Approach to determining the number of clusters in a data set

Ivan Ishchenko, Larysa Globa, Yurii Buhaienko, Andrii Liashenko
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
ivanishchenkoo@gmail.com, lgloba@its.kpi.ua
yura.buhaenko@gmail.com, andrey.lyashenko44@gmail.com

***Abstract*—For different decision-making systems, they build knowledge bases with sets of fuzzy logic rules and when constructing these rules on the basis of statistical information, a complex question is the determination of the number of clusters. The article is devoted to the analysis of methods that allow to automatically determine the number of clusters and their application in decision-making systems. The analysis conducted helped to distinguish the elbow method as the most suitable of all the scanned ones. This was able to find the optimal number of clasters on a test data set.**

***Keywords*—Clustering, decision-making system, data analyses, optimization, fuzzy logic, elbow method, X-means clustering, silhouette method.**

I. Introduction

Today clustering tasks are relevant for many areas of activity. Clustering is intended to divide a set of objects into homogeneous groups, and its purpose is to search for existing structures. This process is used in computer graphics - for image segmentation, for classifying search results, for processing tables and documents, in marketing - for identifying groups of customers, buyers, and goods for developing promotion strategies. At the same time, each data domain has its own particular data sets, for example, in technical data collection systems one has to work with numerical characteristics that have a unique assessment, and for example, when working with user / enterprise data, the data has a completely different format. Based on this, different clustering and data processing algorithms are used.

The problem of determining the number of clusters is one of the main unsolved problems of cluster analysis. The two most used types of cluster analysis procedures are: hierarchical and iterative. For iterative algorithms, the number of clusters is one of the input parameters of the algorithm. For hierarchical procedures, visual analysis of a dendrogram is typical, and the most preferred number of clusters is determined from it. [1]

Despite the apparent diversity, so far no universal algorithm has been found that would be effective for data of different nature.

Most of the existing methods are based on indexes comparing the degree of "scatter" of data within clusters and between clusters, on the calculation of the values of heuristic characteristics (stability functions), showing compliance assigned clusters for selective elements of the set, on the statistics defining the most likely solution, either by estimating the density of distributions. The difference between the levels of association, which can be determined by the dendrogram, is the simplest and most popular solution.

However, this visual analysis of the dendrogram is extremely difficult when:

1) a large number of objects under consideration;
2) implicit expressiveness of the data structure.

For the k-means clustering algorithm, the input parameter k is used, which determines the number of clusters. The parameter k may be erroneous. It depends on the shape and scale of the distribution of points in the data set. The number of clusters can be from one to $n - 1$, where n is the number of objects in the sample. Ie all objects belong to one cluster or each object is a cluster. [2]

If the number of clusters k from a given data set is not obvious or is not specified by an expert, there are methods for it that help to make a decision. These are direct methods and methods of statistical testing: 1. Direct methods: these are optimization of the criterion within cluster sums of squares (the "elbow" method) or the average silhouette. 2. Methods of statistical testing: consists of comparing evidence against the null hypothesis. An example is the statistics gaps. The method is selected depending on the characteristics of the data set.

One of the important issues dealt with in this paper is a problem of the clusters number definition in the process of clustering that based on the statistical data. The number of clusters determines exactly the fuzzy logical rules number that formed the fuzzy knowledge base. Based on this the correct determination of the clusters number has a significant impact on the quality of the resulting fuzzy logic rules and consequently the quality of the fuzzy knowledge base in general.

II. CLUSTERING METHODS ANALYS

A. The eblow method

The elbow method is based on the use of the percentage of variance, which is a function of the number of clusters. The main task is to determine the number of clusters, such that adding another cluster should not provide a significant improvement in data modeling. This means that if we build the percentage of variance explained by clusters against the number of clusters, the first clusters will add a lot of information (they will explain a large variance), with a subsequent change in the number of clusters, the limiting gain starts to decrease sharply, and a clear sectional error appears on the graph. At this point of inflection, the number of clusters is determined, and from this the name "elbow criterion" emerges. But this inflection point may not always be uniquely identified [1]. The percentage of variance is determined by the ratio of the variance between groups to the total variance, also known as the F-test. A slight change in this method shows the curvature of the intra-group dispersion [2,3].

The optimal number of clusters can be determined as follows:

1) Calculate the clustering algorithm (for example, k-means clustering) for different values of k. For example, varying k from 1 to 10 clusters.
2) For each k, calculate the total intracluster sum of the square (wss).
3) Get the wss curve in accordance with the number of clusters k.
4) The location of the bend on the graph is usually considered as an indicator of the corresponding number of clusters. [3]

Consider a multivariate observation $x_i = (x_{i_1}, x_{i_2}, ..., x_{i_p})'$, $i = 1, ..., n$, containing n independent objects measured on p variables. For any partition of the n objects into g clusters (P_g), denote by C_m the set of objects allocated to the mth cluster and by nm the number of objects in C_m, $m = 1, \cdots, g$. Denote by d_i,i' the distance between objects i and i'. The sum of pairwise distances between objects in the mth cluster is given by

$$D_m = \sum_{i,i' \in c_m} d_{i,i'} \tag{1}$$

For a fixed value of g, define

$$W_g = \sum_{m=1}^{g} \frac{1}{2n_m} D_m \tag{2}$$

Note that W_g in (2) is a typical measure of the within-clusters homogeneity associated with P_g, a small value of which reflects a good fit of a classification to the "true" cluster structure of data.

In the above definition of W_g, d_i,i' can be any arbitrary measure of distance. If the squared Euclidean distance is used, simple mathematical derivation shows that W_g is monotonically decreasing in g. Hence, W_g is not informative in choosing the optimal number of clusters by itself. However, for data strongly grouped around G centers, it is expected that the value of function W_g will drop quickly as g increases until it reaches the "true" number of clusters in the data. Intuitively, W_g will decrease at a much slower rate when $g > G$ since with more than G centers, objects belonging to the same cluster will be partitioned.[4,5]

Therefore, an "elbow" point in the curve of W_g may indicate the optimal estimate of the number of cluster in data.

In estimating the number of clusters in a data set, methods based on the W_g criterion are aimed at appropriately determining the "elbow" point in W_g, where W_g is sufficiently small. The idea of the gap method is to compare the curve of W_g from the original data to the curve of the expected) under an appropriate null reference distribution. The best estimate of the cluster number is $\hat{g}$ if W_g falls farthest below the expected curve at $g = \hat{g}$. Defining the gap statistic as

$$Gap_n(g) = E_n^* \{log(W_g)\} - log(W_g) \tag{3}$$

the estimate g is the value of g which maximizes $Gap_n(g)$.

An essential step of the gap method is to generate suitable reference data sets which are used to obtain the benchmark of the within-clusters dispersion for comparison. The reference data can be generated by incorporating information about the shape of the data distribution. By definition, application of the gap method does not depend on the clustering method used. For example, Tibshirani et al. implemented the gap method under the contexts of both K-means and hierarchical clustering methods in their research. Simulation studies showed that the gap method is a potentially powerful procedure in estimating the number of clusters for a data set. Moreover, the gap method has the advantage over most of the other estimating methods that it can be used to test the null hypothesis about homogeneous non-clustered data against the alternative of clustered data.

However, a deficiency of the gap method in finding the correct number of clusters has been demonstrated in more recent studies. For example, the gap method failed to detect the 4-cluster structure in the simulated data which contain well-separated clusters generated from distinct exponential distributions. In microarray data analysis, Dudoit and Fridlyand developed the Clest method and compared it with several other existing methods including the gap method. They noted that the gap method tends to overestimate the number of clusters. One possible reason for such a deficiency in using the gap method may be because W_g, a statistic summarizing the within-clusters homogeneity, is not suitable in measuring the clustering adequately.[5]

B. The silhouette method

The average data power criterion is a criterion for estimating the natural number of clusters. The silhouette of a data instance is a measure of how closely it is compared with the data within its cluster and how weakly it is compared with the data of the neighboring cluster, that is, the cluster, the average distance from which is the lowest. The silhouette value maximally close to 1 means that the base point belongs to the corresponding cluster. When the silhouette value is close to -1, it can be concluded that the point does not belong to the given cluster.

Optimization techniques, such as genetic algorithms, are useful in determining the number of clusters that give the largest silhouette. It is also possible to scale the data so that the silhouette is maximized with the correct number of clusters. In general, this means that it measures the quality of clustering and determines how well each object is located within its cluster. The higher the average width, the better the clustering. The mean silhouette method calculates the mean observation silhouette for different values of k. The optimal number of clusters k is one that maximizes the average silhouette in the range of possible values for k. [3,6]

The algorithm is similar to the elbow method and can be computed as follow:

1) Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2) For each k, calculate the average silhouette of observations (avg.sil).
3) Plot the curve of avg.sil according to the number of clusters k.
4) The location of the maximum is considered as the appropriate number of clusters. [3]

C. Silhouette statistic

Kaufman and Rousseeuw proposed the silhouette index as to estimate the optimum number of clusters in the data. The definition of the silhouette index is based on the silhouettes introduced by Rousseeuw, which are constructed to show graphically how well each object is classified in a given clustering output.[5] To plot the silhouette of the mth cluster, for each object in C_m, calculate $s(i)$ as

$a(i)$ = average dissimilarity of object i to all other objects in the mth cluster

$d(i, C)$ = average dissimilarity of object i to all other objects in cluster C, $C \neq Cm$

$$b(i) = min_{C \neq C_m} d(i, C)$$

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

The silhouette index, denoted by $s(\bar{g})$, is defined as the average of the $s(i)$ for all objects in the data. $s(\bar{g})$ is called the average silhouette width for the entire data set, reflecting the within-cluster compactness and between-cluster separation of a clustering. Compute $s(\bar{g})$ for $g = 1, 2, \cdots$. The optimum value of g is chosen such that $s(\bar{g})$ is maximized over all g:

$$\hat{G} = arg\ max_g s(\bar{g}).$$

D. Gap statistic method

The gap statistic has been published by R. Tibshirani, G. Walther, and T. Hastie (Standford University, 2001). The approach can be applied to any clustering method. The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic)[7,8]. This means that the clustering structure is far away from the random uniform distribution of points. [3]

The algorithm works as follow:

1) Cluster the data under investigation for fixed cluster number, k, where $k = 1, 2, \cdots$. Compute W_k for all values of g;
2) Generate B reference data sets in the way described above. Cluster each of the B reference data sets and calculate $W_b^*(k), b = 1, 2, \cdots, B$ and $k = 1, 2, \cdots$. Compute the gap statistic

$$Gap(k) = (\frac{1}{B}) \sum_b log(W_b^*(k)) - log(W(k))$$

3) Compute the standard deviation

$$sd_k = \left[\left(\frac{1}{b} \right) \sum_b \left\{ (W_b^*(k)) - \bar{l} \right\}^2 \right]^{\frac{1}{2}},$$

where

$$\bar{l} = (\frac{1}{B}) \sum_b log(W_b^*(k))$$

4) Define $s_k = sd_k \sqrt{1 + \frac{1}{B}}$ The optimum number of clusters is given by the smallest k such that

$$Cap(k) \geq Cap(k+1) - s_{k+1}\ [9, 10, 11].$$

III. Estimation of different methods for finding the number of clusters

A software solution was created to find the number of clusters in the data sample on different methods. Obtained results (Fig. 1):Elbow method: 4 clusters solution suggested

- Silhouette method: 2 clusters solution suggested
- Gap statistic method: 4 clusters solution suggested

The silhouette plots display a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from –1 to 1, where 1 means that points are very close to their own cluster and far from other clusters, whereas –1 indicates that points are close to the neighbouring clusters.

Gap statistic is a goodness of clustering measure, where for each hypothetical number of clusters k, it

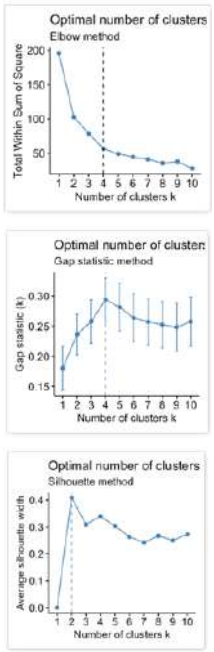

Figure 1. Number of clusters for different algorithms

compares two functions: log of within-cluster sum of squares (wss) with its expectation under the null reference distribution of the data. In essence, it standardizes wss. It chooses the value where the log(wss) is the farthest below the reference curve, ergo the gap statistic is maximum.

The elbow method maps the within-cluster sum of squares onto the number of possible clusters. As a rule of thumb, you pick the number for which you see a significant decrease in the within-cluster dissimilarity, or so called "elbow".

According to these observations, it's possible to define $k = 4$ as the optimal number of clusters in the data. As we can see from the three approaches we can to a certain extent be sure of what an optimal value for the number of clusters can be for a clustering problem. There are few other techniques which can also be used.

CONCLUSIONS

The article analyzes a number of clustering algorithms and their application in decision-making systems. In a cluster analysis, the fundamental problem is to determine the value of the number of clusters, which has a deterministic effect on clusterization results. However, the limitation in current applications is that there is no convincingly acceptable solution to the problem with the best cluster because of the high complexity of real data sets.

Choosing the appropriate clustering method is another important step in clustering. The k-medium clustering is one of the most popular clustering technologies used in practice.

According to the results of the study, it can be concluded that both the k-medium method and the method of agglomeration hierarchical clusterization can be successfully used for clustering in various application areas, with the results of this clusterization being close. The main disadvantage of the k-medium method is that it is necessary to predefine k - the number of clusters and standards, which is not always possible to make rational. The method is very sensitive to these initial approximations of the values of the centers. To eliminate this problem, you can use the method of gradually increasing the number of clusters.

The disadvantage of elbow and average silhouette techniques is that they measure only the general characteristics of clustering. A more complex way is to use the gap statistics, which provides a statistical procedure for formalizing a heuristic elbow / silhouette to estimate the optimal number of clusters.

In this article, we describe various methods for selecting the optimal number of clusters in the data set. Such methods include elbows, silhouette, and statistical methods of rupture.

Future work: Future work is to carry out research on the possibility and quality of the resulting solution in order to receive sets of rules for forms of fuzzy knowledge bases for decision-making systems in technical systems.

REFERENCES

[1] David J. Ketchen Jr; Christopher L. Shook. The application of cluster analysis in Strategic Management Research: An analysis and critique. Strategic Management Journal, 1996. Vol 17 (6), pp 441–458

[2] D. Pelleg; AW Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. International Conference on Machine Learning no17, 2000. pp 411–416

[3] David J. Ketchen. Determining the number of clusters in a data set. Available at: https://wiki2.org/en/Determining ¯the ¯number ¯of ¯clusters ¯in ¯a ¯data ¯set (accessed 2018, Nov)

[4] R.C. de Amorim & C. Hennig. Recovering the number of clusters in data sets with noise features using feature rescaling factors. Information Sciences, 2015. pp 126–145.

[5] Chatti Subbalakshmi, G. Rama Krishnab, S. Krishna Mohan Raoc, P. Venketeswa Raod. A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set. Procedia Computer Science, 2015. Vol 46. pp 346–353.

[6] Can, F.; Ozkarahan, E. A. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. ACM Transactions on Database Systems, 1990. Vol 15, pp 483.

[7] Charrad, Malika, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 2014. Vol 61, pp 1–36.

[8] Kaufman, Leonard, Peter Rousseeuw. "Finding Groups in Data: An Introduction to Cluster Analysis." Wiley, 2005. 368p.

[9] Andrzej Piegat. "Fuzzy Modeling and Control" Heidelberg, Physica–Verlag a Springer–Verlag Company, 2001. 728p.

[10] Mingjin Yan. Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion. Nov, 2005 Blacksburg, Virginia. pp.26–27.

[11] Voroncov K.V. Algoritmy klasterizacii i mnogomernogo shkalirovaniya. Available at: http://www.ccas.ru/voron/download/Clustering.pdf (accessed 2018, Nov)

ПОДХОД К ОПРЕДЕЛЕНИЮ КОЛИЧЕСТВА КЛАСТЕРОВ В НАБОРЕ ДАННЫХ

Ищенко И. А., Глоба Л. С., Бугаенко Ю. М., Ляшенко А. В.

Аннотация – Для различных технических систем принятия решений создают базы знаний с наборами правил нечеткой логики. При построении таких правил на основе статистической информации сложным вопросом является определение количества кластеров. Статья посвящена анализу методов, позволяющих автоматически определять количество кластеров с целью их применения в системах принятия решений. Проведенный анализ математических методов, позволяющих автоматически определять количество кластеров при построении нечеткой базы знаний, а значит и количество нечетких правил, позволяет выделить метод «локтя» как наиболее подходящий. Метод позволил найти оптимальное количество кластеров в наборе тестовых данных.

Received 10.12.18