A New Approach to Determine the Optimal Number of Clusters Based on the Gap Statistic

Jaekyung Yang, Jong-Yeong Lee, Myoungjin Choi, Yeongin Joo

A new approach to determine the optimal number of clusters based on the Gap statistic

Jaekyung Yang[1[0000-0002-4904-1351]], Jong-Yeong Lee [1[0000-0001-7819-005X]], Myoungjin Choi[2[0000-0003-2919-8252]], Yeongin Joo [1[0000-0002-6061-7876]]

[1] Dept. of Industrial and Information Systems Engineering, Jeonbuk National University, Jeonju, Jeonbuk 54896, Republic of Korea
[2] Howon University, 64 Howondae 3gil, Impi, Gunsan city, Jeonbuk 54058, Republic of Korea
jkyang@jbnu.ac.kr

Abstract. Data clustering is one of the most important unsupervised classification method. It aims at organizing objects into groups (or clusters), in such a way that members in the same cluster are similar in some way and members belonging to different cluster are distinctive. Among other general clustering method, k-means is arguably the most popular one. However, it still has some inherent weaknesses. One of the biggest challenges when using k-means is to determine the optimal number of clusters, k. Although many approaches have been suggested in the literature, this is still considered as an unsolved problem. In this study, we propose a new technique to improve the gap statistic approach for selecting k. It has been tested on different datasets, on which it yields superior results compared to the original gap statistic. We expect our new method to also work well on other clustering algorithms where the number k is required. This is because our new approach, like the gap statistic, can work with any clustering method.

Keywords: Clustering, Number of Clusters, Data Mining.

1. Introduction

There are still many open challenges in the clustering task. Those challenges are getting even worse in the current big data era, where data is collected from many sources at high speed. This paper focuses on answering the question: how to decide on the number of clusters k? Being one of the oldest question in the clustering literature, the question has been tackled by hundreds of researchers with many solutions that have been proposed. Among these solutions, the gap statistic is one of the most modern approaches. It is backed by the rigorous theoretical foundation and has been shown to outperform many other heuristic-based approaches such as elbow or silhouette. However, there are still several drawbacks to the original design of the gap statistic, which limits its applicability in real applications. This paper introduces a new technique to mitigate those limitations. The technique can improve the effectiveness of the gap statistic in multiple

dimensions. The gap statistic that uses the newly proposed technique is called the “new gap” for short. The following few subsections describe literature reviews

The Elbow Approach

The oldest method called ‘elbow’ has been proposed to determine the number of clusters for k-mean clustering algorithm [6]. This is a visual method. The idea of the elbow method is to run clustering method on the dataset for a range of values of *k* (for example from 1 to 10), and for each value of *k* calculate clusters and internal index (it could be the sum of squared error (SSE), the percentage of variance, etc.). Then plot a line chart of the internal index for each value of *k*. At some value of k the value of internal index drops dramatically, and after that, it reaches a plateau when k is increased further. This is the best *k* value we can expect. Fig 1 illustrates how the elbow method work. In Fig 1, the line chart goes down rapidly with k increasing from 1 to 2, and from 2 to 3, and reaches an elbow at $k = 3$. After that, it decreases very slowly. Looking at the chart, it looks like maybe the right number of cluster is three because that is the elbow of this curve.

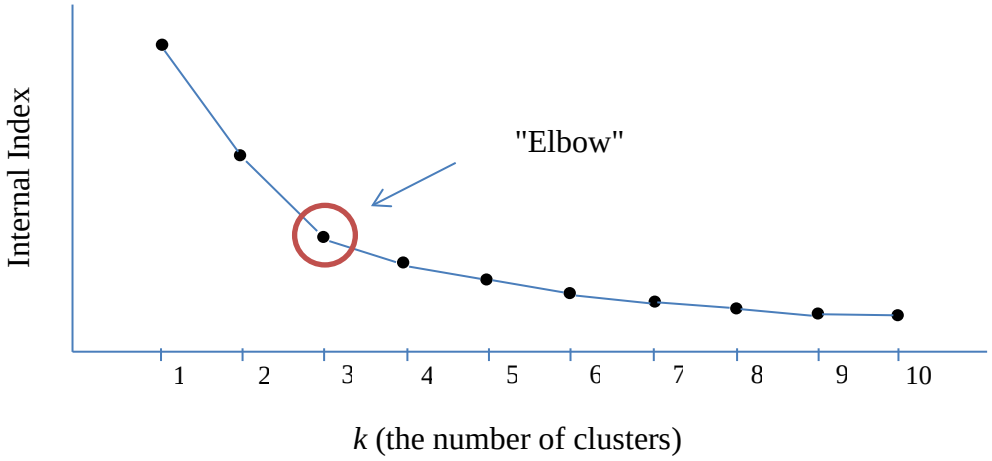


Fig. 1. Identification of Elbow point

However, the elbow method does not always work well. Sometimes, there are more than one elbow, or no elbow at all.

Average Silhouette Approach

Average silhouette method computes the average silhouette of observations for different values of *k* [2][3]. The optimal number of clusters *k* is the one that maximizes the average silhouette over a range of possible values for *k* [7]. Given a clustering result with *k* clusters ($k > 1$), we can estimate how well an observation *i* is clustered by calculating its silhouette statistic $s^k(i)$. Let $a(i)$ be the average distance from observation *i*

to other points in its cluster, and $b(i)$ be the average distance from observation i to points in its the nearest cluster, then the silhouette statistic $s^k(i)$ is calculated by:

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

A point is well clustered if $s^k(i)$ is large. The average silhouette score $avgS(k)$ gives an estimation of the overall clustering quality when clustering the dataset into k clusters:

$$avgS(k) = \frac{1}{n}\sum_{i=1}^{n} s^k(i)$$

, where n is the number of data points.

Therefore, we select k so that it maximizes the average silhouette score. However, this average silhouette is only a heuristic metric, which can be shown to perform poorly in many cases. Note that avgS(k) is not defined at $\underline{k} = 1$.

Hartigan Statistic

Hartigan (1975) proposed the statistic [1]:

$$H(k) = \frac{\frac{W_k}{W_{k+1}} - 1}{n - k - 1}$$

, where W_k is the average within-cluster sum of squares around the cluster means. The formula to calculate W_k is given in the next section about the gap statistic.

The idea is to start with $k = 1$ and keep adding a cluster until $H(k)$ is sufficiently large. Hartigan suggested the "sufficiently large" cut-off is 10. Hence the estimated number of clusters is the smallest $k \geq 1$ such that $H(k) \leq 10$.

Gap Statistic

Gap statistic was introduced in 2001 by Tibshirani et.al. [4] and is still a state-of-the-art method for estimating k. It has been shown to outperform the elbow, average silhouette, and Hartigan methods in both synthesized and real datasets [4][5]. The method works by assuming a null reference distribution. It then compares the change in within-cluster dispersion with the expected change if the null distribution is true. If when $k = K$ and the within-cluster dispersion starts decreasing slower than the expected rate of the reference distribution, the gap statistic returns k as the expected number of clusters. The formal definition of the gap statistic is given as follows:

Let $d_{ij} = \|x_i - x_j\|^2$ denotes the Euclidean distance between observation i and j, D_r is the sum of the pairwise distance for all points in a given cluster C_r containing n_r points.

$$D_r = \sum_{i \in C_r} \sum_{j \in C_r} d_{ij}$$

Then measure of compactness of clusters W_k is the average within – cluster sum of squares around the cluster means:

$$W_k = \sum_{r=1}^{k} \frac{1}{2\, n_r} D_r$$

The purpose of clustering is with a given *K* finding the optimal W_k, when *k* increases, W_k decreases. But the speed reduction of W_k also decreases. The idea of elbow method is to choose the *k* corresponding to the "elbow" (finding k that point has the most significant increase in goodness-of-fit). The problems when using elbow method is no reference clustering to compare, and the differences $W_k - W_{k-1}$'s are not normalized for comparison.

The main idea of the gap statistic is to standardize the graph of $\log(W_k)$ by comparing it with its expectation under an appropriate null reference distribution of the data. Estimate of the optimal number of clusters is then the value of k for which $\log(W_k^{data})$ falls the farthest below this reference curve $\log\left(W_k^{null}\right)$:

$$Gap_n(k) = E_n^*\{\log\left(W_k^{null}\right) - \log(W_k^{data})\}$$

With E_n^* is the expectation under a sample size of n from reference distribution, we estimate $E_n^*\{\log\left(W_k^{null}\right)\}$ by an average of B copies $\log(W_k^{null})$, each of which is computed from a Monte Carlo sample from reference distribution. Cluster the Monte Carlo samples into *k* groups and compute $logW_{kb}$, $b = 1, 2 \ldots, B$, $k = 1, 2 \ldots, K$. Compute the (estimated) gap statistic:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} logW_{kb}^{null} - \log(W_k^{data})$$

Those $logW_{kb}^{null}$ from the B Monte Carlo replicates exhibit a standard error $sd(k)$ which, accounting for the simulation error, is turned into the quantity

$$s_k = \sqrt{1 + \frac{1}{B}}.sd(k)$$

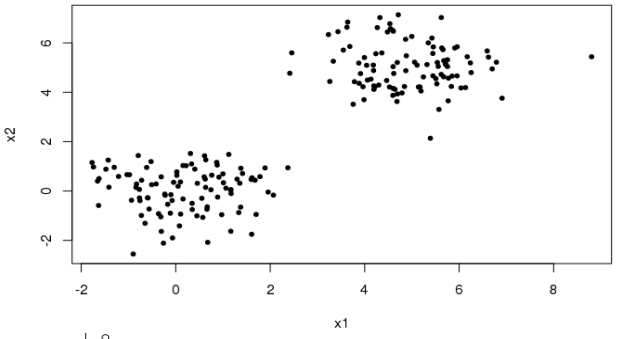
Finally, the optimal number of cluster *K* is the smallest *k* such that

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

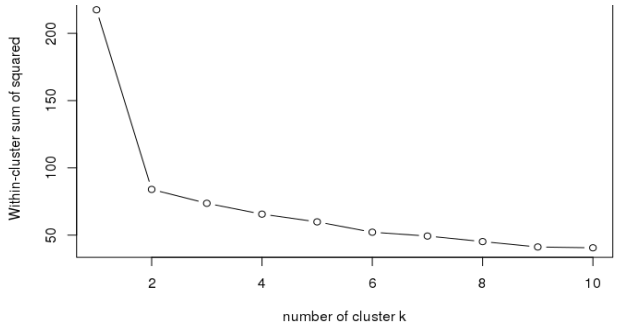
The above rule to select k is presented in the original gap statistic paper and called the "Tibs2001SEmax" rule in the R clustering implementation of the gap statistic. Since 2001, several other alternatives to this rule have been proposed, such as the "firstSEmax" rule [8] or the "globalSEmax" rule [9]. In this study, the Tibs2001SEmax rule in all experiments was used as the baseline approach. In this paper, the term "gap statistic" refers to the function Gap(*k*) with the Tibs2001SEmax is used as the *k*-selecting rule.

Fig. 2 provides an example of how the gap statistic works. Fig. 2a plots the example dataset with two well-separated clusters. Fig. 2b shows the line representing the within sum of squares W_k^{data}, which is a downward trend in number of cluster *k*. Fig. 2c shows the log of the expected rate $\log(W_{kb}^{null})$ using an assumed null distribution (uniform distribution in this case). Fig. 2d shows the gap statistic, which is calculated by subtracting the log expected rate $\log(W_k^{null})$ for the $\log(W_{kb}^{data})$. The optimal number of *k* is the smallest *k* such that there is a significant chance that *Gap*(*k*) is higher than *Gap*(*k*+1), which is $k = 2$ in this case. Tibshirani used one standard deviation s_{k+1} to determine when the chance is significant.

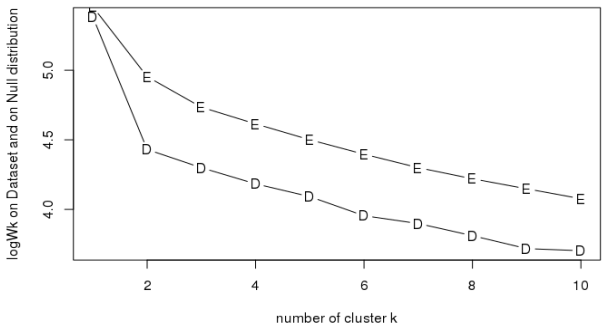
a)


x2
x1
-2
0
2
4
6
8

b)


Within-cluster sum of squared
number of cluster k
50
100
150
200
2
4
6
8
10

c)


logWk on Dataset and on Null distribution
number of cluster k
4.0
4.5
5.0
2
4
6
8
10
D
E

d)

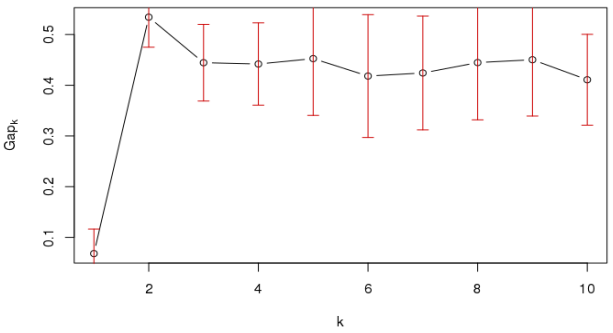

Gapk
k
0.1
0.2
0.3
0.4
0.5
2
4
6
8
10

Fig. 2. How the gap statistic works on a dataset with two well-separated clusters

2. Methodology

Although being backed by a rigorous theoretical foundation (unlike other heuristic-based methods like elbow or silhouette), the Gap statistic still has several drawbacks that limit its applicability to practical applications. In this section, we conduct several experiments with synthesized datasets to demonstrate those limitations. Based on the insights learned from those experiments, we then introduced a new technique to improve the gap statistic.

2.1 The Gap Statistic Limitations

By design, the gap statistics can only work well when all the clusters in the dataset are well-separated from each other. However, this is rarely the case in practice, where clusters usually overlap up to a certain degree. This "non-overlapping" assumption is one of the main reason that limits the gap statistics effectiveness in real applications. Fig. 3 shows how the gap statistics fail to identify the correct K in simple synthesized datasets, that the clusters only barely overlap each other.

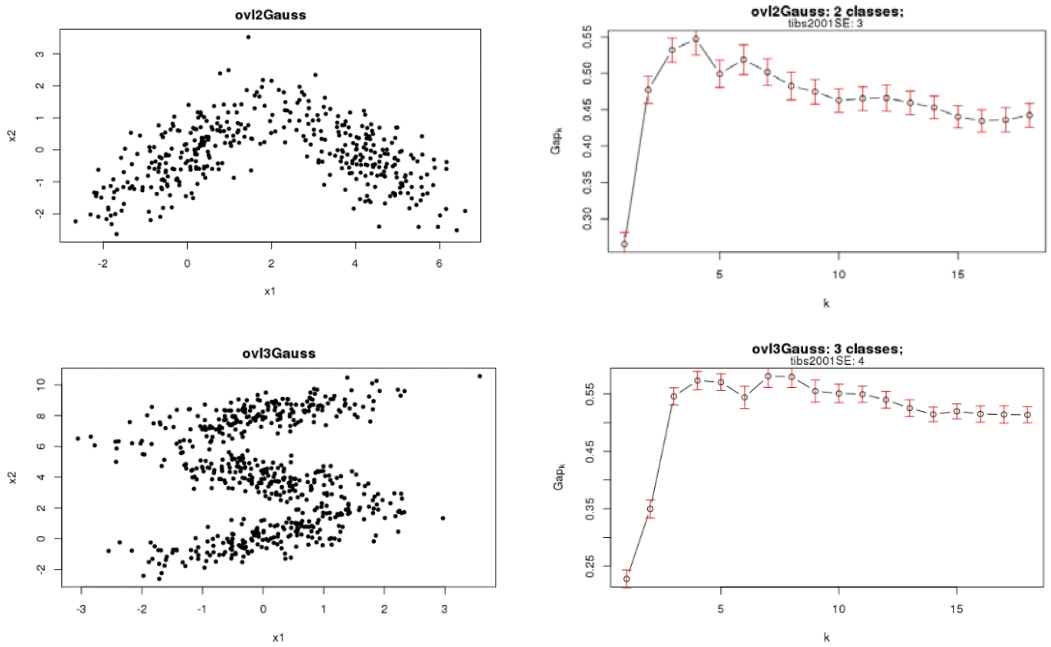


Fig. 3. Overlapping clusters problem with gap statistic

(a) the ovl2Gauss dataset: 400 data points in 2 dimensions that sampled equally from the two 2D Gaussian distributions: $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$ and $\mathcal{N}\left(\begin{bmatrix}4\\0\end{bmatrix}, \begin{bmatrix}1 & -0.7\\-0.7 & 1\end{bmatrix}\right)$.
(b) gap statistic with Tibs2001SE rule suggests k = 3 instead of 2 for the ovl2Gauss.

(c) the ovl3Gauss dataset: 600 data points in 2 dimensions that sampled equally from the three Gaussian distributions: $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$, $\mathcal{N}\left(\begin{bmatrix}0\\8\end{bmatrix}, \begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right)$, and $\mathcal{N}\left(\begin{bmatrix}0\\4\end{bmatrix}, \begin{bmatrix}1 & -0.7\\-0.7 & 1\end{bmatrix}\right)$
(d) gap statistic with Tibs2001Se rule suggests k = 4 instead of 3 for the ovl3Gauss.

However, clusters should not overlap with each other too much. Otherwise, the notion of "cluster" will become very fuzzy. This is because the data density in the overlapping area is the sum of the data density of the two clusters in that area. This can potentially make the overlapping area become another cluster. In some applications, we indeed want to recognize that overlapping space as a cluster, while that behavior is unexpected in other applications. Fig. 4 illustrates this confusion in the case of two strongly overlapping clusters.

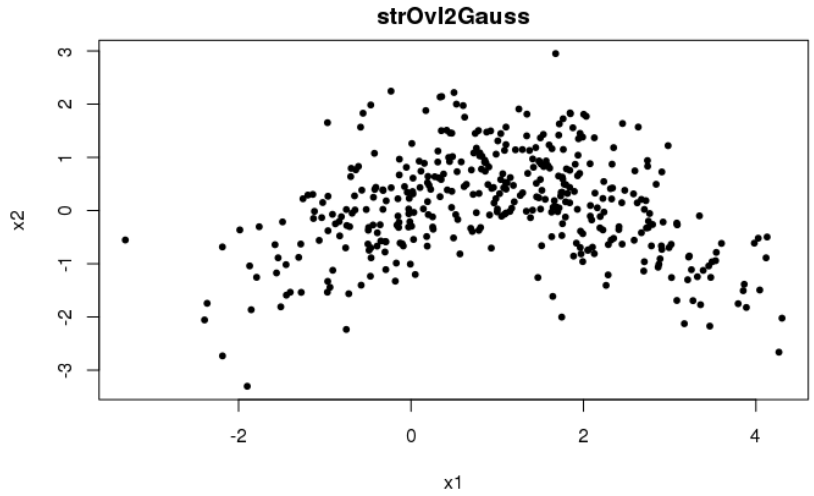


Fig.4. Two strongly overlapping clusters can be correctly seen as one, two, or 3 clusters.

Besides the non-overlapping assumption, the gap statistic also assumes that there is no hierarchical clustering structure in the dataset. This means in the dataset; there is no cluster that consists of many smaller clusters. In addition, the gap statistics require a lot of computing power to compute the expected W_k under the null reference distribution $E_n^*\{\log(W_k)\}$. It has to sample the null reference distribution B times ($B \geq 50$), for each sample b, we run the clustering algorithm. In this case, the clustering algorithm is PAM, which takes $O(n^2)$ with n is the number of data points. In total, the complexity of the algorithm to estimate the $E_n^*\{\log(W_k)\}$ is $O(Bn^2)$. This would make it impossible to apply gap statistic on dataset with more than several thousands of data points.

2.2 The new gap

As described in the previous section, the gap statistic method has largely three limitations. However, we only focus on the overlapping issue to produce a new gap. The other limitation issues will be covered in the further research.

The 1stDaccSEmax Rule for Overlapping Clusters. The Tibs2001SEmax rule returns the smallest k such that the gap at that point has a significant chance (one standard error) to be higher than the next gap. As shown in the previous section, this rule is very sensitive to overlapped clusters. In fact, when there are overlapping clusters in the dataset, the gap does not decrease but slightly increase after $k = K$ (where K is the real number of clusters in the dataset). This results in over-estimation of K.

Therefore, instead of using the gap statistic directly, we propose to use the deceleration of the gap statistic (Dacc statistic for short). The Dacc is calculated as follows:

$$Dacc(k) = [Gap(k) - Gap(k-1)] - [Gap(k+1) - Gap(k)]$$
$$= 2Gap(k) - Gap(k-1) - Gap(k+1)$$

Fig. 5 shows how the $Dacc(k)$ statistic can be computed from the $Gap(k)$ statistic.

We designed this statistic based on the insight that when *k* is going from 1 to *K*, the *Gap*(*k*) increases with constant or accelerated speed, up to the point where *k* = *K*. At that point, the *Gap*(*k*) will suddenly slow down its speed of increasing or start to decreasing (negative speed). Fig. 6 illustrates how the *Dacc*(*k*) looks like in different scenarios.

$$Dacc(k) = [Gap(k) - Gap(k-1)] - [Gap(k+1) - Gap(k)]$$

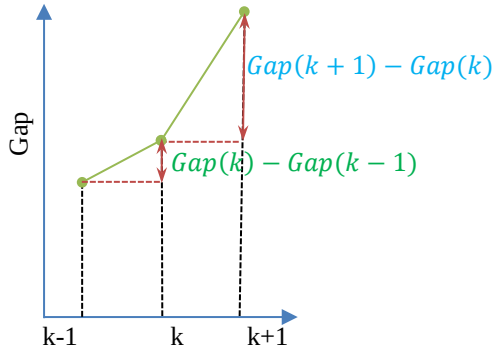


Fig.5. How to compute Dacc(*k*) from Gap(*k*)

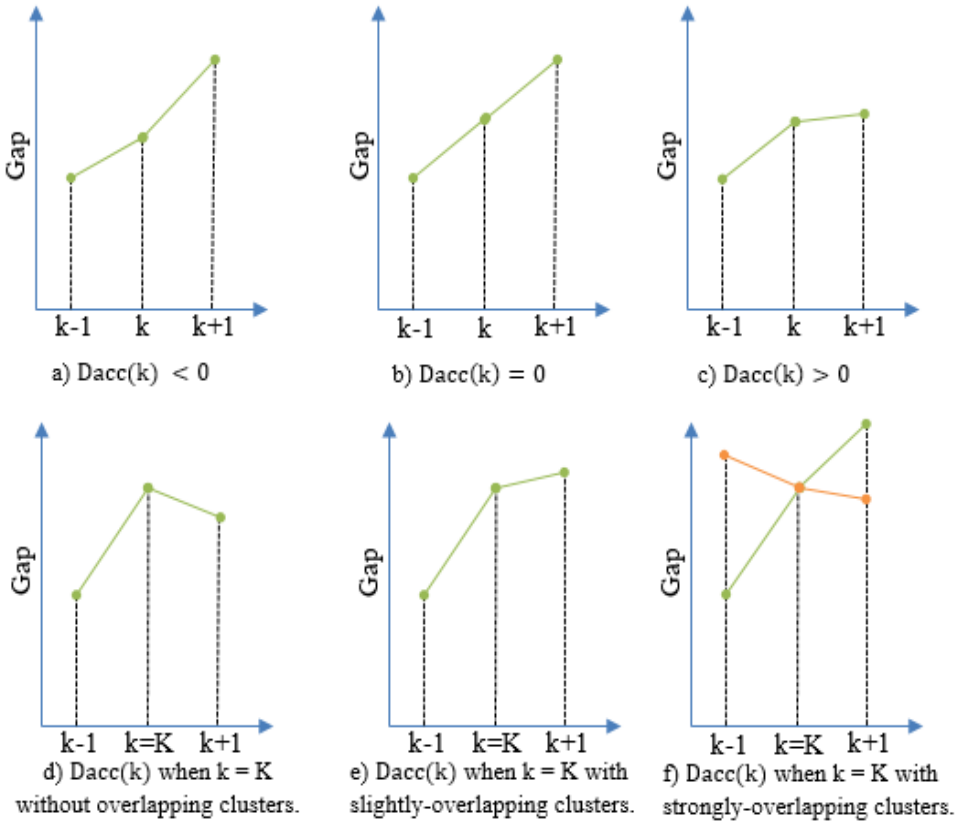


Fig.6. The Dacc(k) value in different scenarios;

Fig. 6a-c) Different cases where $Dacc(k) < 0$; $Dacc(k) = 0$ and $Dacc(k) > 0$

Fig. 6d) In dataset with K non-overlapping clusters: *Gap*(k) increases when $k < K$, reaches its first local maxima at $k = K$, and starts decreasing when $k = K + 1$. Therefore, $k = K$ is also the first local maxima of *Dacc*(k)

Fig. 6c) In dataset with K clusters where some clusters slightly overlap each other: *Gap*(k) still increases from k = K to $k = K$+1, making *Gap*($k = K$) no longer the first local maxima. However, since the overlapping area is small (slightly-overlapping assumption), the increasing speed from *Gap*(K) to *Gap*(K + 1) is significantly smaller than the increasing speed from *Gap*(K - 1) to *Gap*(K), making the Dacc statistic still maximize at $k = K$. Therefore, the *Dacc*(k) is more robust than the *Gap*(k) in a dataset with slightly-overlapping clusters.

Fig. 6f) In dataset with K clusters where some clusters strongly overlap each other: the definition between clusters becomes very fuzzy. Two strongly overlapping clusters can

be correctly considered as one, two, or three clusters. Therefore, both Dacc and Gap statistic behave unpredictably in this case.

To take into account the sampling error occurring when estimating the expected W_k under the null distribution, I incorporate the standard error s_k to the $Dacc(k)$ to get the $DaccSE(k)$ as follows:

$$DaccSE(k) = [(Gap(k) - 0.5s_k) - (Gap(k-1) + 0.5s_{k-1})] - [(Gap(k+1) + 0.5s_{k+1}) - (Gap(k) - 0.5s_k)]$$

$$DaccSE(k) = 2Gap(k) - Gap(k-1) - Gap(k+1) - 0.5s_{k-1} - 0.5s_{k+1} - s_k$$

As we can see, the higher the sampling errors at k - 1, k, or k + 1, the more DaccSE penalizes the Dacc estimation. Note that I used half standard error in the *DaccSE*(k) formula. We can choose to use different factor for the standard error based on how "aggressive" or "conservative" you want the DaccSE to behave. Fig. 7 illustrates how the DaccSE(k) is calculated. While the Dacc is calculated based on the green line, the DaccSE is calculated based on the dashed orange line. The DaccSE penalizes the *Gap*(k - 1), *Gap*(k) and *Gap*(k + 1) estimation according to how big the s_{k-1}, s_k, and s_{k+1} are.

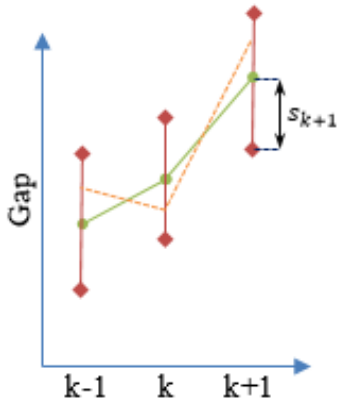


Fig. 7. How the DaccSE(k) is derived from *Gap*(k) and s_k.

The *Gap*(k) chart can have multiple peaks, especially when the dataset has a hierarchical clustering structure. Therefore, instead of selecting k where the *DaccSE*(k) reaches its global maxima, we select the k where *DaccSE*(k) reaches its first local maxima. This is similar to the idea of searching for the first local maxima of the Tibs2001SEmax rule introduced in the original gap paper. This new rule is called the 1stDaccSEmax rule. Generally, the 1stDaccSEmax rule keeps looking for k with the highest positive DaccSE, with k sequentially running from k = 2 to k = kmax and stop at the point where $Gap(k)$ higher than $Gap(k) - s_{k+1}$. Fig. 8 shows how the 1stDaccSEmax rule works in different situations.

Note that although the *DaccSE*(k) statistic does not define when k = 1, the 1stDaccSEmax rule can still detect if there is no cluster in the dataset. This can happen in two situations, which are illustrated in Fig. 8. In Fig. 8b, Gap(1) > Gap(2) by a margin

bigger than s_2. Therefore, we stop looking for *k* right from the beginning and return *k* = 1 right away. In Figure 8c, all the DaccSE is negative (there is no *k* at which the gap decreases). Therefore, we also return *k* = 1 in this case.

Fig. 9 shows the effectiveness of the 1stDaccSEmax rule on synthesized datasets with overlapping clusters

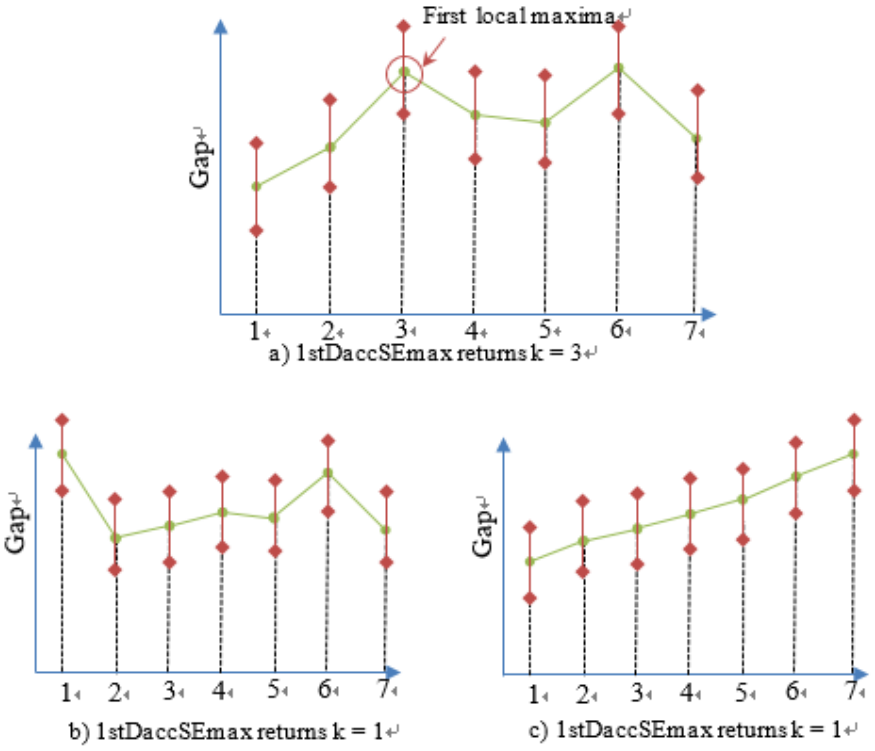


Fig.8. How the 1stDaccSEmax works in different kinds of Gap charts

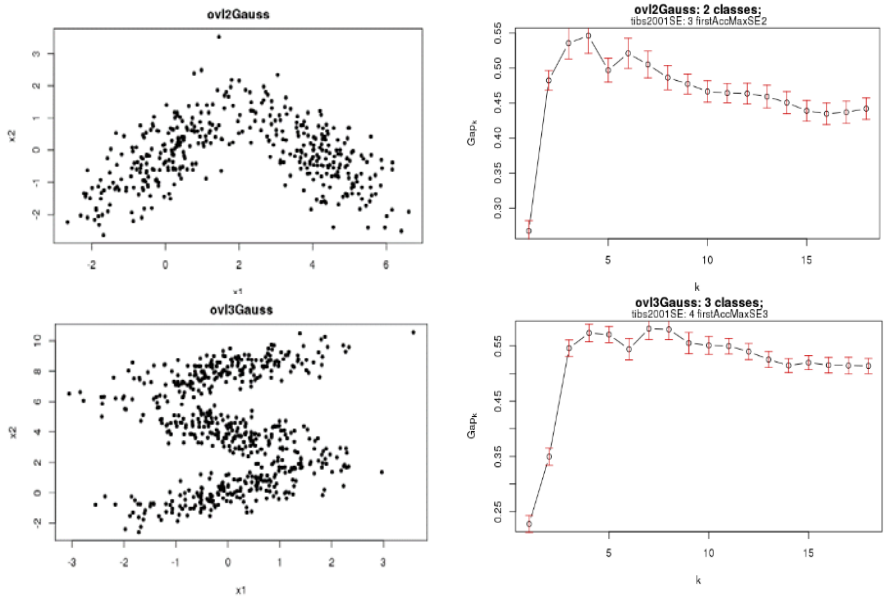


Fig.9. Apply the 1stDaccSEmax rule on synthesized overlapping clusters datasets.
Fig. 9a) The ovl2Gauss dataset
Fig. 9b) Tibs2001SEmax suggests $k = 3$ because $Gap(k)$ still increases from Gap(2) to Gap(3) due to the overlapping. The 1stDaccSEmax predicts correctly that $k = 2$, because the decrease at $k = 3$ is smaller than the decrease at $k = 2$.
Fig. 9c) The ovl3Gauss dataset
Fig. 9d) The Tibs2001SEmax predicts wrongly that $k = 4$ due to the overlapping issue. The 1stDaccSEmax rule correctly predicts that $k = 3$.

3. Conclusion

This study focuses on improving the gap statistic for the task of predicting the number of clusters k of a dataset. It identifies and demonstrates three main limitations of the gap statistic, including the overlapping clusters problem, the hierarchical clustering structure problem, and the big dataset problem. Based on these insights, we proposed the new technique to tackle the overlapping problem: the 1stDaccSEmax rule. The performance of the new method is evaluated with several synthetic datasets. It is believed that the performance of the new gap method would be shown to be better than all other traditional approaches. The further numerical experiments will be done on several real datasets with some other new techniques to overcome the other gap limitations.

ACKNOWLEDGMENT

- This work was supported by the Industry-Academic Cooperation R&D grant (Dec. 24 2018 ~ Dec. 23 2019) funded by the LX Spatial Information Research Institute (SIRI).
- This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2017R1D1A1B03034475).

References

1. Hartigan John A. (1975) Clustering algorithms. New York Wiley - Wiley series in probability and mathematical statistics xiii, 351 p.
2. Wang, Fei, Franco-Penya, Hector-Hugo, Kelleher, J., Pugh, J. & Ross, R. (2017). An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity. 10.1007/978-3-319-62416-7_21.
3. Rousseeuw J. P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20 (1987) 53-65
4. Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. Journal of the Royal Statistical Society B, 63, 411–423.
5. Chiang, M., Mirkin, B. (2010). Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. Journal of Classification. 27. 3-40. 10.1007/s00357-010-9049-5.
6. Trupti M. Kodinariya, Dr. Prashant R. Makwana (2013), Review on determining number of Cluster in k-means Clustering. International Journal of Advance Research in Computer Science and Management Studies. Volume 1, Issue 6, November 2013 pg. 90-95
7. Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
8. Maechler, M. (2012), firstSEMax rule in R document. [Online]. Available: https://www.rdocumentation.org/collaborators/name/Martin%20Maechler
9. Dudoit, S. and Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset, Genome Biology 3, Article number: research0036.1