2008 20th IEEE International Conference on Tools with Artificial Intelligence

Knee Point Detection on Bayesian Information Criterion

Qinpei Zhao, Mantao Xu, Pasi Fränti
Speech & Image Processing Unit
Department of Computer Science, University of Joensuu
Box 111, Fin-80101 Joensuu
FINLAND
{zhao, franti}@cs.joensuu.fi, mantao.xu@carestreamhealth.com

Abstract

The main challenge of cluster analysis is that the number of clusters or the number of model parameters is seldom known, and it must therefore be determined before clustering. Bayesian Information Criterion (BIC) often serves as a statistical criterion for model selection, which can also be used in solving model-based clustering problems, in particular for determining the number of clusters. Conventionally, a correct number of clusters can be identified as the first decisive local maximum of BIC; however, this is intractable due to the overtraining problem and inefficiency of clustering algorithms. To circumvent this limitation, we proposed a novel method for identifying the number of clusters by detecting the knee point of the resulting BIC curve instead. Experiments demonstrated that the proposed method is able to detect the correct number of clusters more robustly and accurately than the conventional approach.

1. Introduction

One of the main difficulties for cluster analysis is that, the correct number of clusters for 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. Obviously, the clustering validity criteria must be carefully defined not only according to a presumably known data distribution of clusters but also to the specification of the input datasets. More importantly, those clustering validity criteria serve as a tool to measure the goodness of groups in clustering as well as a principle for selecting the "best" number of clusters meanwhile. A number of efforts have been made in the previous literatures, e.g., Milligan and Cooper [1] presented a comparison study over thirty validity indices for hierarchical clustering algorithms whereas Dimitriadou et al [2] conducted their comparison study over fifteen validity indices for the case of binary data.

However, one class of clustering methods, model-based clustering, has received considerable attention recently, in a framework of the estimation of Bayesian likelihood or the estimation of Bayesian parameters, e.g. the well-known EM algorithm. The model-based clustering combines both the advantage of the optimal model parameter estimation in model selection and the advantage of selecting the most appropriate number of mixture components [3]. In particular the mixture model approach allows for an approximation of Bayes factor [4] even if clusters are in distinctively different models. Thanks to Banfield and Raftery's intuitive approximation of twice logarithm of Bayes factor, called "AWE", the number of clusters can be identified directly according to the classification likelihood. The approximation of Bayes factor can be extended to a more general principle, *Bayesian Information Criterion* (BIC) [5-8] for the sake of selecting an appropriate number of model parameters or the number of clusters.

In order to seek an optimal number of clusters particularly for a large-scale clustering problem, one could apply an intuitively heuristic approach instead of using an optimization algorithm. A remarkable example is that of Thorndike [9] who identified the optimal number of clusters such that a flattening of the clustering validity curve or a knee point can be observed. In contrast to finding the maximum or minimum of clustering validity index, the knee point detection algorithm is more practical because most of clustering validity indices are monotonically decreased or increased [10] with the number of clusters. Clearly, seeking a maximum or minimum is intractable. The monotony of clustering validity indices hinges on the fact that the likelihood of

1082-3409/08 $25.00 © 2008 IEEE
DOI 10.1109/ICTAI.2008.154

the training data is undesirably improved when the number of clusters is increasing, which mainly results in overtraining problem if the number of parameters is too large. Of course, one could apply the successive difference of the clustering validity index, to seek the optimal number of clusters. However, most of those heuristic decision approaches are highly subjective or heuristic. For instance, the first decisive local maximum of BIC can be viewed as a good number of clusters but the resulting number of clusters is often inaccurate due to inefficiency of clustering optimization procedures. To overcome these difficulties, we propose a simple knee point detection algorithm for BIC in automatic detection of the number of clusters. The knee point detection algorithm is quite intuitive and heuristic since the clustering validity curve monotonically decreases or increases after the knee point. For simplicity of determining the number of clusters, we re-formulate BIC in the framework of partitioning based clustering.

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 proposed method is introduced in Section 2.3. The experiments on the proposed method are presented in Section 3. The results on different kinds of datasets demonstrate that the proposed method improves the original BIC knee point detection algorithm. Conclusions are drawn in Section 4.

2. Proposed Method

2.1 Preliminary

The problem of determining the 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 the optimization on the criterion.

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 monotony with respect to the number of clusters, m, whereas some clustering validity indices may embrace a local maximum or local minimum in the curve. Regardless of the monotony of the underlying clustering validity curve, in most cases, a significant local change could be observed on the curve, which is the so-called *knee* or *jump point*.

Locating the knee point in the validity index curve has not been well-studied. A straightforward approach is to compute difference of successive index values, for example, calculating the difference between previous and current values of the index. Other method, such as L-method [11] has been proposed to find the knee point of the curve by the boundary between the pair of straight lines that most closely fit the curve in Hierarchical / segmentation clustering. For some indices, the local maximum or minimum value will be considered as the knee point. However, if there are several local maximal (minimal) values, the challenge is to decide which one is the most suitable one to indicate the information of the data sets. According to the experimental results in our study, BIC indicates a good estimation in determining the number of clusters in partitioning based clustering. To improve the accuracy of BIC, a good knee point detection method is needed instead of taking the first local maximum.

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 [12]. The problems of determining the number of clusters and the clustering problem are solved simultaneously.

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

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

where, $L(\theta)$ is the log-likelihood function of data θ 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:

$$\sum_i = \frac{1}{n_i - m} \sum_{j=1}^{n_i} \| x_j - C_i \|^2 \qquad (\mathbf{2})$$

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

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

Suppose that $pr(x_j)$ is the probability of the j^{th} data point in the data sets, and the variable d is the dimension of the data set. Then, log-likelihood of data belonging to the i^{th} cluster can be derived as follows:

$$L(\theta_i) = \log \prod_{j=1}^{n_i} (pr(C_i) \bullet pr(x_j)) = \sum_{j=1}^{n_i} \log(pr(C_i)\, pr(x_j))$$

$$= \sum_{j=1}^{n_i} \log(\frac{n_i}{n} \frac{1}{(2\pi)^{d/2} \sum_i^{1/2}} \exp(-\frac{\| x_j - C_i \|^2}{2\sum_i})) \quad (4)$$

$$= \sum_{j=1}^{n_i} (\log \frac{n_i}{n} - \log((2\pi)^{d/2} \sum_i^{1/2}) - \frac{\| x_j - C_i \|^2}{2\sum_i})$$

$$= n_i \log n_i - n_i \log n - \frac{n_i * d}{2} \log(2\pi) - \frac{n_i}{2} \log \sum_i - \frac{n_i - m}{2}$$

To extend the log-likelihood of each cluster to all of the clusters, the fact is applied that the log-likelihood of the whole data set is the sum of the log-likelihood of the individual cluster. Therefore the total log-likelihood will be:

$$BIC = \sum_{i=1}^{m} (n_i \log \frac{n_i}{n} - \frac{n_i * d}{2} \log(2\pi) - \frac{n_i}{2} \log \sum_i - \frac{n_i - m}{2}) - \frac{1}{2} m \log n \quad (5)$$

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 [5]. 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 a better choice for deciding which local maximum has the strongest evidence for the correct number of clusters.

2.3 Knee Point Detection of BIC

In this section, we analyze the drawback of knee point detection on BIC by using the first decisive local maximum as the number of clusters. Successive difference on BIC is also analyzed. We then propose our knee point detection method on BIC called *DiffBIC* method in partitioning based clustering.

2.3.1 Existing Methods. There is a slight option difference on how to find the optimal value of BIC for cluster validity except the first decisive local maximum. However, our experimental findings indicate several local maximums in the BIC curve (see Fig.1) due to the fact that the clustering performance is highly subjective to the initial clustering guess or partition. Hence, the resulting first decisive local maximum could often be the local maximum approaching or very close to the initial guess. This can be observed in the BIC curve for dataset s3 in Fig.1: the first decisive local maximum is achievable at m=4 albeit the right number of clusters m is 15 where there is a more significant change of BIC (not a conventional knee point). The difference values of BIC for dataset s3 and s4 also reveal that detection of knee point for BIC may be faced with the same challenge as the first local decisive maximum. A more objective method of detecting the knee point of BIC curve is therefore demanded.

Several alternative techniques on knee point detection methods have been proposed in the literature. Successive difference of two adjacent points is one possible way and it can be calculated as: SD(n) = BIC(n-1)+BIC(n+1)-2*BIC(n); where n is the current point. However, it can locate the knee point only locally as it considers only several successive points in the curve as shown in Fig.1. According to the figure, we can find the highest differences for each dataset with successive difference at the points m_{opt}(s1)=15, m_{opt}(s2)=15, m_{opt}(s3)=4 and m_{opt}(s4)=5. The detected points offer the most significant changes of BIC but without taking into account of BIC value itself. Eventually, this method is not always reliable in particular when a local maximum close to the initial guess can be quickly obtained by clustering algorithms.

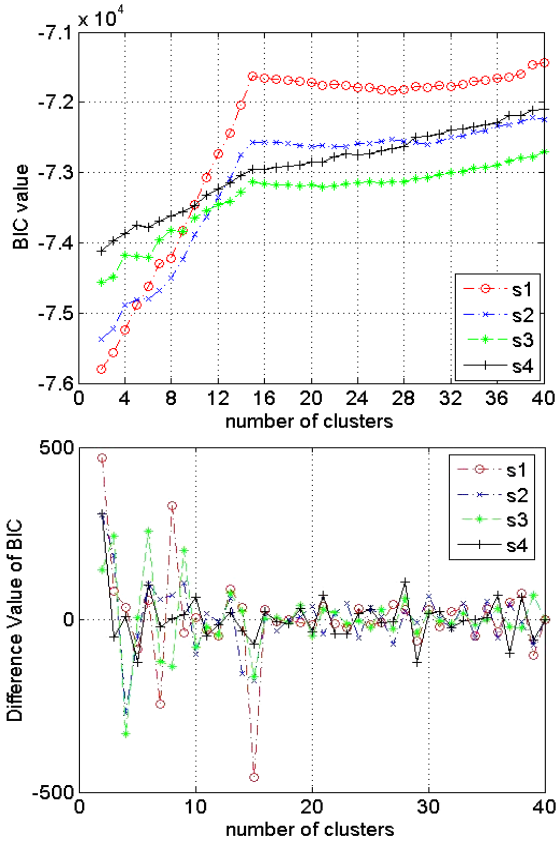


Figure 1. The original BIC curve (up) for datasets s1 to s4 obtained by RLS clustering algorithm [14] and successive difference of BIC (down).

2.3.2 DiffBic Function. We propose to combine both the information on BIC and the number of clusters m. The value of original BIC contains the information about the quality of clustering for each number of clusters. The knee point of BIC has to be the one that reflects this information overall. Two main features should be

satisfied i.e. the detected knee point can indicate the most significant change, and be as large as possible. The proposed method is designed based on these two features.

Given the range of m: $[m_{min}, m_{max}]$ where $m_{max} >> m_{opt}$ to contain the optimal m, obtain BIC value for each m. Normalize the obtained BIC value into the range of $[m_{min}, m_{max}]$ to get C_1. Then C_1 is divided by the number of clusters m getting the value Cm. This is further normalized into the same range of $[m_{min}, m_{max}]$ to obtain C_2. With the normalizations, C_1 and C_2 are under the same range. BIC_{max} and BIC_{min} in (6) represent the maximal and minimal value among the BIC values. Besides, Cm_{max} and Cm_{min} are respectively the maximal and minimal value among the Cm values.

$$C_1 = (m_{max} - m_{min})\,(BIC - BIC_{min})/(\,BIC_{max} - BIC_{min})$$

$$Cm = C_1 / m$$

$$C_2 = (m_{max} - m_{min})\,(Cm - Cm_{min})\,/(Cm_{max} - Cm_{min}) \qquad (6)$$

The value of Cm calculates the ratio between the normalized BIC value and the number of clusters, which reveals the global trend of the BIC curve as is shown in Fig.2. Each Cm value represents the angle α, which makes $\tan(\alpha)=C_1/m=Cm$. Whenever there is a local maximum in the original curve, angle α will indicate a difference.

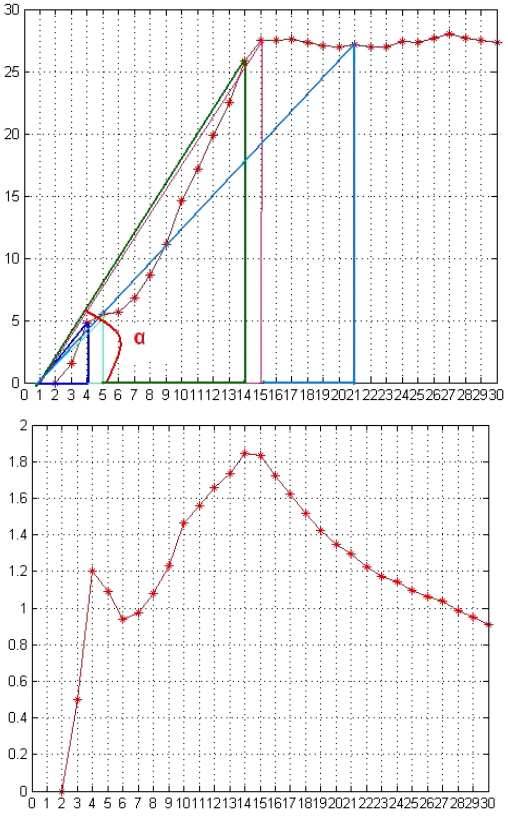


Figure 2. How the value of Cm (Normalized BIC value divided by the number of clusters) reveals the global trend. Normalized BIC curve (up); Result of Cm (down).

We consider two cases that the original BIC curve has globally increasing trend (case1) or decreasing trend (case2). Basically, a large BIC value is preferred to be the optimal m. In case1, the value depends on m_{max}, meanwhile in case2, on the other side, it depends on m_{min}. In case1, C_2 reaches several local maximums. When C_2 find the point that indicates the most significant change, it will not have an increasing trend anymore. The largest value of C_2 is considered as the most significant change. Thus, the sum of C_1 and C_2 will be calculated to reach the maximum information. In the other case, the original BIC has a decreasing trend, which makes C_2 to show a decreasing trend. As both of C_1 and C_2 are decreasing, the absolute subtraction of them is calculated to reach the most significant change. In both cases, two is divided in order to set the $DiffBic$ value into the same range of C_1.

$$DiffBic = \begin{cases} (C_1 + C_2)/2 \ldots\ldots\ldots\ldots case1 \\ |\,C_1 - C_2\,|/2 \ldots\ldots\ldots\ldots case2 \end{cases} \qquad (7)$$

2.3.3 Max Refinement. The range of m: $[m_{min}, m_{max}]$ is user-defined, which is assumed to contain the optimal m. Basically the most reliable way is $m_{max} = n$, n is the size of the dataset. However, m_{max} will be set as a more reasonable value in practice because of the heavy computation when $m_{max} = n$. In this paper, we define the m_{max} large enough, and then a max refinement is carried out.

There will be intersections across the C_1 and $DiffBic$ value in (7) because of the normalizations whenever the trend of the original BIC is increasing or decreasing. The positions of the intersection are affected by the setting of m_{min} and m_{max}. We assume that m_{max} is large enough to contain m_{opt}. With the assumption that $m_{max} \geq m_{opt}$, the first intersection $m=max'$ where $max' \neq m_{min}$ and $max' > m_{opt}$ exists. The value of max' can be thought as the refinement to m_{max} value. With this max refinement, the range of m can be reduced to $[m_{min}, max']$. There are two reasons for max refinement designing. One is that the original range setting is arbitrary; and the refined range is a smaller range that already contains the optimal value. The other is that BIC has an increasing or decreasing trend with the increment of the number of clusters, the points after the intersection has less information. Refine the original range $[m_{min}, m_{max}]$ into smaller one $[m_{min}, max']$ can make the decision accurately.

Finding the maximum value of $DiffBic$ in the new range: $[m_{min}, max']$, the optimal number of clusters is obtained by the proposed method. As Fig.3 shows, we get the second case for datasets s1 to s4. For each dataset, an intersection can be found to refine the max value. The maximum value of the proposed method is thought as the optimal number of clusters. According to this, the results from the proposed method is: m_{opt}(s1)= 15, m_{opt}(s2)= 15, m_{opt}(s3)= 15 and m_{opt}(s4)= 15.

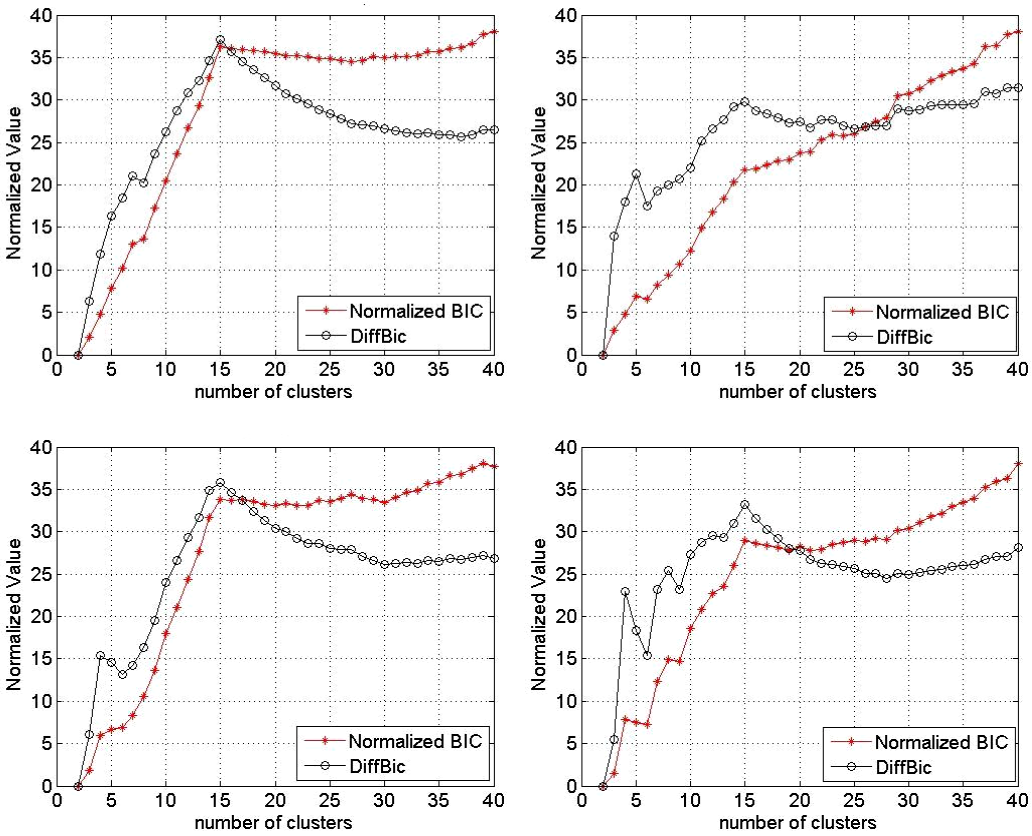


Figure 3. The results come from the proposed method for datasets s1 to s4 (left to right, up to down) with RLS clustering algorithm. Normalized BIC is represented as C1 in the context; DiffBic is the result from the proposed method.

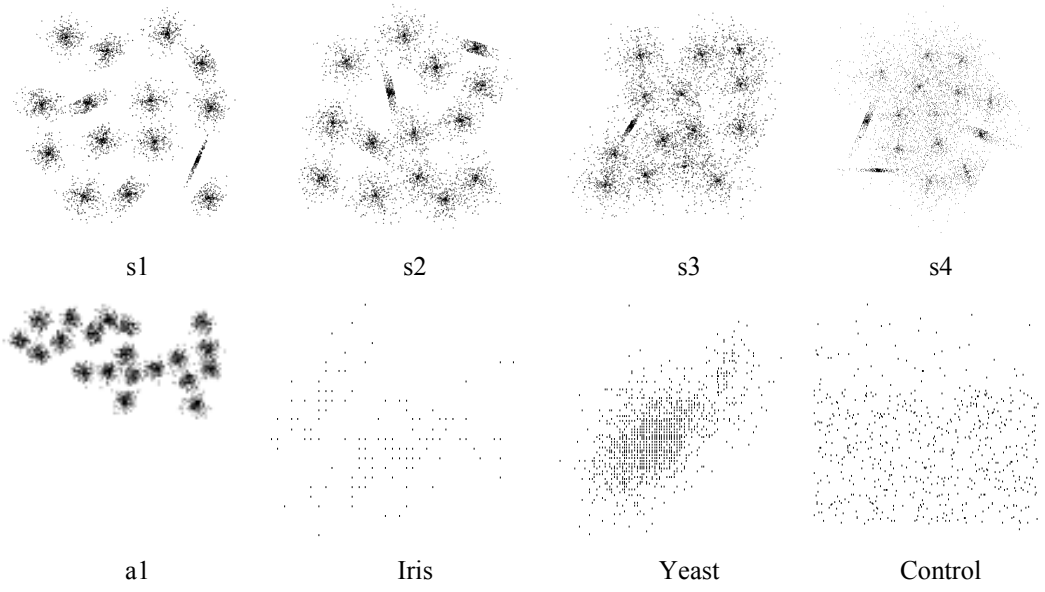


Figure 4. Two-dimensional visualization of the datasets for experiments.

We have further experiments on the proposed method with more datasets here. Both artificially generated datasets and real datasets are tested. The two dimensional view of the datasets is shown in Fig.4.

3. Experimental Results

The datasets s1 to s4 are generated with varying complexity in terms of spatial data distributions, which have 5000 vectors scattered around 15 predefined clusters with varying degrees of overlapping. The dataset a1 is generated in 2-dimensional Gaussian distribution. Datasets Iris, Yeast and Control are obtained from the UCI Machine Learning Repository. Iris contains 3 classes of 50 instances each, where each class refers to a type of iris plant. Yeast is originally used for protein localization sites prediction. The class distribution from a rule-based expert system indicates the optimum number of clusters as 10. However, as the size of 6 clusters among them is too small, our clustering algorithms reach 5 clusters as the optimal clustering. Dataset Control contains 600 examples of control charts synthetically generated by the process of Alcock and Manolopoulos (1999). There are six different classes of control charts.

The data sets can be found here:

- s1-s4, a1: http://cs.joensuu.fi/~isido/clustering/
- Iris, Yeast, Control: www.ics.uci.edu/~mlearn/MLRepository.html

Table 1. Data sets with their properties including the size of the dataset, dimension, the number of clusters and how they have been generated.

Data Set	Size	Dimension	No. of Clusters	Generated
s1-s4	5000	2	15	synthetic
a1	3000	2	20	synthetic
Iris	150	4	3	real
Yeast	1484	8	5	real
Control	600	NA	6	real

As cluster validity criterion is related to clustering algorithm, we test the revised BIC on both K-means and Randomized Local Search (RLS) [14] clustering algorithms. The RLS method is run using 5000 iterations and 2 K-means iterations within the algorithm. Meanwhile, in the K-Means clustering algorithm, 20 iterations are used for synthetic datasets (s1-s4, a1), 200 iterations for Iris and Yeast, and 500 iterations for control dataset. The proposed knee point detection method is then applied to the calculated BIC value. The results from different datasets by the proposed method with K-means and RLS clustering algorithms are summarized in Fig.5 and Fig.6.

The results from different datasets with RLS clustering algorithm are all visible and correct. However, the results from the K-means clustering algorithm are not good for real datasets even if the number of iterations is well-tuned. The datasets Control gets the result m_{opt}(control)=5. This can not prove the failure on our knee point detecting method; the actual reason is the K-Means clustering algorithm itself. Table 2 shows the results from different knee point detection methods on BIC.

4. Conclusions

Determining the number of clusters is one of the most difficult problems in cluster analysis. 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, a new knee point detecting method of BIC is proposed in this paper. As the proposed method takes advantage of the information of criterion and number of clusters, it is reliable to get the optimal results. Experimental results on different kinds of data sets also prove its effectiveness.

Table 2. The number of clusters obtains from different knee point detection method on BIC. BIC represents the first local maximum. SD is the successive difference on BIC. Cm is the value that gets from the normalized BIC value divided by the number of clusters, taking the maximum as its optimal value. DiffBic represents the proposed knee point detection method.

Method	Data Sets							
	s1	s2	s3	s4	a1	Iris	Control	Yeast
BIC	15	4	4	5	3	3	2	2
SD	15	15	4	5	3	17	2	2
Cm	15	14	4	14	3	NA	2	2
KP	15	15	15	15	20	3	6	5

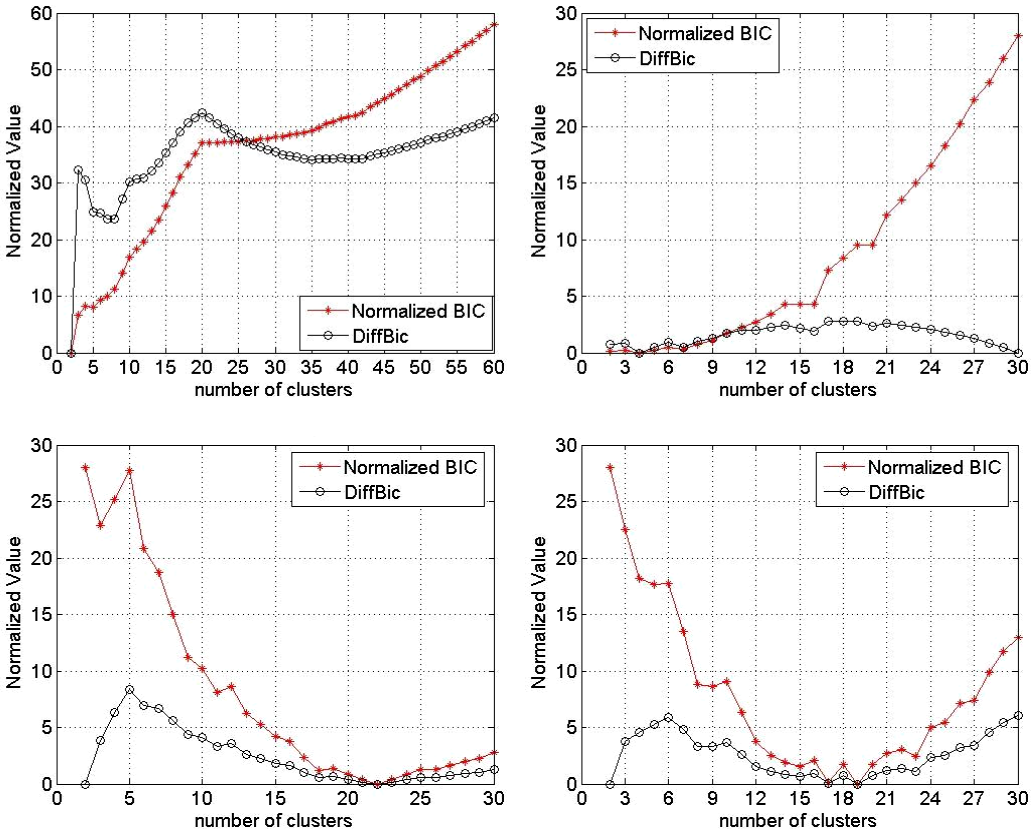
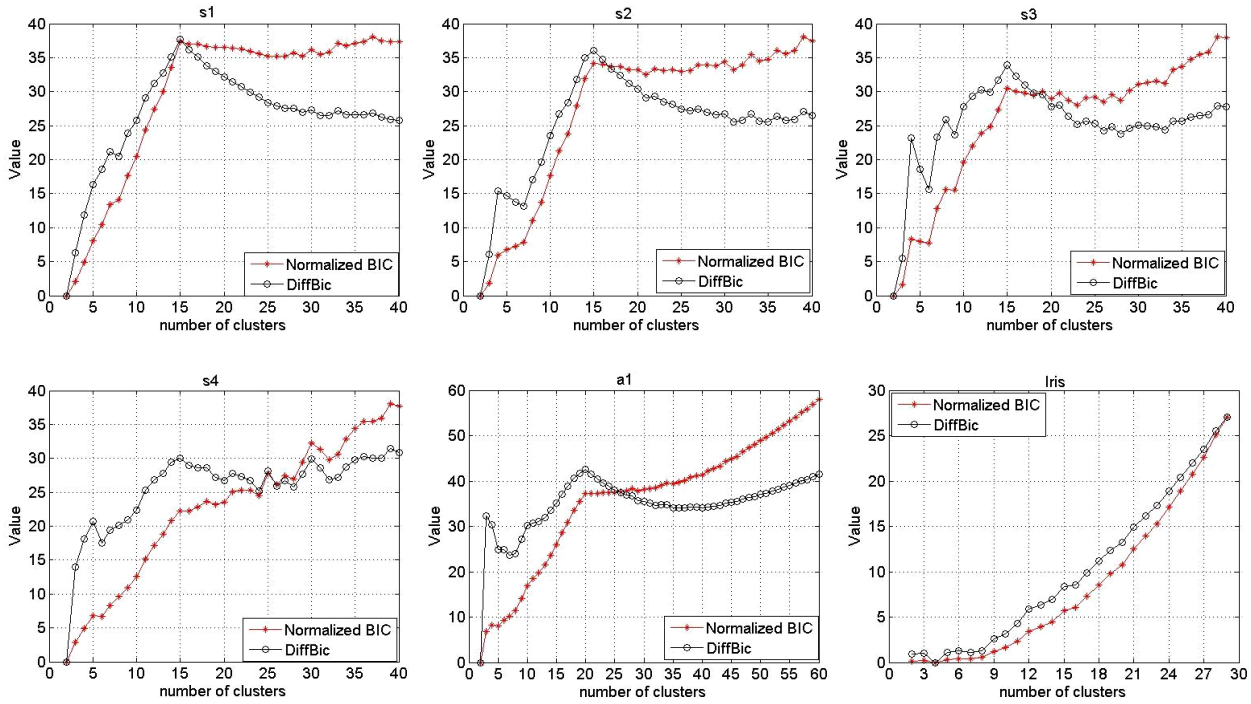


Figure 5. Results on different datasets (a1, Iris, Yeast, Control from left to right, top to down) with RLS clustering algorithm; Normalized BIC is represented as C1 in the context; DiffBic is the result from the proposed method.



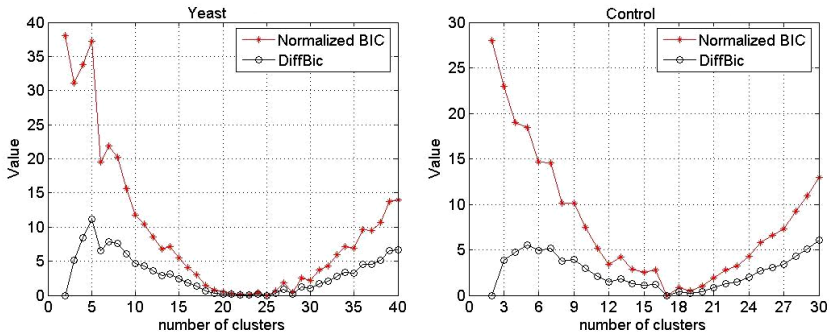


Figure 6. Results on different datasets (a1, Iris, Yeast, Control from left to right, top to down) with K-Means clustering algorithm; Normalized BIC is represented as C1 in the context; DiffBic is the result of the proposed method.

5. Acknowledgements

This research is supported by CIMO fellowship.

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] X.L. Hu and L. Xu. Investigation on several model selection criteria for determining the number of cluster. *Neural Information Processing*, Vol. 4, No.1, July 2004.

[4] R.E. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, Vol.90, No.430, pp. 773-795, 1995.

[5] C. Frayley and A. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer Journal*, Vol.41, No.8, pp. 578-588, 1998.

[6] A. Dasgupta and A. Raftery. Detecting features in spatial point process with clutter via model-based clustering. *Journal of the American Statistical Association*, 93, pp. 294-302, 1998.

[7] 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.

[8] S.S. Chen and P.S. Gopalakrishnan. Clustering via the Bayesian information criterion with applications in speech recognition. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol.2, pp. 645-648.

[9] R.L. Thorndike. Who belongs in the family? *Psychometrika*, Vol. 18, 267-276, 1953.

[10] 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.

[11] 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.

[12] J.D. Banfield and A.E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, Vol. 49, pp. 803-821, 1993.

[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.