper we have proposed to use Dbscan algorithm as a starting step for K-means and XK-Means algorithm in finding initial number of clusters K. After using Dbscan algorithms for finding K, this k is then fed to XK-Means algorithms as initial clusters which then find the initial cluster centers for these clusters, then partitioning of data into most optimal clusters by using steps of simple K-Means algorithm. We have then evaluated the performance of proposed approach on iris data set.

Our proposed approach for automatically obtaining clusters from a given dataset is based on three steps as explained below:

A. Obtaining the number of clusters.
B. Initialization of cluster center seeds.
C. Partitioning data into respective clusters.

A. Obtaining the number of clusters

Density-Based Spatial Clustering and Application with Noise (Dbscan) was a clustering algorithm based on density. The Dbscan algorithm can identify clusters in large spatial data sets by looking at the local density of database elements, using only one input parameter. Furthermore, the user gets a suggestion on which parameter value that would be suitable. Therefore, minimal knowledge of the domain is required. The Dbscan can also determine what information should be classified as noise or outliers. In spite of this, its working process is quick and scales very well with the size of the database almost linearly. By using the density distribution of nodes in the database, Dbscan can categorize these nodes into separate clusters that define the different classes. Dbscan can find clusters of arbitrary shape its working process is quick and scales very well with the size of the database – almost linearly. By using the density distribution of nodes in the database, Dbscan can categorize these nodes into separate clusters that define the different classes. Dbscan can find clusters of arbitrary shape. Dbscan algorithms require two parameters: epsilon (eps) and minimum points (minpts).it starts with an arbitrary starting point that has not been visited. It then finds all the neighbor points within distance eps of the starting point. If the number of neighbors is greater than or equal to minpts, a cluster is formed. The starting point and its neighbors are added to this cluster and the starting point is marked as visited. The algorithm then repeats the evaluation process for all the neighbors' recursively. If the number of neighbors is less than minpts, the point is marked as noise. If cluster is fully expanded (all points within reach are visited) then the algorithm proceeds to iterate through the remaining unvisited points in the dataset.

```
Steps of Dbscan Algorithm
          For each o ε D
         If o is not yet classified then
         If o is a core object then
Collect all object density reachable from
O and assign them to a new cluster
Else
Assign o to noise
```

B. Initializing Cluster Centers

This novel algorithm is known as 'exploratory k-means/XKmeans algorithm is given by [3].The objective of this method is to reduce the computational load and enhance precision. This method has targeted the weakness of k-Means in its ineffectiveness in exploring better search paths that may lead to optimal solution. So, this method suggests to integrate an exploratory component into the k-means algorithm can be attributed to the general lack of exploration mechanism which results in search paths permanently trapped in local minimum if the initial setting is not in a course leading to the global solution.XK-means algorithm overcomes this flaw by incorporating exploratory mechanism. This method is inspired by the concept of particle swarm optimization (PSO), which states the importance of integrating exploration and exploitation in a search process.XK-means algorithm preserves the basic framework of k-means and on top of it an exploratory vector is added to each centroid before k-means iteration. This algorithm has been evaluated with real gene expression data sets. Its performance is compared against the k-means and Pk-means methods. It is noted that amalgamation of exploratory capability of the search process, while preserving favorable exploitation property of the k-means algorithm. This proposed method first samples a potential solution in the problem space in random manner while evaluating the correctness of the solution (in terms of MSE) this algorithm will swiftly explore other potential solutions in the neighboring area. This leads to better grouping of the clusters and improving speed of the grouping.
Steps:

1) Determine the number of centroids, k, and initialize the set of centroids randomly.

2) Initialize inertia magnitudes

$$b_i \ and \ a_i.$$

The terms $a_i \ and \ b_i$ are non zero and positive and are referred to as lower and upper inertia magnitudes in dimension i. They are related as:

$$b_i^* = \alpha b_i \ \ Where \qquad \alpha = [0,1] \ \ a_i = \beta b_i \ \ where \ \beta = [0,1]$$

3) Update the inertia magnitudes according to equations:

$$a_i = \beta b_i \ \ where \ \beta = [0,1]$$
$$b_i^* = \alpha b_i \ Where \ \alpha = [0,1]$$
$$b_i^* \ is \ the \ next \ value \ of \ the \ b_i$$

4) Add exploratory vector to explore each centroid using equation:

$$z_j^* = z_j + v_j$$

v

$\mathbf{v_j}$ is a D- dimensional exploratory vector at current iteration and Quantity of its dimensions is defined as:

$$v_{j,i}| = rand(a_i, b_i) * \ randsign(i)$$

5) Assign each data point to the cluster with the closest (Euclidean) distance between them.

6) Compute new centroid according to the classic K-Mean.

7) Evaluate MSE (mean square error).

8) If stopping criteria has reached then stop else update the inertia magnitudes again i.e. go to step 3.

C. Partitioning data into respective clusters

K Means is a classic clustering method that has been widely adopted in numerous engineering applications. Its popularity is supported by the simplicity of the algorithm, as well as the fast convergence rate especially in high-dimensional problems. K-Means is a method to partition a given set of N data points into K groups (called clusters) in D-dimensional Euclidean space. The partitioning in the space is based on certain similarity metrics which is usually Euclidean distance. The objective of the partitioning process is to minimize the error, generally expressed as the mean squared error (MSE), in approximating each data point with its nearest centroid.

Input: k: the number of clusters
D: a data set containing n objects.
Output: A set of k clusters.

Method:

1) Arbitrarily choose k objects from D as initial cluster centers:
2) Repeat
3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the object in the cluster;
4) Update the cluster means ,i.e., calculate the mean value of the objects for each cluster;

5) Until no change.

IV. PERFORMANCE EVALUATION

TABLE I
FINDING NO.OF CLUSTERS BY DBSCAN

Eps	Minpts	No. Of Cluster
1.5	3	3
2	3	5
2.5	3	7
3	3	10
3.5	3	10

TABLE II
K-MEANS ALGORITHM FOR K=3

Cluster No.	Iris Setosa	Iris Versicolor	Iris Virginica
1	45	0	8
2	3	35	10
3	0	8	31

TABLE III
XK-MEANS ALGORITHM FOR K=3

Cluster No.	Iris Setosa	Iris Versicolor	Iris Virginica
1	47	0	0
2	2	38	5
3	1	2	39

V. CONCLUSION

K-Means greatly depends on initial cluster centers. Selection of appropriate value of K and cluster center objects is a challenging issue. Lot of research has been already done on it and much more is yet to be done. The proposed method can help in choosing better values of K which in turn results in better clustering. We have employed improved K-means i.e. XK-Means algorithm for testing proposed approach. Results show that there is considerable improvement in clustering. This however can be further improved by using more efficient cluster initialization techniques.

REFERENCES

[1] Bradley, P. S., & Fayyad, U.” *Refining initial points for k-means clustering”*, In:Proc. of the 15th int. conf. on machine learning (pp. 91–99),1998.

[2] MacQueen, J.(1967).”*Some methods for classification and analysisof multivariate observations”*,In: Proc. Of the 5th Berkeley symposium on mathematical statistics and probability (pp. 281–297).

[3] S.S. Khan, A. Ahmad, *“Cluster center initialization algorithm for K -Means clustering”*, Patter Recognition Letters 25 1293–1302,2004.

[4] Su, T., & Dy, J. G.*” In search of deterministic methods for initializing k-means”*, (2007).

[5] Lu, J. F., Tang, J. B., Tang, Z. M., & Yang, J. Y. “*Hierarchical initialization approach for k-means clustering”*, Pattern Recognition Letters, 29(6), 787–795,2008.

[6] Onoda, T., Sakai, M., & Yamada, S. *“Careful seeding method based on independent components analysis for k-means clustering”*. Journal of Emerging Technologies in Web Intelligence, 4(1), 51–59,2012.

[7] K. Mumtaz et al. / (IJCSE) International Journal on Computer Science and Engineering, *“A Novel Density based improved k-means Clustering Algorithm – Dbkmeans”*,ISSN : 0975-3397 213 Vol. 02, No. 02, 2010, 213-218.

[8] Kalpana D. Joshi et al,”*Modified K-Means for Better Initial Cluster Centres”*,International Journal of Computer Science and Mobile computing Vol.2 Issue. 7, July- 2013, pg. 219-223.

[9] Y.K. Lam, P.W.M. Tsang.”*eXploratory K-Means: A new simple and Efficient algorithm for gene clustering”* / Applied Soft Computing 12(2012) 1149–1157

[10] Forgy,E.(1965*).”Clusteranalysis of multivariate data: Efficiency vs. interpretability of classification”*.Biometric 768.

[11] Likas, A., Vlassis, N., & Verbeek, J.”*The global k-means clustering algorithm”*, 2003.

[12] Norušis, M. J. (2011). IBM SPSS statistics 19 statistical procedures companion .Addison Wesley.

[13] Anderberg, M. R. (1973*).” Cluster analysis for applications”*. Academic Press

[14] Al-Daoud, M. (2005).”*A new algorithm for cluster initialization”*, In: Proc. Of the 2nd World Enformatika conf. (pp. 74–76).

[15] Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M.”*Robust partitional clustering by outlier and density insensitive seeding”*. Pattern Recognition Letters.30(11),994-1002.

[16] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000).” *LOF*: *Identifying density based local outliers”*. ACM SIGMOD Record, 29 (2), 93–104.

[17] Huang, C. M., & Harris, R.W. (1993).”*A comparison of several vector quantization codebook generation approaches”*. IEEE Transactions on Image Processing, 2 (1), 108–112.