International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 6 (2014), pp. 577-582

http://www. irphouse.com

Selection of n in K-Means Algorithm

Gaurav Kant Chaturvedi[1], Vineet Jain[2], Divya Sharma[3] and Pooja Madkan[4]

[1, 2, 3] Department of CSE and IT, ITM University Gurgaon, INDIA
[4]Department of CSE, Amity University, UP, INDIA

ABSTRACT

One of the most popular and widely used algorithms in data clustering is K-means algorithm. However, one of its main downsides is that user has to specify number of clusters that is n, before the algorithm is to be implemented anywhere. This paper reviews existing methods for choosing the available number of clusters for the algorithm as well as the factors which are affecting the selection and proposed measure in the end of the paper which assist the selection. This paper provides an analysis of the results by using the proposed measure which helps in determining the number of clusters for the K-means algorithm for diverse range of data sets.

Keywords- data clustering and k-means algorithm

1. INTRODUCTION

K-Means clustering is an example of a partitioning algorithm. The grouping of data points is based on similarity between these points .on similarity, but the similarity between the clusters formed is dependent on the number clusters the algorithm is told to find. So if the algorithm is told to find too little number of clusters than the actual similarity between data points, dissimilar points may be included within the same cluster. Conversely, choosing too many clusters may lead to many clusters being very similar.

This paper proposes a method which is based on the information that is obtained while implementing the clustering operation of K –means which will help for selecting the number of clusters n. This proposed method focuses on the evaluation criteria to measure and suggest as well the suitable values for n, and thereby rendering unfit for the use of trial and error.

2. SELECTION OF K-MEANS ALGORITHM

Some techniques for selecting the value of n in K-Means clustering algorithm include:

a) Specifying the value of n within a range or set.
b) Value of n being specified by the user.
c) Value of n is determined in a later step of the algorithm.
d) Value of n is equal to the number of generators.
e) Determined by statistical methods.
f) Determined through visualization.

The performance of the algorithm depends on the value of n chosen. So for flexibility one can use, (a), instead of a predefined value of n, a set of values for n as it is important to have a value large enough such that it reflects the characteristics of datasets and also a number much smaller than the number of data points [1].On the other hand, if the user chooses to specify the value of n his/her self (b), to find a coherent clustering result, many iterations of the algorithm might be needed. If the value of n is chosen at a later step in the algorithm (c), it is being employed as a black box [2]. The number of clusters, n, is determined by the specific requirements of the main processing algorithm. (d) is used to determine the value of n only when we deal with synthetic data sets , which are mostly useful when testing algorithms. These datasets are populated using uniform distribution generators; so the number of clusters is equal to the number of generators with the assumption that any cluster resulting from the algorithm will cover all the objects yielding from that generator. The drawback with this is that there may be cases where a data point generated from generator Ga might be covered by cluster B (shown in fig A and B) and vice versa.

Several statistical techniques are available (e) for selecting n. These techniques work on the underlying assumption about the distribution of the datasets [3] but this technique does not explain the distortion inside a cluster , so , a cluster created by using this method may not correspond to a cluster in a partitioning clustering and vice versa .

Therefore, statistical methods are not applicable in one class. Most real world problems do not satisfy this assumption.

The easiest technique for determining the value of n is visual verification. It is also easy to explain. Visual examples may be used to present probable clustering outcomes or shortcomings of an algorithm. Visual technique primarily involves discerning data distribution graphically on the basis of positioning, shape and size.

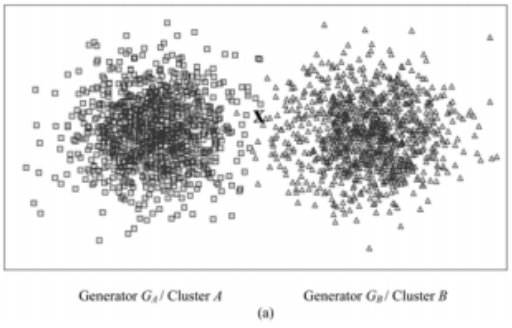

Figure 1[3]

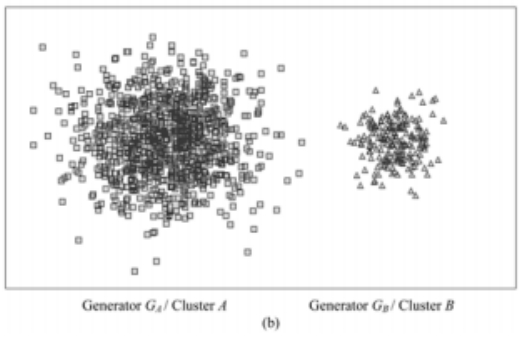

Figure 2[5]

3. ANALYSIS OF K-MEANS

The A function f(k) is used to evaluate the clustering result to select number of clusters. This function takes into account certain factors:

a. Evaluation criteria should be closely related to the clustering criteria. For example, such criteria could be minimization of the randomness of clusters.
b. Level of Detail: Having low level of detail is useful in gaining an overview of an object. By increasing the level of detail, one could get more information about the object but with the added expense of having to process more data. In clustering, a similar approach could be used where a dataset having k objects is grouped together and form a number of clusters between 1 and k which would correspond to the lowest and highest levels of detail.
c. Internal distribution versus global impact: Irregularities could be located using clustering and also to regions where data is concentrated can be identified. But, contrary to common sense, not every region having high concentration can be called a cluster. For it to be classified as such, it is imperative to not only analyse its inner distribution but also its interdependency with other object groupings also. In K-Means, distortion of a cluster can be calculated by using a function which consist of three things that is: data population, distance between objects and the centre of the cluster [5] which is given as:

$$I_j = \sum_{t=1}^{N_j} [d(x_{jt}, w_j)]^2$$

[3] Ij is the distortion of the cluster j, Wj is the centre and Nj is the number of clusters belonging to j. Xjt is the tth object which belongs to a cluster j, and d (Xjt , Wj) is the distance between the object Xjt and the centre . Individual cluster has its specific sort of distortion and its consequence on the dataset can be calculated by its contribution to sum of all the distortions i.e. Sk. This information is useful in determining if a particular region qualifies as a cluster or not.

The performance of function f(k) is verified by applying a series of experiments on the artificially generated data sets . All data are normalized before the K-Means algorithm is applied with the K ranging values that varies from 1 to 19 and f(k) is calculated by taking whole distortion of the clusters . In the results, one can notice that the minimum values of f(k) do not vary significantly from the average value for any given recommendation by the user .

The implementation of K-Means clustering Algorithm shown in the graph has been done using RStudio v 3.0.0, using extensively the R Programming language and tools offered by RStudio. The Graph represents a 2 Dimensional view of the Clustering results on a 4 Dimensional data set when number of clusters is set to be 4.

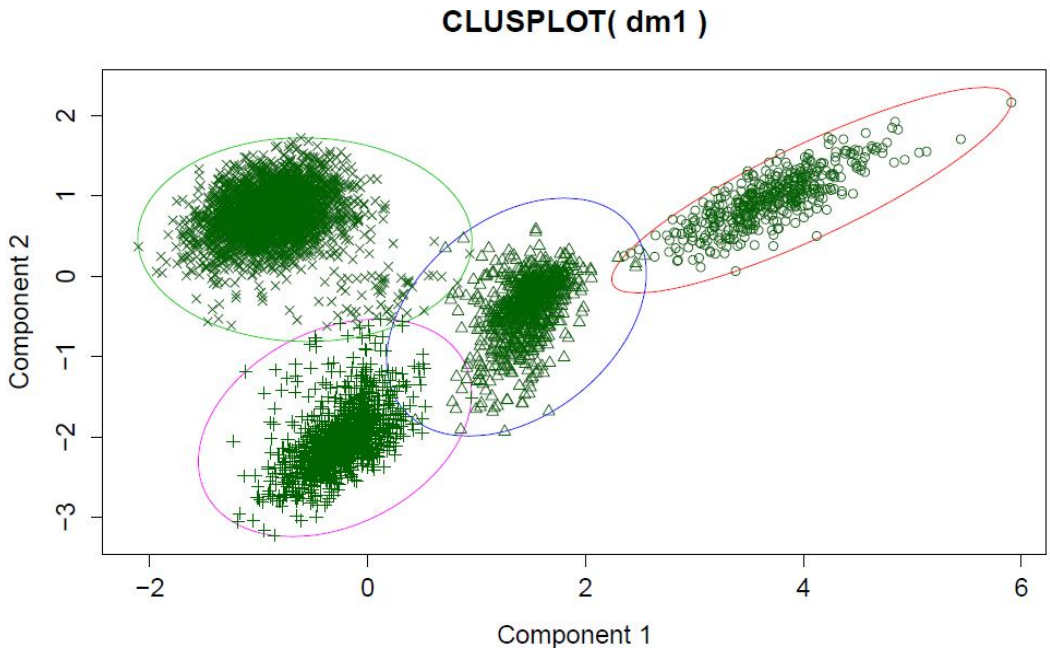

Figure 3: Result Graph

4. CONCLUSION

After every given technique for determining the value of n in K-Means algorithm has some inconsistencies to some degree, with some results being appropriate on most datasets but being poor on one or two datasets. This is due to the way data in these datasets is distributed. Finding a general method that works universally for all datasets is difficult so each dataset requires its own due diligence as to which method needs to be used.

A new method to find the number of clusters for K-Means algorithm has been proposed. This method takes into account information reflecting the performance of the algorithm. It also propose multiple values of n for such cases when different clustering results can be achieved by applying different required levels of detailing. Much more research is required to confirm the proficiency of this method when applied to datasets with highly complex distributions like “Temperature/K”.

Acknowledgements: It is with great sense of satisfaction that I present my real venture in computer science in form of a report. I wish to express my heartfelt thanks to all those who assisted me during the research.

5. REFERENCES

[1] M. B Al-Daoud, N. B. Venkateswarlu , and S.A.Roberts (1996), New methods for the initialization of clusters, Pattern Recognition Lett. Volume 17, pp. 451–455.

[2] J. Han and M .Kamber (2000), Data Mining: Concepts and Techniques, *Morgan Kaufmann*, San Francisco, California.

[3] D T Pham, S.S Dimov and C D Nguyen (2004), Selection of K in K- Means Clustering, *In Proceedings of IMechE Vol. 219 Part C*: J. Mechanical Engineering Science

[4] D.Pelleg, and A.Moore (2000), X-means: extending K-means with efficient estimation of the number of clusters, *In Proceedings of the 17th International Conference on Machine Learning*, Stanford, California, pp.727–734.

[5] K.Alsabti, S.Ranka, and V. Singh(1998), An efficient K-means clustering algorithm, *In Proceedings of the First Workshop on High-Performance Data Mining*, Orlando, Florida.