Knee Point Detection in BIC for Detecting the Number of Clusters

Qinpei Zhao, Ville Hautamaki, Pasi Fränti

Department of Computer Science, University of Joensuu
Box 111, Fin-80101 Joensuu
FINLAND
{zhao, villeh, franti }@cs.joensuu.fi

Abstract. Bayesian Information Criterion (BIC) is a promising method for detecting the number of clusters. It is often used in model-based clustering, in which a decisive first local maximum is detected as the number of clusters. In this paper, we re-formulate the BIC in partitioning based clustering algorithm, and propose a new knee point finding method based on it. Experimental results show that the proposed method detects the correct number of clusters more robustly and accurately than the original BIC, and performs well in comparison to several other cluster validity indices.

1. Introduction

Cluster analysis is to group a collection of patterns, which is usually represented as a vector of measurements, or a point in a multidimensional space, into clusters according to a clustering similarity function or a clustering validity index. The output of clustering over the same dataset could be very different if the input parameters for clustering vary. This is due to the fact that variation of clustering parameters has changed the behaviour and the execution of clustering substantially. An essential input parameter for clustering is the number of clusters that best fits a given dataset. Thus, a common question prior arising before clustering is how many clusters are present in a given set of objects. Moreover, most clustering algorithms face several common issues in execution of clustering: if different partitions are obtained for a given dataset, then amongst the resulting partitions, which one is the most suitable or optimal one.

A number of measures have been well developed for this problem in literature [1-12]. In general, they can be categorized into three types: *external criteria*, *internal criteria* and *relative criteria*. An external criterion evaluates the result of clustering based on a pre-specified structure. Meanwhile an internal criterion is based on quantities that involve the vectors of the data set themselves. The idea behind the third one is to evaluate a clustering structure by comparing it to other clustering results, obtained by the same algorithm but with different number of clusters. The basis of external and internal criteria is statistical testing and the hypothesis can lead to computationally complex procedure. The relative criterion, On the other hand, does not involve statistical tests, and is used more often.

Milligan and Cooper [1] have provided a comparison of thirty validity indices for data sets by using only hierarchical clustering algorithms. Dimitriadou et al [2] presents another comparison of fifteen validity indices for binary data sets. Based on a typical definition of clusters, where the points within the same cluster are close to each other, while the clusters themselves are far from each other, several measures have been proposed. Calinski and Harabasz [3] proposed the F-statistic method, which takes advantage of within-cluster variance and between-cluster variance. Dunn's index [4] considered both the diameter of each cluster and the distance between clusters. As the diameter will be severely affected by noise, the Dunn's index may not perform very well as a cluster validity index. This issue has been addressed in [5]. Davies-Bouldin [6] is another well known index, which is based on the idea that for a good partition inter cluster separation as well as intra cluster homogeneity and compactness should be high.

Because different kind of clustering algorithms often have different properties, different types of measures based on specific clustering algorithms have been proposed. For example, Xie-Beni index [7] was originally proposed to identify separation for fuzzy c-partitions. It depends on the data set, geometric distance measure, distance between cluster centroids, and more importantly on the fuzzy partition generated by any fuzzy algorithm. When dealing with model-based clustering, Banfield and Raftery used a heuristically derived approximation to twice the log Bayes factor [9] called the "AWE" to determine the number of clusters in hierarchical clustering based on the classification likelihood. When EM is used to find the maximum mixture likelihood, a more reliable approximation to AWE called *Bayesian Information Criterion* (BIC) [8] is applicable. A new K-means based algorithm incorporating model selection was proposed in [10]. This so-called X-means algorithm uses BIC to make local decisions that maximize the posterior probabilities of the model under the assumption that the models are spherical Gaussians. Because of the effectiveness of BIC in model-based clustering, we re-formulate BIC to determine the number of clusters in partitioning based clustering.

Some of the indices can be easily used to determine the number of clusters by finding the minimal or maximal value, but several of them cannot. A criterion with within-group sum-of-squares objective function trace (W) was proposed by Krzanowski [11], in which the plot of index value against number of clusters was monotonically decreasing. They considered using the successive difference of the function to find the optimum value. Yet, in the visual "number of clusters vs. criterion metric" graph there often is a clear knee point (or jump point) that can be used to detect the number of clusters, see Fig.1. In principle, the problem of finding the knee point can be attacked by successive difference method. But the successive difference method only considers some adjacent points and local trend of the graph which may lead uncorrect results. We therefore propose to measure the knee point based on the angles of the local significant changes in the successive difference results, and demonstrate that by this method, the performance of the BIC method can be improved.

The rest of the paper is organized as follows. The problem formulation is given in Section 2.1. The BIC method in partitioning based clustering is renewed in Section

2.2, and the angle-based method is introduced in Section 2.3. The proposed method is compared to several existing methods in Section 3. The results demonstrate that the proposed knee point finding method improves the original BIC method, which takes the first local minimum as the number of clusters, and outperforms most of the existing method on the data sets tested. Conclusions are drawn in Section 4.

2. Proposed Method

We proposed a knee point finding method for BIC in partitioning based clustering, which is called *angle-based method*. The next section describes the proposed method.

2.1 Preliminary

The problem of determining number of clusters is defined here as follows:

Given a fixed number of clusters $m \geq 2$, and a specific clustering algorithm, find the clustering that best fits for the data set with different parameters. The procedure of identifying the best clustering scheme involves the following parts:

- Select a proper cluster validity index.
- Repeat a clustering algorithm successively for number of clusters, m from a predefined minimum to a predefined maximum.
- Plot the "number of clusters vs. criterion metric" graph and select the m at which the partition appears to be "best" in terms of at which the criterion is optimized.

Based on this procedure, one can identify the best clustering scheme. The problem remains that how to select the optimal m for the validity index. Mean square error (MSE), for example, exhibits a decreasing with respect to m increasing. Meanwhile, some indexes show the maximum or minimum in the curve. No matter what kind of case we have, there exists the significant local change in the curve, which is so-called knee or jump point.

Locating the knee point in the validity index curve is not well-studied. A straightforward approach is to take difference of successive index values, for example, calculating the difference between previous and current values of the index. Other method like L-method [12] is propsed to find the knee point of the curve by the boudary between the pair of straight lines that most closely fit the curve. For some indexes, the maximum or minimum value will be considered as the knee point. However, if there are several local maximum (minimum) values existing, the challenge is to decide which one is the most suitable one to indicate the information of the data sets. According to our study, BIC indicates a good prospect in determining the number of clusters in partitioning based clustering. To improve the accuracy of BIC, a good knee point finding method instead of taking the first local maximum is needed.

2.2 Bayesian Information Criterion (BIC)

The *Bayesian Information Criterion* (BIC) has been successfully applied to the problem of determining the number of components in model-based clustering by Banfield and Raftery. The problems of determining number of clusters and the clustering method are solved simultaneously.

We derive the formula of BIC based on Kass and Wasserman [13].

$$BIC = L(\theta) - \frac{1}{2} m \log n \tag{1}$$

where, $L(\theta)$ is the log-likelihood function according to each model, m is the number of clusters and n is the size of the data set. Under the identical spherical Gaussian assumption, the maximum likelihood estimate for the variance of the i^{th} cluster is:

$$\Sigma_i = \frac{1}{n_i - m} \sum_{j=1}^{n_i} \| x_j - C_i \|^2 \tag{2}$$

where n_i is the size of each cluster, x_j is the j^{th} point in the cluster and C_i is the i^{th} cluster. For m clusters, the sum of log-likelihood of each cluster is as follows.

$$L(\theta) = \sum_{i=1}^{m} L(\theta_i) \tag{3}$$

Define $pr(x_i)$ as the probability of the i^{th} point in data sets, and $C_{p(i)}$ is the cluster corresponding to the partitioning. The variable d is the dimension of the data sets. Then, log-likelihood of the i^{th} cluster can be derived as follows:

$$\begin{aligned} L(\theta_i) &= \log \prod_{i=1}^{n_i} pr(x_i) = \sum_{i=1}^{n_i} \log pr(x_i) \\ &= \sum_{i=1}^{n_i} \log\left(\frac{n_i}{n} \frac{1}{(2\pi)^{d/2} \Sigma^{1/2}} \exp\left(-\frac{\| x_i - C_{p(i)} \|^2}{2\Sigma_i}\right)\right) \\ &= \sum_{i=1}^{n_i} \left(\log \frac{n_i}{n} - \log((2\pi)^{d/2} \Sigma_i^{1/2}) - \frac{\| x_i - C_{p(i)} \|^2}{2\Sigma_i}\right) \\ &= n_i \log n_i - n_i \log n - \frac{n_i * d}{2} \log(2\pi) - \frac{n_i}{2} \log \Sigma_i - \frac{n_i - m}{2} \end{aligned} \tag{4}$$

To extend the log-likelihood of each cluster to all of the clusters, we use the fact that the log-likelihood of the points that belong to every clusters is the sum of the log-likelihood of the individual ones. So the total log-likelihood will be:

$$BIC = \sum_{i=1}^{m} \left(\log n_i - n_i \log n - \frac{n_i * d}{2} \log(2\pi) - \frac{n_i}{2} \log \Sigma_i - \frac{n_i - m}{2}\right) - \frac{1}{2} m \log n \tag{5}$$

We use this BIC formula globally for each number of clusters in a predefined range. In general, m should be as small as possible according to [8]. Their strategy for the number of clusters is that a decisive first local maximum indicates strong evidence for the model size. However, according to our experiments, a good knee point detection method would be better choice for deciding which local maximum has stronger evidence for the correct number of clusters.

2.3 Angle-based Method

Some existing validity indices indicate the structure of data sets very well and contribute a lot to the problem. However, we can not directly obtain the correct number of clusters from them when they decrease or increase monotonously or only have some significant local changes. In this case, the structure of the dataset can be revealed by using a good knee point detection method. One efficient way is to calculate the difference between previous and afterward index values. There will be peaks at the points with significant local changes in the difference curve. It's also possible to consider more points of the curve in successive difference.

$$\mathrm{DiffFun}(m) = \mathrm{F}(m\text{-}1) + \mathrm{F}(m+1) - 2*\mathrm{F}(m) \tag{6}$$

where *DiffFun* is the difference function, *F(m)* is the index value and *m* is the current number of clusters. It takes use of the previous, afterward and current values simultaneously. The disadvantage of the successive difference method is that it only considers several points instead of the whole curve, allowing the index to find only local changes without a global perspective. If there are several local changes, then it may give a wrong result.

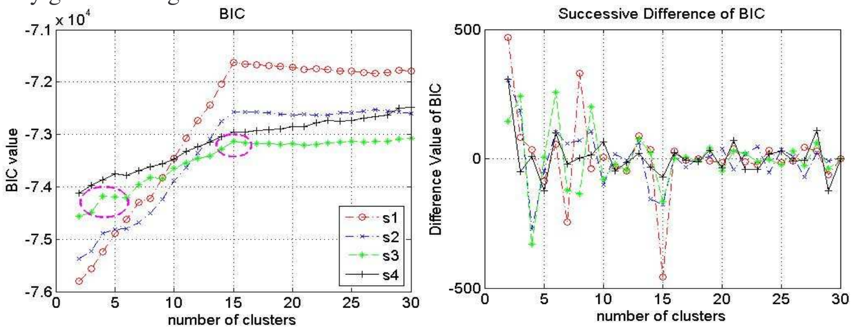

Fig. 1. Number of clusters vs. criterion metric graph of BIC (left), and its successive difference (right).

In Fig.1, the calculated BIC values are plotted using four data sets with different degrees of cluster overlapping. There are at least two obvious jumps in each curve. The first decisive local maximum is usually considered to be the number of clusters in the original BIC. The successive difference graph also gives strong support on this rule. The problem is the second local change (*m*=15) in the BIC curve also indicates strong evidence on the number of clusters in a global view. To decide which one is the optimal number, we take use of the angle property of a curve and propose an angle-based method to define locate the optimal local knee (jump) in a graph of BIC.

Given a function *F(m)* of BIC where m is in the range [*min*, *max*]. Calculate the successive difference in terms of formula (6) to get the function difference *DiffFun*. And detect n local significant changes by finding the first n minimum values in *DiffFun*. Here $n \le m/2-1$ because at least 2 points can generate 1 trough. Sort the local minimum values in decreasing way. Start from the point with bigger troughs; calculate the angle of them by (7).

$$\text{Angle} = \text{atan}(1/|F(m)-F(m-1)|)+\text{atan}(1/|F(m+1)-F(m)| \qquad (\mathbf{7})$$

```
Angle-based Method on Knee Point Finding Problem

Input:   Graph($m$)  ($m$[$min$, $max$])
Output:  Number of clusters $m$
Initialize:
   Current_Value = Graph($min$);
   Previous_Value = Graph($min$);
   After_Value = Graph($min$);
Begin:
for $m$ = $min$ to $max$
   Current_Value = Graph($m$);
   After_Value = Graph($m$+1);
   DiffFunc = Previous_Value + After_Value - 2*Current_Value;
   Previous_Value = Current_Value;
end
Find first $n$ local minimums in DiffFunc
LocalMin[$n$] = ($m$, Current_Value, Previous_Value, After_Value);
for each $n$ with decreasing order of LocalMin value,
   angle[$n$] = AngleCalc(Current_Value, Previous_Value, After_Value);
   Stop when the first maximum among the angles appears.
end
return $m$ with the first maximum angle;
```

Fig. 2. Pseudo-code of the angle-based method

It will stop when the first maximum angle appears, which indicates the trend of the curve globally, because it takes use of both the successive difference and angle property.

3. Experimental Results

We use here four two-dimensional artificially generated data sets denoted as s1 to s4 and one four-dimensional real data set Iris (Fig.3). The data sets s1 to s4 are generated with varying complexity in terms of spatial data distributions, which have 5000 vectors scattered around 15 predefined clusters with a varying degrees of overlap. Iris is obtained from the UCI Machine Learning Repository. It contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The data sets can be found here:

- s1-s4: cs.joensuu.fi/~isido/clustering/
- Iris: www.ics.uci.edu/~mlearn/MLRepository.html

As the measures have to be tested on a certain clustering algorithm, we run K-means and Randomize Local Search (RLS) [14] clustering with *m*=[2,30] in the case of s1-s4, and *m*=[2,10] in the case of Iris. To emphasize the effectiveness of the proposed method, we compare it with other measures:

- Dunn's index (DI) + maximum
- Davies-Bouldin's Index (DBI) + minimum
- Xie-Beni (XB) + minimum
- Bayesian Information Criterion (BIC) + first local maximum
- Angle-based BIC (ABIC).

Among them, DI, DBI and XB select the number of clusters either as the minimum or maximum value of the measure. We also report the results of the original BIC using the first local maximum as the number of clusters, and the proposed method using the angle-based method.

Table 1. Results using RLS (with 5000 RLS iterations and 2 K-means iterations).

Index	Data Set				
	s1	s2	s3	s4	Iris
DI	15	7	16	25	2
DBI	15	15	8	13	2
XB	15	15	4	13	2
BIC	15	4	4	5	3
ABIC	15	15	15	15	3

Table 2. Results using K-means (20 iterations)

Index	Data Set				
	s1	s2	s3	s4	Iris
DI	2	2	2	2	2
DBI	15	15	11	16	2
XB	15	15	4	13	2
BIC	15	4	4	5	3
ABIC	15	15	15	15	3

In Table 1 and Table 2, we list the number of clusters found by the different measures, data sets and clustering algorithms. Fig.4 and Fig.5 visualize the results for other four measures with RLS and K-means respectively. In Fig.6, we shows the result of each step in our method with data sets s4 and Iris.

- DI gives clear maximum for the easiest data set (s1) but fails with the more challenging ones. When K-means is applied with 20 iterations, it fails completely even with s1.

- DBI finds the correct minimum for s1 and s2, but the results for s3 and s4 indicate minimum somewhere around 10 and 15 and the detected minimum points are incorrect (8, 13 and 2).
- XB takes the minimum as the number of clusters, which is clearly visible in the case of the easiest data set (s1). Correct result is also found for s2, but again, the index fails with the more demanding sets (s3, s4 and Iris).
- The original BIC, which considers the first decisive local maximum as the number of cluster gets the correct number only for s1 and Iris.
- The proposed ABIC provides accurate results in all cases.

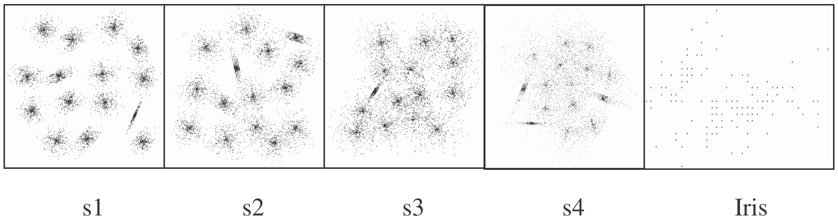

Fig. 3. The data sets used for testing: s1-s4 and Iris.

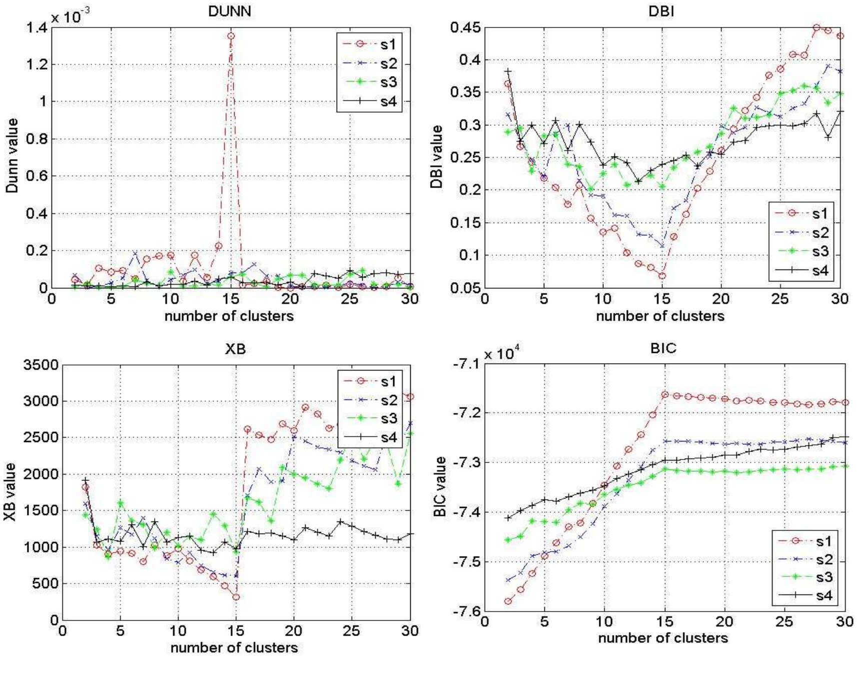

Fig. 4. Comparison of the other measures on s1 to s4 with RLS clustering algorithm.

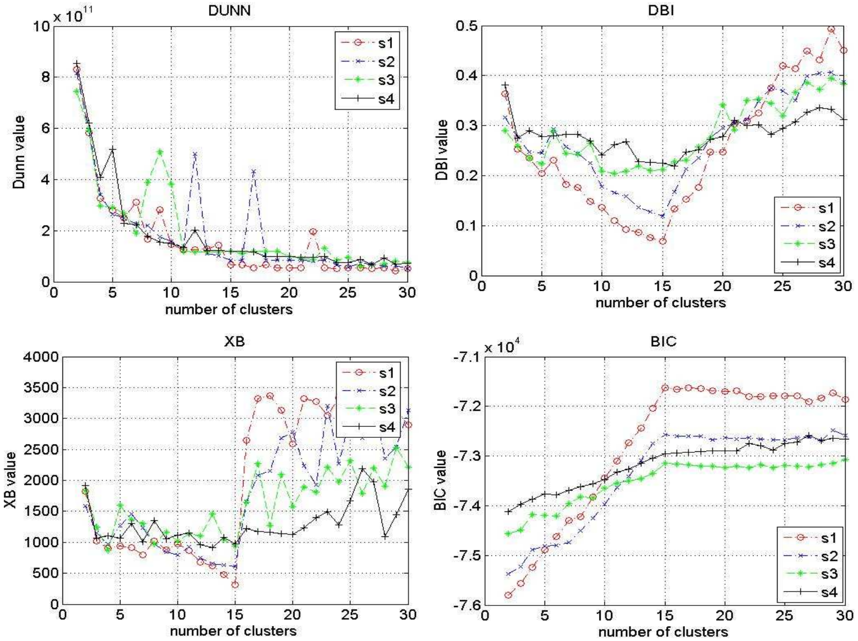

Fig. 5. Comparison of the other measures on s1 to s4 with K-means clustering algorithm.

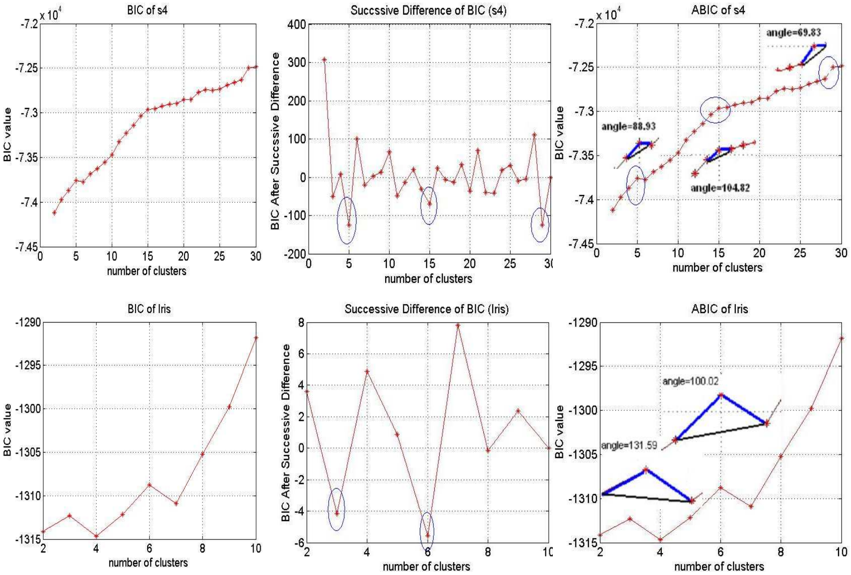

Fig. 6. Steps of the angle-based method on data sets s4 and Iris; BIC curve (left), the successive difference of BIC (middle) and the angles of the local significant changes (right).

4. Conclusions

We re-formulate BIC in partitioning based clustering, which shows good prospect for determining the number of clusters. The original method to decide the knee point of BIC is to take the first decisive local maximum, which is not accurate enough according to our experiments. To improve the BIC for getting more reliable results, an angle-based method for knee point finding of BIC is proposed in this paper. As the proposed method takes use of the global trend of the index curve, it's reliable to get the number of clusters. Experimental results also prove its effectiveness compared with other measures.

Reference:

1. G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, Vol.50, pp. 159-179, 1985.
2. E. Dimitriadou, S. Dolnicar, and A. Weingassel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, Vol.67, No.1, pp. 137-160, 2002.
3. T. Calinski, and J. Harabasz. A dendrite method for cluster analysis. *Communication in statistics*, Vol.3, pp. 1-27, 1974.
4. J.C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetica*, Vol.4, pp. 95-104, 1974.
5. J.C. Bezdek, N.R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, Vol. 28, No.3, pp. 301-315, 1998.
6. D.L. Davies and D.W. Bouldin. Cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.1, No.2, pp.95-104, 1979.
7. X.L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.13, No.8, pp. 841-847, 1991.
8. C. Frayley and A. Raftery. How many clusters? Which clustering method? answers via model-based cluster analysis. Technical Report no. 329, Department of Statistics, University of Washington, 1998.
9. R.E. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, Vol.90, No.430, pp. 773-795. 1995.
10. D.Pelleg, A.Moore: X-means: Extending K-means with efficient estimation of the number of clusters. *Proceeding of the 17th International Conference on Machine Learning*, pp.727-734, 2000.
11. W.J. Krzanowski, Y.T.Lai, A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, Vol.44, No.1 (Mar., 1988), pp.23-34.
12. S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Proceeding of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pp. 576-584, 2004.
13. R.E. Kass and L. Wasserman. A reference Bayesian test for nested Hypotheses and its relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, Vol. 90, No. 431, pp.928-934. 1995.
14. P. Fränti and J. Kivijärvi. Randomized local search algorithm for the clustering problem. *Pattern Analysis and Applications*, Vol.3, No.4, pp. 358-369, 2000.