Determining the number of clusters in a data set via repeated data clustering into two clusters

Jerzy Korzeniewski
University of Lodz, Lodz, Poland jurkor@wp.pl

Abstract

In the paper a new algorithm of determining the number of clusters in a data set of objects described with continuous variables is presented. The idea consists in repeated division of data set (or the clusters resulting from the previous step) into two clusters. We accept the division (raising the number of clusters by one) if a division quality measure exceeds the threshold and we forget it (reverting to the data set structure from before the division) if the measure falls below the threshold. One can apply different division quality measures but a new one based on the Rand index is proposed. The performance of the new index is compared with that of the leading indices constructed so far i.e. Calinski-Harabasz index and the gap index, in examples of selected data sets from the UCI repository.

Keywords: cluster analysis, number of clusters, Rand index, gap statistic, Calinski-Harabasz index.

1. Introduction

Determining the number of clusters in a data set is an important step of cluster analysis. Most of the indices used for the task are of a wrapper-optimization kind i.e. they try to find optimal value of the cluster number for an established grouping method. Most popular ones are the following indices: Baker-Hubert, Caliński–Harabasz, Dunn, Davies-Bouldin, Hartigan, Hubert-Levine, Krzanowski-Lai, Gap index. The quality of these indices was studied a couple of times e.g. *Milligan i Cooper* (1985), *Migdał-Najman i Najman* (2005, 2006), *Korzeniewski* (2005). In a broad study by Milligan and Cooper the Caliński-Harabasz (1971) index turned out to be the best. However, this study was performed nearly 30 years ago and, since then, new proposals were made, their authors claiming their superiority over other indices. An example of a newer index with a good opinion can be the Gap index (*Tibshirani et al.,* 2001). These two indices will be a reference point in the process of constructing a new index.

The value of the Caliński-Harabasz index for the number of clusters k is given by

$$CH(k) = \frac{tr(\mathbf{B}_k) \big/ (k-1)}{tr(\mathbf{W}_k) \big/ (n-k)} \quad , \qquad (1)$$

where $\mathbf{B}_k$ – between-cluster covariance matrix, $\mathbf{W}_k$ – within-cluster covariance matrix, n – the number of data set objects. The proper number of clusters is k maximizing expression (1). As can be observed, by means of this index one cannot decide whether a data set should be divided into any clusters at all.

The value of the Gap index for the number of clusters k is determined by means of the expression

$$Gap(k) = \frac{1}{B} \sum_b \log\left(tr(\mathbf{W}_{kb})\right) - \log\left(tr(\mathbf{W}_k)\right) \quad , \qquad (2)$$

$\mathbf{W}_{kb}$ is the within-cluster covariance matrix with the original random variable being replaced by the variable generated from the uniform distribution over the interval given by the sample spread of the original variable; B – number of repetitions of

uniform variables generations. The proper number of clusters is the smallest k fulfilling the condition

$$Gap(k) \geq Gap(k+1) - s_{k+1}, \qquad (3)$$

where $s_k = sd_k \sqrt{1+1/B}$, where sd_k denotes the standard deviation of B values of $\log\left(tr\left(\mathbf{W}_{kb}\right)\right)$. Instead of the uniform distribution one can try another technique of estimating the logarithm of the trace of covariance matrices which do not form any cluster structure, making use of the shape of the marginal distributions of the original variables. It is visible that the formula of the index can also be used to determine whether the data set should be divided into any clusters i.e. for k equal to 1. In the investigation we applied $B = 100$.

2. New index formulation

The idea of a new index which we intend to propose in this section is the multiple division of a data set (or a part of it) into two clusters and keeping the division when it turns out to be sufficiently good i.e. when both clusters are separated well enough. We will use the classical k-means method (k=2), with multiple random choice of starting points but using any other grouping procedure is also possible. The measure of the quality of the division we define as follows. Let us assume that a fixed subset of the data set was divided in the primary stage, into two clusters S_1 and S_2. In the next stage we consider a new subset given by the smaller of the two clusters and a third of the bigger (more numerous) cluster. The new subset is also divided into two clusters and the measure of the quality of the primary division is the corrected Rand index measuring the consistency of both divisions. We will denote this measure by $R(S_1, S_2)$. When the fixed subset is the whole data set then the value $R(S_1, S_2)$ is the final quality measure. However, when apart from the clusters S_1 and S_2 there were other clusters (or parts not qualified as clusters), then, the value $R(S_1, S_2)$ cannot be the final measure because it may happen that one of the clusters S_1 and S_2 (or both) is "close" to the remaining part of the data set. The value $R(S_1, S_2)$ may be accepted as the final measure when at least one of the clusters S_1 and S_2 has the same or bigger value of the division quality with all other clusters. Therefore, the final measure of the quality of the division into two clusters, when assuming that apart from the clusters S_1 and S_2 there are k-2 other clusters, is the number:

$$R(S_1, S_2, k) = \max_{i=1,2} \left\{ \min_{\substack{j=1,\ldots,k \\ j\neq 1}} R(S_1, S_j), \min_{\substack{j=1,\ldots,k \\ j\neq 2}} R(S_2, S_j) \right\} \qquad (4)$$

In order to apply measure (4) to determine the number of clusters in the data set we have to propose an algorithm of repeated divisions into two clusters, because the succession of the divisions (the mode of searching through the data set) may be meaningful for the final result. The threshold for measure (4) from which we accept the division of a part of data set into two clusters will be the number 0.4. Let us propose the following algorithm of the repeated data set division into two clusters.

Step 1. Set K=1 i.e. treat the whole data set as one cluster.
Step 2. Each cluster numbered k, (k=1,2,…,K) divide into two clusters and find the value m_k of measure (4) for this division if the numbers of both smaller clusters constitute at least 5% of n.

Step 3. From among all values m_k select the biggest number corresponding to cluster k_0 and (if this number is greater than 0.4) increase K by 1, at the same time replacing the cluster k_0 with two smaller clusters into which it was divided. Go to step 2. If the greatest of the numbers m_k is smaller than 0.4 go to step 4.

Step 4. Each cluster numbered k, (k=1,2,…,K) divide into two clusters (first division) and find the value of measure (4) for the follow–on division (second division) of each of these two clusters if the numbers of both clusters from the second division constitute at least 5% of n. From the two values of measure (4) choose the bigger one corresponding to cluster k_0 (from the second division) and, if it exceeds 0.4 then change the division of cluster k dividing it into cluster k_0 and the remaining two clusters. Go to step 2. If the bigger from the two values of measure (4) is smaller than 0.4 for all clusters k, k=1,2,…,K then merge the two first division clusters and go to step 5.

Step 5. Perform the same procedure as in step 4 dividing each cluster into three clusters. If one of these three clusters can be divided into two clusters then go to step 2. If none of the three clusters cannot be divided into two clusters then merge the three clusters from the initial division and go to step 6.

Step 6. The current K is the result pointing to the proper number of clusters. If visiting step 6 for less than 20 times go to step 1.

Step 7. Repeat steps 1-6 20 times. From the 20 candidates for the number of clusters pick the dominating number which will be the final number of clusters.

One has to remember that measure (4) corresponds only to dividing a cluster into two smaller clusters. The divisions from step 4 and 5 into two or three clusters are tentative divisions and are not assessed by means of measure (4). However, each of the clusters resulting from these tentative divisions is tried to be divided into two clusters and these divisions are assessed and the best one is picked. These tentative divisions are merged back with the exception of cluster k_0 (if such was found). The introduction of tentative divisions into two or three clusters is necessary because the only multistage division into two or three clusters may not give good results in the case of a cluster structure consisting of more than 5 clusters. In such a case dividing the whole data set into only two clusters may not return good value of measure (4) because each of the two clusters cannot be separated from the other one as they both are made up of a couple of small clusters. If we use a tentative division into two and, subsequently three, clusters we have better chances of finding a cluster which will be well enough separated from other clusters.

3. Data sets

In order to check the quality of the new proposal we applied it to some well known data sets from the UCI repository. We used the following data sets.

Iris data set. Objects are iris flowers constituting three different clusters. Number of objects: 150, number of variables: 4, true number of clusters: 3.

Glass data set. Objects are samples of glass. Number of objects: 214, number of variables: 9, true number of clusters: 5.

Wines data set. Objects are kinds of wine. Number of objects: 178, number of variables: 13, true number of clusters: 3.

Ionosphere data set. Objects are observation data. Number of objects: 351, number of variables: 34, true number of clusters: 2.

Concrete data set. Objects are samples of concrete. Number of objects: 1030, number of variables: 8, true number of clusters: 5.

4. Results and conclusions

On all data sets we performed all three compared indices. For the Caliński-Harabasz and Gap index we used *k*-means with random starting points as the grouping procedure. We repeated grouping 100 times and remembered the one with the smallest within group variance. The investigated range of possible numbers of clusters was from 2 to 15. We got the results presented in Table 1.

Table 1. The true and assessed numbers of clusters for all five data sets.

Data set	True number of clusters	Assessed number of clusters		
		New	CH	Gap
Iris	3	2	3	7
Glass	5	5	2	7
Wines	3	3	2	3
Ionosphere	2	8	11	2
Concrete	5	9	2	11

Source: own research.

The results are quite promising for the new proposal. The new index went totally wrong only in two out of five data sets. The two indices compared were not better. Therefore, in our opinion, a comprehensive simulation investigation is needed to asses the new proposal more adequately.

References

Caliński R.B., Harabasz J., (1971), "A dendrite method for cluster analysis", *Communications in Statistics* 1974, vol. 3.

Korzeniewski J., (2005), „Propozycja nowego algorytmu wyznaczającego liczbę skupień", *Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Taksonomia* 12.

Migdał-Najman K., Najman K. (2005), „Analityczne metody ustalania liczby skupień", *Taksonomia 12: Klasyfikacja i analiza danych – teoria i zastosowania*, Nr 1076, Wrocław.

Milligan G., Cooper M., (1985), "An examination of procedures for determining the number of clusters in a data set", *Psychometrika* 1985, No 2.

Najman K., Migdał-Najman K., (2006), „Wykorzystanie indeksu Silhouette do ustalania optymalnej liczby skupień", *Wiadomości Statystyczne*, 6.

Tibshrani R., Walther G., Hastie T., (2001), "Estimating the Number of Clusters in a Datase via the Gap Statistic", *Journal of the Royal Statistical Society* 32.