Comparative Analysis of Fuzzy C Means and Fuzzy C Means++

Manika garg
Assistant Professor, Department of Computer Science, IITM Janakpuri, New Delhi, India.

Abstract

Cluster analysis is one of the most useful means for identifying relations and patterns in the area of data mining. It can be defined as partitioning of large volumes of data into various clusters that share some property or attribute. The most common clustering algorithm is k means. The more improved version of k means that incorporates fuzzy feature is fuzzy c means. To overcome some of the limitations of fuzzy c means, fuzzy c means ++ was introduced which was based on effective seeding mechanism of k means++ algorithm. The latter algorithm showed remarkable results but with some more limitations of its own. In this paper, we discuss both the methods and their algorithms in detail. We discuss the advantages and limitations of each in various scenarios.

Keywords: Fuzzy, C, Data Mining, K means++.

1 Introduction

Data mining is the analysis of datasets that are observational, aiming at finding out unsuspected relationships among datasets and summarizing the data in such a noble fashion that are both understandable and useful to the data users . Data mining involves use of various techniques like clustering, classifications, visualizations etc. Data clustering is a technique for data analysis used to partition data into various sets such that the members of same set share some trait.

The k-means method is a widely used clustering technique that seeks to minimize the average squared distance between points in the same

Cluster. K means is a hard clustering method.In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

Fuzzy c-means ++ is a type of fuzzy clustering method that claims to solve some of the limitations of fuzzy c means algorithm. It results in faster convergence times and better quality results but does involves some limitations of its own.

In this research paper, Fuzzy C-Means and fuzzy C-Means++ clustering algorithms are analysed based on their clustering efficiency.

2 Background

K-Means or Hard C-Means clustering is basically a partitioning method applied to analyse data and treats observations of the data as objects based on locations and distance between various input data points. Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters.

Each cluster is characterized by its centre point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial distances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. But use of several replicates with random starting point leads to a solution i.e. a global solution. In a dataset, a desired number of clusters K and a set of k initial starting points, the K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose co-ordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters.

Algorithmic steps for K-Means clustering [12]

1) Set K – To choose a number of desired clusters, K.

2) Initialization – To choose k starting points which are used as initial estimates of the cluster centroids. They are taken as the initial starting values.

3) Classification – To examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.

4) Centroid calculation – When each point in the data set is assigned to a cluster, it is needed to recalculate the new k centroids.

5) Convergence criteria – The steps of (iii) and (iv) require to be repeated until no point changes its cluster assignment or until the centroids no longer move.

A specific way of choosing centres for the k-means algorithm was proposed by David Arthur and was called K means ++.The K-means++ method (Arthur et al., 2007), the basis of this work, initializes the cluster centres of the K-means algorithm by selecting points in the dataset that are further away from each other in a probabilistic manner. This method both avoids the problems of the standard method and improves speed of convergence, being theoretically guaranteed to be O (log k), and hence competitive with the optimal solution. In particular, let D(x)denote the shortest distance from a data point to the closest centre we have already chosen.

Algorithmic steps for K-Means++ clustering

Take one centre c1, chosen uniformly at random X.

Take a new centre ci, choosing x ∈ X with probability ⟦D(x)⟧^2/(∑_(x∈X)▒⟦D(x)⟧^2).

Repeat Step 2, until we have taken k centres altogether.

4. Proceed as with the standard k-means algorithm.

We call the weighting used in Step 2 simply “D2 weighting”.

We will focus on more general versions of above two algorithms which are fuzzy c-means and fuzzy c means++ discussed in next sections.

hanging indents so that if the heading is too long to fit on a single line, the indent will be maintained. Headings should use initial capital letters, as shown.

The body text should be in 10 point Times New Roman, with the paragraphs justified on both left and right margins, and with 6 points spacing before each paragraph.

3 Fuzzy C-Means Clustering

Bezdek introduced Fuzzy C-Means clustering method in 1981, extend from Hard C-Mean clustering method. FCM is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature analysis, clustering and classifier design. FCM is widely applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition.

With the development of the fuzzy theory, the FCM clustering algorithm which is actually based on Ruspini Fuzzy clustering theory was proposed in 1980’s. This algorithm is used for analysis based on distance between various input data points. The clusters are formed according to the distance between data points and the cluster centres are formed for each cluster.

In fact, FCM is a data clustering technique in which a data set is grouped into n clusters with every data point in the dataset related to every cluster and it will have a high degree of belonging (connection) to that cluster and another data point that lies far away from the centre of a cluster which will have a low degree of belonging to that cluster.

Algorithmic steps for Fuzzy C-Means clustering

Algorithm 1: Fuzzy C-means (FCM)

Given X = {x_i } where i= 1 to N and k, return U and R

1: procedure FCM (Data set X, Clusters k)

2: U^0is randomly initialized

3: repeat

4:⟦ r⟧_j=μ_ij^m ∑_(i=1)^n▒⟦μ_ij^m x_i ⟧ where j=1,2…..k

5:μ_(ij=1/(∑_(k=1)^c▒(⟦|x⟧_i-r_j |)/(⟦|x⟧_i-r_k |)⟧^(2/(m-1))))

6: until |U^(k+1)-U^k |<e

7: end procedure

FCM iteratively moves the cluster centres to the right location within a dataset. To be specific introducing the fuzzy logic in K-Means clustering algorithm is the Fuzzy C-Means algorithm in general. In fact, FCM clustering techniques are based on fuzzy behaviour and they provide a technique which is natural for producing a clustering where membership weights have a natural interpretation but not probabilistic at all. This algorithm is basically similar in structure to K-Means algorithm and it also behaves in a similar fashion.

But, this method has some limitations. It has higher computational cost as compared to K means algorithm. Computational complexity is quadratic in the number of clusters O(NC^2P) where N is the number of data points, C is the number of clusters and P is the dimension of data points. Most

importantly, Random selection of initial centroids gives different results every time. There is no defined method about how to select those initial centroids. Number of iterations hugely depends on the selection of initial centroids.

4 Fuzzy C-Means ++ Clustering

Initializing representatives by selecting random points from the input dataset results in a sub-optimal starting strategy for the standard version of the algorithm. The idea behind the proposed Fuzzy C-means++ scheme (Algorithm 2) is to choose points that are spread out in the data set as representatives and update the membership matrix accordingly before commencing Steps 2–4 magenta) and thus requiring a much lower number of steps to converge. Note that the more well behaved (non-overlapping) the data set is, the more predictable (less volatile) is the initialization outcome of the Fuzzy C-means++ scheme.

Algorithm 2: Fuzzy C-means (FCM++) Initialization.

Given a set of N data points X = {x_i } where i= 1 to N and k representing the clusters number and p representing the spreading factor, return a set R of initial centres.

8: procedure FCM++ (Dataset X, Clusters k)

9 : R :=R∪ random point from dataset

10: while size of R < k do:

11: sample x ∈ X with probabilityd^p/(∑_(x∈X)▒d^p) .

12: R := R ∪ x

13: end while

14: end procedure

The proposed method, shown above, picks the first representative at random from the dataset and adds it to the set of representatives R. This point r1 determines a probability distribution for each other point ri in the dataset: the bigger the distance from r1 to ri the higher the chance of ri being picked as the next representative. Using the distanced^p where d^p (x,R) denotes the distance (raised to power p) from a point x ∈ X to its closest representative in R) allows for controlling the spreading factor of the algorithm through the parameter p. A small value for p will pick points closer together while a larger p will pick points that are further away as initial points. In the extreme case of p = 0, each point will have a random chance of being picked next and the method is similar to random initialization. Conversely, if we choose a p that is very large, we would likely pick outliers as starting points. A trade-off has to be made and the next section considers ways to choose p depending on the data. The set R is updated in this way until k representatives are chosen. This method has many advantages over earlier mentioned methods. Fuzzy C-means++ achieves lower cost functions values at convergence. The XieBeni index indicating cluster quality is lower (better) for Fuzzy C-means++ across the whole range of ks. It gives faster convergence time and higher quality solutions.

However, the choice of p parameter representing the spreading factor is not defined properly. There is no proper value of p which agrees with all the data sets. Sometimes, the algorithm generates bad clusters, because the algorithm still depends on choosing the initial cluster c1. The ambiguity in choosing the initial cluster makes the algorithm stochastic. This means that the results produced were considerably different across several analysis run using the same initial conditions. Also, it Scales poorly for large datasets. It was observed that the difference in the results grows as the datasets contain more data points and has higher feature dimensionality. It is inherently serial algorithm that is time consuming for large data sets.

5 Conclusion

The paper compares fuzzy c-means and fuzzy c-means++ clustering algorithms. Fuzzy c-means++ clustering produces approximate results as fuzzy c means but with lesser number of iterations. However, there is still ambiguity in the algorithm in selecting the value of spreading factor. Also, there is no defined way of selecting the first random point. But overall, we can say that fuzzy c-means++ performs much better than fuzzy c-means with slightly more calculations.

References

[1] Adrian Stetco⇑, Xiao-Jun Zeng, John Keane.Fuzzy C-means++: Fuzzy C-means with effective seeding initialization

[2] Arthur, D., Arthur, D., Vassilvitskii, S., &Vassilvitskii, S..k-Means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, Vol. 8, pp. 1027–1035, 2007.

[3] Asuncion, A., & Newman, D. J.,.UCI machine learning repository.University of California Irvine School of Information, 2007.

[4] Celebi, M. E., Kingravi, H. A., & Vela, P. A.,.A comparative study of efficientinitialization methods for the k-means clustering algorithm. Expert Systems with Applications, 40, 200–210. http://dx.doi.org/10.1016/j.eswa.2012.07.021, 2013.

[5] Doring, C., Lesot, M.-J., & Kruse, R.,.Data analysis with fuzzy clustering methods. Computational Statistics & Data Analysis, 51, 192–214, 2006

[6] Fukuyama, Y., &Sugeno, M., A new method of choosing the number of clusters for the fuzzy c-means method, Proc. 5th Fuzzy Syst. Symp., pp. 247–250, (1989)..

[7] Gath, I., &Geva, A. B.,Unsupervised optimal fuzzy clustering.IEEETransactions on Pattern Analysis and Machine Intelligence, 11, 773–780, 1989.

[8] Kolen, J. F., & Hutcheson, T.,.Reducing the time complexity of the fuzzy Cmeans algorithm. IEEE Transactions on Fuzzy Systems, 10, 263–267, 2002.

[9] MacQueen, J.B.. Some methods for classification and analysis of multivariate, 1967

[10] observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281–297.

[11] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., &Leisch, F..Miscfunctions of the department of statistics, 2014.

[12] Pal, N. R., &Bezdek, J. C.,.On cluster validity for the fuzzy C-means model. IEEE Transactions on Fuzzy Systems, Vol. 3, pp. 370–379, 1995.

[13] Peizhuang, W., Pattern recognition with fuzzy objective function algorithms, 1983.

[14] R Development Core Team,: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/, 2013.

[15] Wang, X., Wang, Y., & Wang, L.,Improving fuzzy C-means clustering basedon feature-weight learning. Pattern Recognition Letters, 25, 1123–1132.http://dx.doi.org/10.1016/j.patrec.2004.03.008, 2004.

[16] Xie, X. L., &Beni, G.,.A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13, 841–847, 1991. http://dx.doi.org/ 10.1109/34.85677.

[17] Yang, Q., Zhang, D., &Tian, F.,.An initialization method for fuzzy C-meansalgorithm using subtractive clustering.In 2010 third international conference on intelligent networks and intelligent systems pp. 393–396, 2010.

[18] Zou, K., Wang, Z., & Hu, M.,.An new initialization method for fuzzy C-means algorithm. Fuzzy Optimization and Decision Making, Vol. 7, pp. 409–416, 2008.