ISSN: 2321-7782 (Online)

Volume 1, Issue 6, November 2013

International Journal of Advance Research in Computer Science and Management Studies

Research Paper

Available online at: www.ijarcsms.com

Review on determining number of Cluster in K-Means Clustering

Trupti M. Kodinariya[1]
Research Scholar in JJT University,
Department of Computer Engineering,
Jhunjhunu, Rajasthan - India

Dr. Prashant R. Makwana [2]
Director - Research
GRMECT Research Center
Rajkot - India

Abstract: Clustering is widely used in different field such as biology, psychology, and economics. The result of clustering varies as number of cluster parameter changes hence main challenge of cluster analysis is that the number of clusters or the number of model parameters is seldom known, and it must be determined before clustering. The several clustering algorithm has been proposed. Among them k-means method is a simple and fast clustering technique. We address the problem of cluster number selection by using a k-means approach We can ask end users to provide a number of clusters in advance, but it is not feasible end user requires domain knowledge of each data set. There are many methods available to estimate the number of clusters such as statistical indices, variance based method, Information Theoretic, goodness of fit method etc...The paper explores six different approaches to determine the right number of clusters in a dataset

Keywords: Akaike's information criterion, Bayesian inference criterion, Clustering, Cross-validation, Elbow Method, Jump Method, Number of Cluster, Silhouette.

I. INTRODUCTION

Clustering, as a generic tool for finding groups or clusters in multivariate data, has found wide application in biology, psychology and economics .One of the main difficulties for cluster analysis is that, the correct number of clusters of different types of datasets is seldom known in practice. However, most of clustering algorithms are designed only to investigate the inherited grouping or partition of data objects according to a known number of clusters. Thus, identifying the number of clusters is an important task for any clustering problem in practice albeit it must be faced with many operational challenges. A tractable way for cluster analysis is to ask the end user to input the number of clusters in advance, which needs the expert domain knowledge over the underlying datasets. On the other hand, many statistical criteria or clustering validity indices have been investigated in the sense of automatically selecting an appropriate number of clusters.

Several algorithms have been proposed in the literature for clustering. The *k*-means clustering algorithm is the most commonly used [1] because of its simplicity. In this paper, we focus on one of problem of K-mean i.e .automation of number of cluster.

In the literature several approaches have been proposed to determine the number of clusters for k-mean clustering algorithm. We focus on six different approaches : i) By rule of thumb; ii) Elbow method; iii) Information Criterion Approach; iv) An Information Theoretic Approach; v) Choosing k Using the Silhouette and vi) Cross-validation.

The paper is organized as follows. A generic version of K-Means is described in Section 2. Section 3 contains a review of methods for finding the right K in K-Means in the published literature. Section 4 concludes the paper.

II. GENERIC VERSION OF K-MEANS ALGORITHM

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group is done. At this point, it is needed to re-calculate k new centroids as centers of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop has been generated. As a result of this loop it may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$W(S, C) = \sum_{k=1}^{K} \sum_{i \in S_k} \| y_i - c_k \|^2 \qquad (1)$$

Where S is a K-cluster partition of the entity set represented by vectors y_i ($i \in I$) in the *M*-dimensional feature space, consisting of non-empty non-overlapping clusters S_k, each with a centroid c_k $(k=1,2,\ldots K)$.

The algorithm is composed of the following steps:

1. Place k points in the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the k centroids.
4. Repeat Step 2 and 3 until the centroids no longer move.

III. DIFFERENT APPROACHES TO SELECTING THE RIGHT NUMBER OF CLUSTERS IN K-MEANS CLUSTERING

There have been a number of different proposals in the literature for choosing the right K after multiple runs of K-Means, among them we focus on following approaches.

A. By rule of thumb

B. Elbow method

C. Information Criterion Approach

D. An Information Theoretic Approach

E. Choosing k Using the Silhouette

F. Cross-validation

A. By rule of thumb

It is a simple method. This method can by apply to any type of data set.

$$k \approx \sqrt{n/2}$$

Where *n* is the number of objects (data points).

B. *Elbow Method*

The oldest method for determining the true number of clusters in a data set is inelegantly called the elbow method [2]. It is a visual method. The idea is that Start with K=2, and keep increasing it in each step by 1, calculating your clusters and the cost that comes with the training. At some value for K the cost drops dramatically, and after that it reaches a plateau when you increase it further. This is the K value you want.

The rationale is that after this, you increase the number of clusters but the new cluster is very near some of the existing. In the fig. 1, the distortion J () goes down rapidly with K increasing from 1 to 2, and from 2 to 3, and then W reach an elbow at K=3, and then the distortion goes down very slowly after that. And then it looks like maybe using three clusters is the right number of clusters, because that's the elbow of this curve. Distortion goes down rapidly until K=3, and really goes down very slowly after that hence number of cluster needed for this data set is 3.

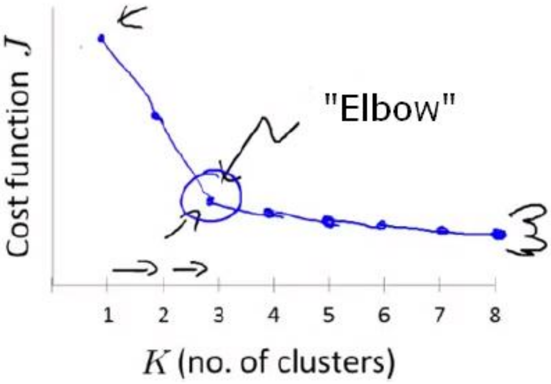

Fig. 1 identification of Elbow point

Problem with elbow method: This "elbow" cannot always be unambiguously identified. Sometimes there is no elbow, or several elbows as shown in fig. 2

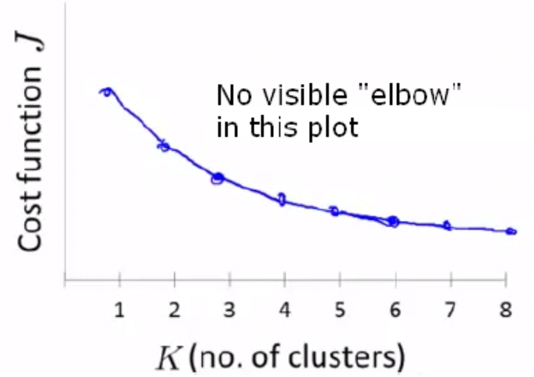

Fig. 2: ambiguity to identifying elbow point

C. *Information Criterion Approach*

As the number of clusters in the mixture model increase results in an increase in the dimensionality of the model, causing a monotonous increase in its likelihood. If we were to focus on finding the maximum likelihood model with any number of clusters, we would ultimately end up with a model in which every data point is the sole member of its own cluster. Obviously, we wish to avoid such a construction, and hence we must choose some criteria that do not depend solely on the likelihood.

An Information Criteria parameter is used for selecting among models with different number of parameters. It seeks to balance the increase in likelihood due to additional parameters by introducing a penalty term for each parameter.

It is model selection based Two-Step cluster procedure, where the user can choose an automatic selection of the cluster number, based any of information criteria.

The model selection techniques are relied upon to determine the number of clusters based on mixture models [3, 4]. Conventionally, model selection is implemented in two phases. In the first phase, we obtain a set of candidate models by some learning principles (usually by maximum likelihood (ML) learning) for a range of models. In the second phase, we select the appropriate model based on some model selection criterion. Popular examples of model selection criteria include Akaike's information criterion (AIC) [5], the consistent Akaike's information criterion (CAIC) [7], and the minimum description length (MDL) criterion [8, 6] which formally coincides with the Bayesian inference criterion (BIC) [9].

AIC and BIC are defined as

$$\text{AIC} = -2 * \ln(\text{likelihood}) + 2 * k$$

$$\text{BIC} = -2 * \ln(\text{likelihood}) + \ln(N) * k$$

Where k is the model degrees of freedom calculated as the rank of variance–covariance matrix of the parameters e(V) and N is the number of observations used in estimation or, more precisely, the number of independent terms in the likelihood. Operationally, N is defined as e(N) unless the n() option is specified.

D. An Information Theoretic Approach

An alternative approach to choosing the number of clusters that makes limited parametric assumptions, can be rigorously theoretically motivated using ideas from the field of rate distortion theory, is both simple to understand and compute, and is highly effective on a wide range of problems. The procedure is based on "distortion" which is a measure of within cluster dispersion.

Sugar and James introduce jump statistic which utilizes the criterion *W* in eq.1 extended according to the Gaussian distribution model [10]. Specifically, the distance between an entity and centroid in eq. 1 is calculated as

$$d(i,\ c_k)=(y_i-c_k)^T\Gamma_k^{-1}(y_i-c_k)$$

Where *Γk* is the within cluster covariance matrix.

The jump is defined as $JS(K) = W_K^{-M/2} - W_{K-1}^{-M/2}$ assuming that $W0^{-M/2} \equiv 0$. The maximum jump *JS (K)* corresponds to the right number of clusters. This is supported with a mathematical derivation stating that if the data can be considered a standard sample from a mixture of Gaussian distributions at which distances between centroids are great enough, then the maximum jump would indeed occur at *K* equal to the number of Gaussian components in the mixture.

E. Choosing k Using the Silhouette

A number of approaches utilize indexes comparing within-cluster distances with between cluster distances: the greater the difference the better the fit; many of them are mentioned in Milligan and Cooper [11].

Two of the indexes are: (a) the point-biserial correlation, that is, the correlation coefficient between the entity-to-entity distance matrix and the binary partition matrix assigning each pair of the entities 1, if they belong to the same cluster, and 0, if not, and (b) its ordinal version proposed by Hubert and Levin [12].

A well-balanced coefficient, the silhouette width, which has shown good performance in experiments, was introduced by Kaufman and Rousseeuw [13-14]. The concept of silhouette width involves the difference between the within-cluster tightness and separation from the rest. Specifically, the silhouette width *s(i)* for entity $i \in I$ is defined as

$$s(i)=\frac{b(i)-a(i)}{\max(a(i),b(i))}$$

Where *a(i)* is the average distance between *i* and all other entities of the cluster to which *i* belongs and *b(i)* is the minimum of the average distances between *i* and all the entities in each other cluster. The silhouette width values lie in the range from—1 to 1. If the silhouette width value for an entity is about zero, it means that that the entity could be assigned to another cluster as well. If the silhouette width value is close to—1, it means that the entity is misclassified. If all the silhouette width values are close to 1, it means that the set I is well clustered.

A clustering can be characterized by the average silhouette width of individual entities. The largest average silhouette width, over different K, indicates the best number of clusters.

F. Cross-validation

Cross-validation is another approach for estimating the number of clusters proposed by Smyth [15]. It is based on cluster stability. This method splits the data in two or more parts. One part is used for clustering and the other part(s) is used for validation.

The idea behind clustering stability is that a “good” algorithm tends to repeatedly produce similar clusterings on data originating from the same source. In other words, the algorithm is stable with respect to input randomization.

A major issue, as pointed out by Krieger and Green [16] is that a clustering model is stable only if the objective function has a unique global minimize. It is demonstrated that the approaches mentioned in the article, fail to determine the appropriate number of clusters, especially if the sample size gets larger and the variable exhibit higher correlation. The clustering stability might not be that desirable for determining the value of k.

Ben-Hur, Elisseeff and Guyon [17] propose to use distribution of pairwise similarity between clusterings of sub-samples of a dataset as a measure of the stability of a partition. In this paper several experiments are run, and all the results coincide with the intuitive selection.

Lange, Roth, Braun and Buhmann [18] proposed a new measure of clustering stability to assess the validity of a cluster model. Good performances have been achieved on both, simulated data and gene expression data sets.

Another great paper on this idea is presented by Ben-David, von Luxburg and Pal [19]. In this work, the authors propose some new definitions of stability and some related clustering notions. The results suggest that the existence of a unique minimize indicates stability, and the existence of a symmetry permuting such minimize indicates instability. The results indicate that stability does not reflect the validity or meaningfulness of the selection of the number of clusters. Instead, the parameters it measures are independent of clustering parameters

A modified cross validation is proposed by wang [20]. He focuses in his paper on introducing some novel criteria for determining the number of clusters. This new selection criterion measures the quality of clusterings through their instability from sample to sample. Here the clustering instability is estimated through cross validation, and the goal of the method is to minimize the instability. The data is divided into two training sets and one validation set to imitate the definition of stability. Then, a distance based clustering algorithm is applied on the independent and identically distributed training sets and the inconsistencies evaluated on the validation set. This method has been proven to be effective and robust on a variety of simulated and real life examples.

IV. CONCLUSION

K-Means arguably is the most popular clustering method. This is why studying its properties is of interest not only to the classification, data mining and machine learning communities, but also to the increasing numbers of practitioners in marketing research, bioinformatics, customer management, engineering and other application areas.

This paper addresses one of the most controversial issues in clustering: the right number of clusters, which some may view as baseless because in many cases, "clusters are not in data but in the viewing eye." In our survey, we explore the case when clusters, though not exactly conventional, are in data.

References

1. R. C. Dubes and A. K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988.
2. Andrew Ng, Clustering with the K-Means Algorithm, Machine Learning, 2012
3. H. Bozdogan. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan, editor, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, volume 2, pages 69–113, Dordrecht, the Netherlands, 1994. Kluwer Academic Publishers .
4. Xu. BYY harmony learning, structural RPCL, and topological self-organizing on mixture models. Neural Networks, 15:1125–1151, 2002
5. H. Akaike. A new look at statistical model identification. IEEE Transactions on Automatic Control, 19:716–723, 1974.
6. A. Barron and J. Rissanen. The minimum description length principle in coding and modeling. IEEE Trans. Information Theory, 44:2743–2760, 1998.
7. H. Bozdogan. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika, 52(3):345–370, 1987.
8. J. Rissanen. Modeling by shortest data description. Automatica, 14:465–471, 1978.
9. G. Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464, 1978.
10. C. A. Sugar and G. M. James. Finding the number of clusters in a dataset: an information-theoretic approach. Journal of the American Statistical Association, 2003.
11. G. W. Milligan and M. C. Cooper. An examination of procedures for determiningthe number of clusters in a data set. Psychometrica, 1985.
12. HUBERT, L.J., and LEVIN, J.R., A General Statistical Framework for Assessing Categorical Clustering in Free Recall", Psychological Bulletin, 1976
13. POLLARD, K.S., and VAN DER LAAN, M.J., A Method to Identify Significant Clusters in Gene Expression Data, U.C. Berkeley Division of Biostatistics Working Paper Series, p. 107, 2002
14. KAUFMAN L., and ROUSSEEUW P., Finding Groups in Data: An Introduction to Cluster Analysis, New York: J. Wiley & Son, 1990.
15. Smyth, P. Clustering using Monte Carlo cross validation. In Proc. 2nd Intl. Conf. Knowl. Discovery & Data Mining (KDD-96), Portland, 1996.
16. Krieger, A. M. & Grenn, P. E. . A cautionary note on using internal cross validation to select the number of clusters. Psychometrika , 1999.
17. Ben-Hur, A., Elisseeff, A. & Guyon, I.. A stability based method for discovering structure in clustered data. In Pac. Symp. Biocomp. 2002
18. Lange, T., Roth, V., Braun, M. & Buhmann, J., Stability-based validation of clustering solutions. NeuralComp, 2004.
19. Ben-David, S., Von Luxburg, U. & Pal, D., A sober look at stability of clustering. In Proc. 19th Ann. Conf. Learn. Theory (COLT 2006), Ed. G. Lugosi and H. Simon, pp. 5–19. Berlin: Springer, 2006
20. Wang, J., Consistent selection of the number of clusters via cross-validation. Biometrika, 2010

AUTHOR(S) PROFILE

Trupti Kodinariya, received the B.E. and M.E. degrees in Computer Engineering from North Gujarat University in 2002 and Dharmasinh Desai Institute of Technology,- Nadiad in 2005, respectively. During 2005-2009, she worked as Sr. lecturer at Charotar University of Science and Technology. She now with Atmiya institute of technology and science -rajkot. She published many papers in International/National Conferences and Journals. Her areas of research are Image Processing, pattern Recognition, Data Mining and Neural Network.

Dr. Prashant Makwana, received the M.Sc. and Phd degrees from Saurashtra University. He worked as Software Developer/Sr.Software Developer/Project manager in IT Company for 11 years (AMAZON IT Industry). He now with GRMECT Research Center as Director. He carried out 17 research projects. He published many papers in International/National Conferences and Journals.