# Sum-of-Square Based Cluster Validity Index and Significance Analysis

Qinpei Zhao, Mantao Xu, Pasi Fränti

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

**Abstract.** Different clustering algorithms achieve different results to certain data sets because most clustering algorithms are sensitive to the input parameters and the structure of data sets. Cluster validity, as the way of evaluating the result of the clustering algorithms, is one of the problems in cluster analysis. In this paper, we build up a framework for cluster validity process, meanwhile a sum-of-squares based index for cluster validity purpose is proposed. For homogeneous data based on independent variables, the proposed clustering validity index is effective in comparison to some other commonly used indexes. We use resampling method in the framework to analyze the stability of clustering algorithm, and the certainty of cluster validity index also.

## 1. Introduction

Clustering is an unsupervised process which tries to discover the unknown structure of data sets accurately. There are quite many clustering algorithms [1] based on different existing strategies. They are developed to satisfy with different needs from data sets. The common sense is that there is no general algorithm applicable to all kinds of data sets. The problem comes up that how to evaluate the effect of clustering algorithms on different data sets. Cluster validity provides the way of validating the quality of clustering algorithms and means of discovering the natural structure of data sets. If cluster analysis is to make a significant contribution, much more attention must be paid to cluster validity issues. Cluster validity measures are the methods, which can not only compare the results of two different sets of cluster analysis to determine the better one, but determining the “correct” number of clusters in the data set as well.

Many different indexes of cluster validity have been proposed. Milligan and Cooper [2] have presented a comparison study over thirty validity indexes for hierarchical clustering algorithms whereas Dimitriadou et al [3] conducted their comparison study over fifteen validity indexes for the case of binary data. Different measures under different situations achieve different answers. We introduce several indexes mentioned in these two literatures to be compared with our method.

We separate the measures in this paper into two types, one is sum-of-square based type, and the other is classic methods. The methods in the first type measure the dispersion of the data points in a cluster and between the clusters respectively. The indexes are:

- Ball and Hall [4], the maximum value of the successive difference determines the optimal number of clusters.
- Calinski and Harabasz [5], the minimum value of the successive difference determines the optimal number of clusters.
- Hartigan [6], the minimum value of the successive difference determines the optimal number of clusters.
- Xu [7], maximum value can be determined as the optimal number of clusters, the successive difference is applicable but not necessary.

The classic measures are mostly proposed in different area and work quite well to some extend. These measures share the advantage that they use the maximum or minimum as the optimal number of clusters.

- Dunn's index [8], maximum of the index value is determined as the optimal number of clusters.
- Davies-Bouldin index [9], minimum of the index value is determined as the optimal number of clusters.
- Xie-Beni's separation index [10], minimum of the index value is determined as the optimal number of clusters.
- Bayesian Information Criterion [11], which is a model selection criteria. The first local maximum is determined as the optimal number of clusters.
- Silhouette Coefficient [12], maximum of the index value is determined as the optimal number of clusters.

Application of resampling method, bootstrapping, subsampling, or cross validation to cluster validity is not new in cluster validity. Peck et al. [13] developed a bootstrap-based procedure to obtain approximate confidence bounds on the number of clusters in the "best" clustering. Ben-Hur et al. [14] present a method that exploits measurements of the stability of clustering solutions obtained by perturbing the data set. Cluster validation by prediction strength [15] is to view clustering as a classification problem, which uses the way of cross validation technique. Dudoit and Fridlyand [16] introduce a prediction-based sampling method, CLEST, in which, the data is first spilt into two non-overlapping sets; then the learning set is clustered and a classifier is built using the obtained labels; the test set is also clustered and the obtained labels are compared using an external index. We make the framework of cluster validity process with resampling methods inside to validate the clustering algorithm and the validity index.

The rest of the paper is organized as follows. We introduce the framework of cluster validity in Section 2. The proposed new index is formulated in Section 3. Experiments on the proposed method are presented in Section 4, which displays the results on both artificial generated and real data sets. Two clustering algorithms are applied in the experiment. A further step on variability and certainty analysis is introduced in Section 5. Conclusions and future work are drawn in Section 6.

## 2. Related Work

Cluster validity is related to the clustering algorithms. The fundamental clustering problem is to partition a given data set into groups, such that the points in the same group are more similar to each other than the points in different groups. Thus, one way for cluster validity is to analyze the within-between group variance.

Let $X = \{x_1, x_2 \ldots x_n\}$ be a set of data with n samples. Suppose the samples in $X$ have hard labels that mark them as representatives of m non-overlapping clusters, says $C = \{C_1, C_2 \ldots C_m\}$. The clustering algorithm is to find the optimal partition $P = \{P_1, P_2 \ldots P_m\}$. The most important parameter among them is the parameter $m$, the number of clusters, as most of the clustering algorithms need the parameter $m$ as the input and the clustering result is also affected by it.

Given the data set $X$, a specific clustering algorithm, and a fixed range of number of clusters, the basic procedure of cluster validity involves:

- Fix the data sets with externally information.
- Repeat a clustering algorithm successively for the number of clusters, $m$ from a predefined minimum $m_{min}$, to a predefined maximum $m_{max}$.
- Get the clustering results: partitions and codebooks. Calculate the index value of each number of clusters.
- Plot the “number of clusters vs. index metric” graph and select the m at which the partition appears to be “best” according to how the index is optimized.
- Compare the detected number of clusters ($m^*$) with the “externally information” to prove the effectiveness of the index.

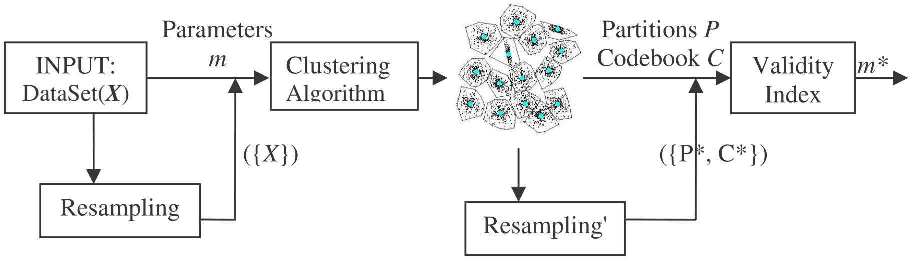


**Fig. 1.** Scheme diagram of cluster validity process

The clustering algorithm can be any kind of algorithms existing. We use the Randomized Swap algorithm (RLS) [17] to help the validity procedure complete. RLS clustering algorithm takes use of both the k-means and the advantage of local search. To eliminate the effect on index from the clustering algorithm, K-means clustering, the most typical clustering algorithm is also tested in this paper.

Based on this procedure, we can easily have the scheme diagram of cluster validity in Fig.1. To estimate the stability of the clustering algorithm, we could use resampling method just as the resampling part shows. On the other hand, in order to exclude the effect of data sets and clustering algorithm, another resampling method is used, as the resampling' part shows. We will introduce this part in section 4 in detail.

Basically, to prove the effectiveness, comparison is needed. The two types' indexes mentioned above are compared to the proposed index in the experiments section.

## 3. Proposed Method

In cluster analysis, the within group variance and between group variance can be calculated by *sum-of-squares within cluster* (SSW) and *sum-of-squares between clusters* (SSB) respectively. We analysis the existing index based on SSW and SSB, and then propose a sum-of-squares based method, so-called WB-index.

The value of SSW is defined as:

$$SSW(C,m) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j\in C_i} \| x_j - C_{P(j)} \| \tag{1}$$

which is minimized over all *m*-partitions *C* in the clustering procedure. According to ANOVA, the *total sum-of-squares* (SST) can be decomposed into two parts that are SSW and SSB for any partition *C*.

$$SSB(C,m) = \frac{1}{n}\sum_{i=1}^{m} n_i \| C_i - \bar{x} \| \tag{2}$$

where $n_{\mathrm{i}}$ is the number of elements in each cluster, and $\bar{x}$ is the mean value of the whole data set, *m* is the number of clusters. Hence, we can now define a generalized *within-between cluster type* (SSWB) as follows, which is a function of the SSW or SSB:

$$SSWB = function(SSW(C,m), SSB(C,m)) \tag{3}$$

**Table 1.** Sum-of-based indexes.

| No. | Index Name | Formula |
|---|---|---|
| 1 | Ball & Hall | $SSW/m$ |
| 2 | Calinski&Harabasz | $CH = \frac{SSB/(m-1)}{SSW/(n-m)}$ |
| 3 | Hartigan | $H-index = -\log(SSW/SSB)$ |
| 4 | Xu | $Xu = d\log(\sqrt{SSW/(dn^2)}) + \log(m)$ |

The sum-of-based methods above (table.1) are all based on the property of SSW and SSB. We study on these indexes in Fig.1. As in Fig1.(a) shows, the trends of normalized SSW and SSW/SSB are almost same, which indicates that the factor of SSW has a more important effect in the ratio of SSW/SSB. In other *WB-type* indexes except for Xu's index, we find that they either monotonously increase/decrease or need additional knee point detection method such as successive difference to get the optimal number of clusters. Xu's index has clear minimum knee point, but as our experiments will show in section 4, it doesn't work well on real data sets.

Thus, we propose a simpler sum-of-square method, WB-index as:

$$WB = m \bullet SSW \,/\, SSB \qquad (\mathbf{4})$$

We emphasize the effect of SSW with multiplying the number of clusters. The advantages of the proposed method are that it determines the number of clusters by minimal value of it without any knee point detection method, and it is quite simple and efficient.

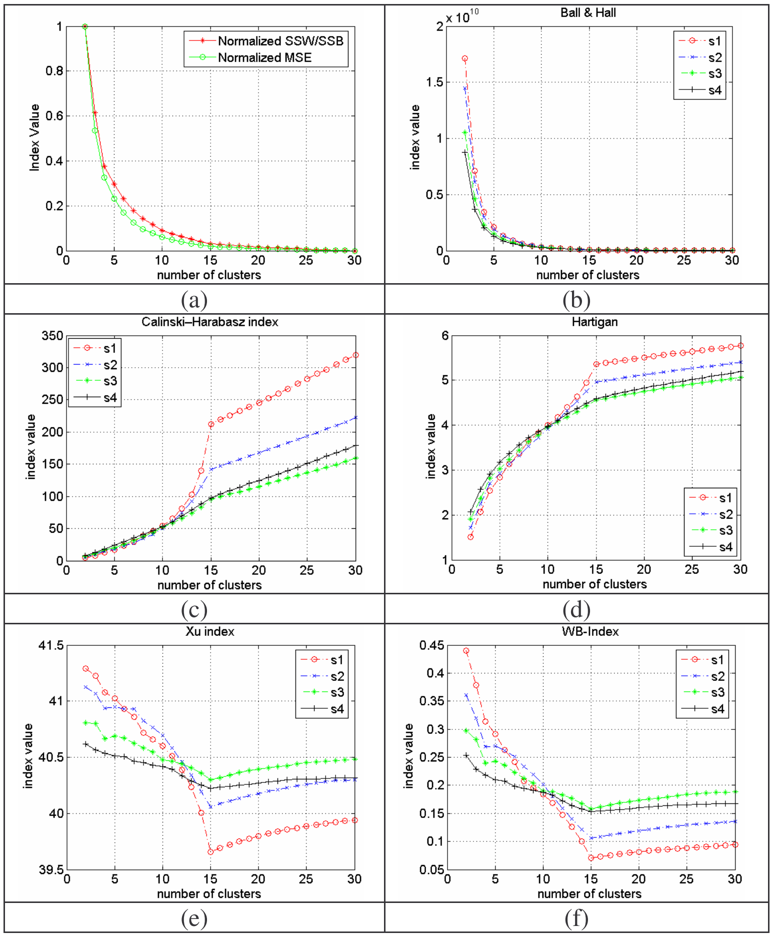


**Fig. 2.** (a). Comparison on SSW and SSW/SSB; (b)-(f). Comparison on several sum-of-square based indexes on four artificial data sets.

## 4. Experimental Results

In this paper, we test the methods with the data sets in table2. 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. The datasets a1 and R15 is generated in 2-dimensional Gaussian distribution. Iris and Breast are the real data sets that obtained from the UCI Machine Learning Repository. Iris is a four-dimensional data set, which contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The second real data set is the Wisconsin breast cancer data set (Wolberg and Mangasarian, 1990).

For comparative purpose, we test the following other five classic measures:

- Dunn's index (DI)
- Davies-Bouldin's Index (DBI)
- Xie-Beni (XB)
- Bayes Information Criterion (BIC)
- Silhouette Coefficient (SC)

In the special case of $m$=1, SSW equals to SST. Clustering algorithm is therefore performed by $m$=[2,30] in the case of S1-S4 and $m$=[2,10] in the case of real data sets.

**Table 2.** Information on data sets in the experiments.

| DataSet | Size | Dimension | # of clusters | Generated |
|---|---|---|---|---|
| s1-s4 | 5000 | 2 | 15 | artificial |
| a1 | 3000 | 2 | 20 | artificial |
| R15 | 600 | 2 | 15 | artificial |
| Breast | 699 | 11 | 2 | real |
| iris/Iris | 150 | 4 | 3 | real |

**Table 3.** Results using RLS (with 5000 RLS iterations and 2 K-means iterations).

| **DataSet** | **BH*** | **CH*** | **Har*** | **Xu** | **DI** | **DBI** | **XB** | **SC** | **BIC*** | **WB-INDEX** |
|---|---|---|---|---|---|---|---|---|---|---|
| s1-s4 | 3 | **15** | **15** | **15** | **15** | **15** | **15** | **15** | **15** | **15** |
| | 3 | **15** | 4 | **15** | 7 | **15** | **15** | **15** | 4 | **15** |
| | 4 | **15** | 4 | **15** | 16 | 8 | 4 | **15** | 4 | **15** |
| | 3 | **15** | 3 | **15** | 25 | 13 | 13 | **15** | 5 | **15** |
| a1 | 3 | **20** | 3 | **20** | 34 | **20** | **20** | **20** | 3 | **20** |
| R15 | 3 | **15** | **15** | **15** | 2 | **15** | **15** | **15** | 8 | **15** |
| Breast | 3 | 3 | 3 | NA | 14 | **2** | **2** | **2** | **2** | **2** |
| Iris | **3** | **3** | **3** | NA | 2 | 2 | 2 | 9 | 6 | **3** |

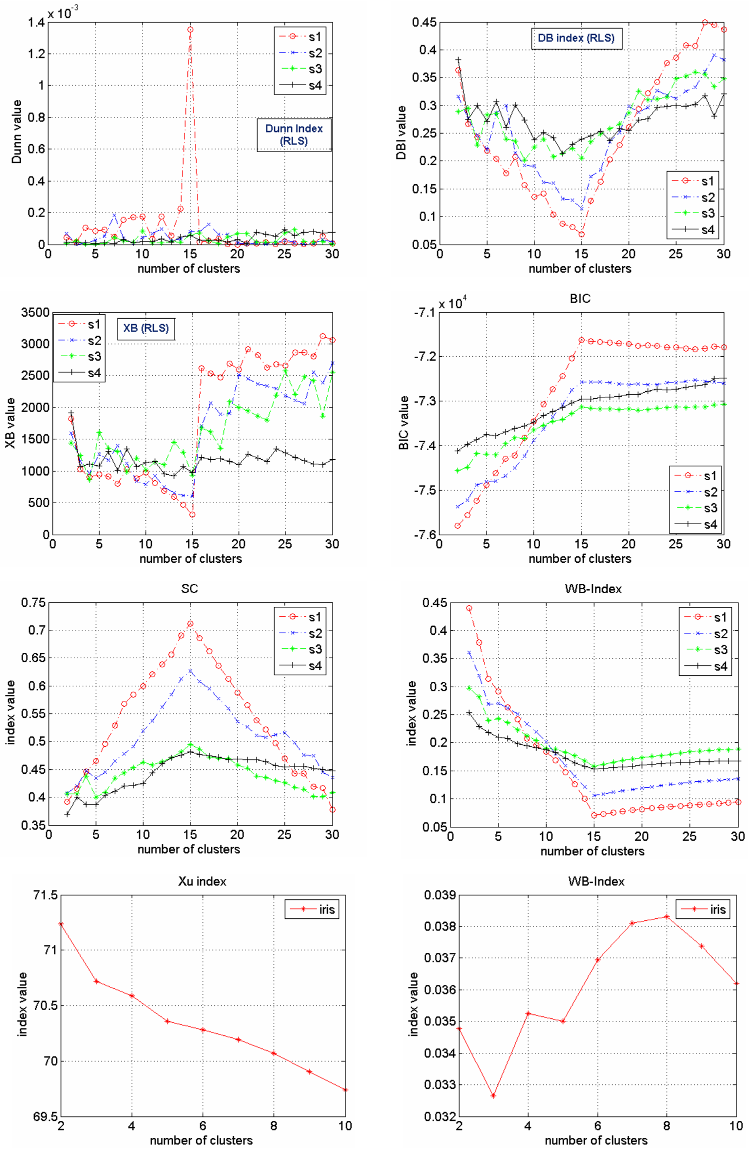


**Fig. 3.** The results on different validity indexes and data sets. The results of Xu's index and the proposed index on real data set Iris are on the last row. It is unable to find the minimum value of Xu's index as the optimal number of clusters because it is monotonously decreasing.

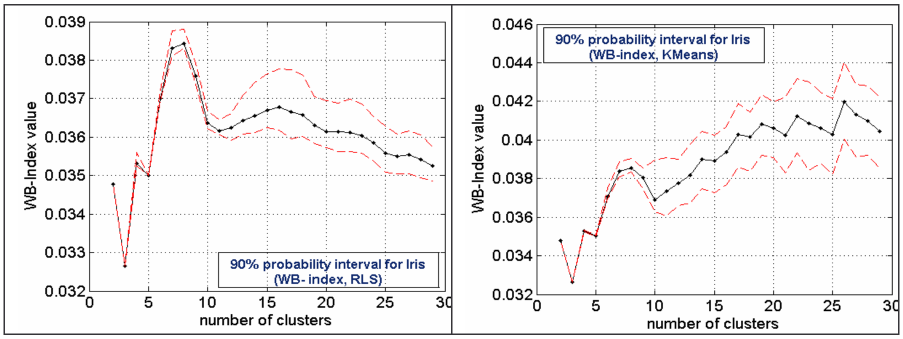
## 4. Significance Analysis

The results of the experiments with different clustering algorithms and data sets indicate that the proposed index provides the accurate estimate of the number of clusters, which indicates the effectiveness on cluster validity. Basically, we can demonstrate the proposed index as it shows in the experiments. However, we want to confirm the results in this section by further analysis.

### 4.1 Variability Analysis

With uncertain distribution of the results, a natural approach for the variability estimation associated with each index value in order to test for statistical significance can use resampling method. As in Fig.1 shows, resample on the original data set (*X*), get a new data set (*X**) and apply the new data set for the validation procedure again. Repeat it *B* times, deal with *B* times index values to get the statistical significance. However, the RLS clustering algorithm is designed with randomization, in which, there is random swapping of the code vectors. Thus, we use a simpler way here to keep the data set unchanged and take use of the randomization of the clustering algorithm by running *B* times to analyze the results.

**Fig. 4**. 90% probability interval of WB-index with RLS and KMeans clustering on data set Iris



Quartile range is one of the measures used to estimate variability. We use it into our scheme to analyze the variability of each index value. With the same setting of input parameters, fix the number of clusters, run the clustering algorithm *B* times to get *B* values on the same number of clusters. Then the 5th and 95th percentiles of the *B* index values are used to get 90% probability range.

Iris data as a real data set is representative to be tested. According to the results, only the proposed method and BIC after knee detection get the correct number on Iris. In this case, both of the clustering algorithms with the same data set and index are tested. We run $B = 100$ times of the clustering algorithm with same input parameters setting. The 90% probability interval with RLS and K-means is shown respectively in Fig.4, the dash line is the boundary of the range. It is clear that the range $m = [2, 3]$ indicates strong evidence that many of clusters in the data set with RLS clustering;

meanwhile, $m$ = [2, 5] with K-means clustering. The range of m is wider with K-means clustering than RLS clustering, from which we can conclude that RLS clustering is more stable than K-means, and the variability on $m$ = 3 is convincingly not so strong on Iris data set.

### 4.2 Certainty Analysis

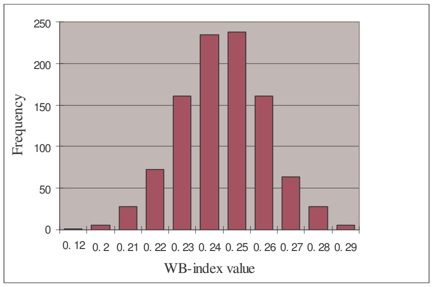
We develop another way to prove the certainty of the proposed index, which takes use of the resampling method. As Fig.1 shows, the effect of the validity index is affected both by the data set and the clustering algorithm. In this case, resampling the data set can not prevent the effect comes from the clustering algorithms. In this sense, we process the resampling method on the partitions getting from the clustering to avoid this problem.

In the first run of the validity procedure, a set of partitions (*P*) is generated. Basically, this set of partitions is the optimal one according to the clustering algorithm. A WB-index value (*WBI*) is obtained on *P*. We permutate the original partitions (*P*) by *B* times, getting {*P**}, and recalculate the index values {*WBI**}. As the optimal value of the WB-index should be as small as possible, we can estimate the certainty by counting the probability that $WBI^* \leq WBI$.

$$P = \frac{No.(WBI^* \leq WBI)}{TotalNo.(WBI^*)} \tag{5}$$

The smaller the probability *P* is, the more certainty the method obtains. It is not practical to calculate all possible permutations because of the time involves. Generally, at least *B* =1000 times permutations should be done. 1000 random permutations were done on the partitions (Fig.5). It indicates the certainty of the index, as the observed optimal value is much smaller than any of the values obtained under permutation.

**Fig. 5.** Distribution of WB-index on Iris data set (m=3) for 1000 permutations of the partitions with RLS clustering. The “optimal” value of WBI is very extreme by reference to this distribution (WBI = 0.032653).



## 5. Conclusions

We represented a framework with resampling step for the estimation on stability of clustering algorithm and variability of validity index in cluster validity process. In addition, we proposed a new sum-of-square based index which indicates simplicity and good prospect comparing to other indexes. Based on the proposed index, we finish a whole process on cluster validity.

Reference:

1. R. Xu, and D. Wunsch, “Survey of clustering algorithms”, *IEEE Transactions on Neural Networks*, Vol. 16, No.3, 2005, pp. 645-678.
2. 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.
3. 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.
4. Ball, G.H., & Hubert, L.J. ISODATA, A novel method of data analysis and pattern classification (Tech. Rep. NTIS No. AD 699616). Menlo Park, CA: Standford Research Institute, 1965.
5. T. Calinski, and J. Harabasz. A dendrite method for cluster analysis. *Communication in statistics*, Vol.3, pp. 1-27, 1974.
6. Hartigan, J.A. *Clustering algorithms*. New York, NY: Wiley, 1975.
7. Xu, L. Bayesian Ying-Yang machine, clustering and number of clusters. *Pattern Recognition Letters*, 18, 1167-1178, 1997.
8. J.C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetica*, Vol.4, pp. 95-104, 1974.
9. 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.
10. 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.
11. 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.
12. L.Kaufman and P.J.Rousseeuw. Finding Groups in data. *An Introduction to cluster analysis*. New York: Wiley, 1990.
13. R. Peck, L. Fisher, and J.V. Ness, “Approximate confidence intervals for the number of clusters”, *Journal of the American Statistical Association*, Vol. 84, No. 405, 1989, pp. 184-191.
14. A. Ben-Hur, A. Elisseeff, and I. Guyon, “A stability based method for discovering structure in clustered data”, *Proceedings of the Pacific Symposium on Biocomputing*, 2002, pp. 6-17.
15. R. Tibshirani, and G. Walther, “Cluster validation by prediction strength”, *Journal of Computational & Graphical Statistics*, Vol. 14, No.3, 2005, pp. 511-528.
16. S. Dudoit, and J. Fridlyand, “A prediction-based resampling method for estimating the number of clusters in a dataset”, *Genome Biology*, Vol. 3, No.7, June 2002.
17. 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.