Extending external validity measures for determining the number of clusters

Qinpei Zhao
School of Computing
University of Eastern Finland
Joensuu, Finland
qinpei.zhao@uef.fi

Mantao Xu
School of Electronics & Information
Shanghai Dian Ji University
Shanghai, China
mantao.xu@gmail.com

Pasi Fränti
School of Computing
University of Eastern Finland
Joensuu, Finland
pasi.franti@uef.fi

***Abstract*—External validity measures in cluster analysis evaluate how well the clustering results match to a prior knowledge about the data. However, it is always intractable to get the prior knowledge in the practical problem of unsupervised learning, such as cluster analysis. In this paper, we extend the external validity measures for both hard and soft partitions by a resampling method, where no prior information is needed. To lighten the time burden caused by the resampling method, we incorporate two approaches into the proposed method: (i) extending external validity measures for soft partitions in a computational time of $O(M^2N)$; (ii) an efficient sub-sampling method with time complexity of $O(N)$. The proposed method is then applied and reviewed in determining the number of clusters for the problem of unsupervised learning, cluster analysis. Experimental results has demonstrated the proposed method is very effective in solving the number of clusters.**

***Keywords*-external cluster validity, clustering, subsampling, image segmentation**

I. Introduction

External validity measures are preferable for evaluating the goodness of clusterings when ground truth labels are available [1]. With the ground truth consisting of class labels assigned to the patterns, the ideal clustering is selected based on how well the cluster labels produced by the algorithm match. External measures are also used to compare the similarity of two clustering results.

Rand Index [2], [3], Jaccard coefficient, Fowlkes and Mallows index [4] are typical external measures, which evaluate the clustering quality by the similarity of the pairs of data objects in different partitions. A study of 16 external measures for K-means clustering has been conducted in [5]. According to the result of this survey, we only investigated Adjusted Rand Index in the experiments.

External measures are mainly designed for hard partitions. Researchers shed light on extensions of external measures for fuzzy results. A fuzzy extension of the Rand index has been introduced [6]. Other measures such as adjusted Rand Index, the Jaccard coefficient, the Fowlkes and Mallows index have also been derived from the same formulation. However, they are as computationally expensive as $O(M^2N^2)$, where M is the number of clusters and N is the data size. Thanks Michele's pioneer solution of fuzzy clustering, the computational cost of external measures has significantly reduced to $O(M^2N)$ time [7].

In clustering, however, prior knowledge of the data is usually not available. To overcome this difficulty, Rand Index was extended to calculate a pairwise stability [8], where the pairwise stability is calculated as the variability of the clustering results by resampling the original data or multiple initializations. For example, bootstrap resampling has been utilized in evaluating the fuzzy partition stability in [9], and its fuzzy extension has been introduced in [6]. However, these methods lead to a high time complexity in general.

Since the goal of image segmentation shares the commonalities with clustering, several clustering methods have been applied in image segmentation successfully [10]. A common way to evaluate the segmentation result is supervised evaluation, in which manually segmented reference images are used as ground truth. However, external information is difficult to acquire and require human assistance. Generating a reference image is also a subjective, and time consuming task. Even given the reference information, it is not guaranteed that the reference is unique. Considering the difficulty, a framework for a similarity measure is suggested in [11], where the measure is based on an objective comparison between the results from image segmentation algorithms and several manual segmentations.

In this paper, we mainly extend the external measures by a resampling method to the case of cluster analysis that no ground truth is available. The proposed method combined both the benefits of resampling method and fast implementation of external measures in determining the number of clusters, which is applicable for both hard and soft partitions in clustering problems. With numerous clustering algorithms and varies of image types, evaluation of the segmentation result is an open question. We employed the proposed method on segmentation evaluation to prove the validity of the method.

II. External measures

Clustering aims at partitioning a set of N and d-dimensional data points $X = \{x_1, x_2, ..., x_n\}$ into M

978-1-4577-1675-1

clusters. The partition is defined as:

$$P = [p_{ij}]_{N\times M}; \sum_{j=1}^{M} p_{ij} = 1 \quad (1)$$

Here P is a $N \times M$ partition matrix, p_{ij} represents the probability of the i^{th} point belonging to the j^{th} cluster. In hard clustering, p_{ij} is either 0 or 1, while in soft clustering $p_{ij} \in (0, 1)$. Given two partitions P and G, external validity measures are used to measure the similarity of two clusterings by the proportion of pairs of vectors that agree by belonging either to the same cluster or to different clusters in both partitions.

A. Hard partitions

External validity measures can be computed from the contingency matrix in $O(M^2 + N)$ time for hard partitions. A contingency matrix is defined as:

$$C_{ij} = \sum_{t=1}^{N} I(P(t) = i \wedge G(t) = j) \quad (2)$$

where I is the indicator function, t is the data point, and $i, j < M$ are the group labels. The quantities a, b, c, d are defined as follows:

$$\begin{aligned} a &= \sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij}^2 - N \\ b &= \sum_{j=1}^{M}(\sum_{i=1}^{M} C_{ij})^2 - \sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij}^2 \\ c &= \sum_{i=1}^{M}(\sum_{j=1}^{M} C_{ij})^2 - \sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij}^2 \\ d &= \sum_{i=1}^{M}\sum_{j=1}^{M}(C_{ij}\sum_{l\neq i}^{M}\sum_{s\neq j}^{M} C_{ls}) \end{aligned} \quad (3)$$

These calculate the number of points that belongs to the same cluster in P and G (a); belongs to the same cluster in P but to different in G (b); inverse of b (c); are in different groups in P and G (d). Terms a and d measure the amount of agreement of P and G, whereas terms b and c measure the amount of disagreement. The Adjusted Rand index is now derived by:

$$ARI = \frac{2 \times (a \times d - b \times c)}{(c \times c + b \times b + 2 \times a \times d + (a + d) \times (c + b))} \quad (4)$$

The definitions of Rand index, Jaccard coefficient and Fowlkes-Mallows indices are all based on Eq. 3, see [6], [7] for the exact definition.

B. Efficient extension to soft partitions

An efficient extension of external measures into soft partitions in [7] is based on an update definition of contingency matrix. In soft partitions, each point has a membership/probability value to each cluster. The calculation of contingency matrix for soft partitions is defined as:

$$C_{ij} = \sum_{t=1}^{N} (P_{ti} + G_{tj})^{\alpha} \quad (5)$$

where, t is the data point, i, j are the number of cluster. The value α is used to boost the influence of higher memberships and reduced the influence of lower memberships. We discuss its setting in Section IV. Given two soft partitions of the same data set, the contingency matrix is calculated according to Eq. 5. The calculations of a, b, c, d are the same as in Eq. 3 with the only difference of the calculation of a, which is defined as follows:

$$a = \sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij}^2 - \sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij} \quad (6)$$

The time complexity of the soft version is $O(M^2N)$.

III. Determining the number of clusters

For determining the number of clusters, the basic idea is to test whether the points in a data set are randomly structured or not. Resampling techniques such as bootstrapping [12], subsampling [13], cross validation [14], sampling by Monte Carlo method [15] have been utilized as a solution for this problem. They allow one to simulate the process of estimating the probability density function of a validity measures using random numbers, i.e. the resampling-based method discover the structure of the data by simulation. However, as a non-parametric method the resampling method requires high computation.

We propose a resampling-based method for determining the number of clusters in a more efficient way. The idea of this method is to estimate $index(k)$ by comparing an index value on the original data I_x with an expectation under index values on an appropriate null reference distribution of the data I_u. The estimated optimal number of clusters is the value that minimizes $index(k)$.

$$index(k) = E_b[I_u] - I_x \quad (7)$$

where E_b denotes expectation under a sample with size B from the reference uniform distribution. This idea is similar with the gap-statistic [15]. The most significant difference is that the proposed method is applied to external measures, which are working differently with internal measures. The proposed method is applicable both to hard and soft clusterings. To speed up the method, an efficient extension of soft external measures and a sub-sampling method are employed.

The proposed method is described in Algorithm 1. First, we perform a sub-sampling algorithm [16] with $O(N)$ time complexity on the original data set to reduce the computation. The parameters setting is the same as in [16]. Clustering

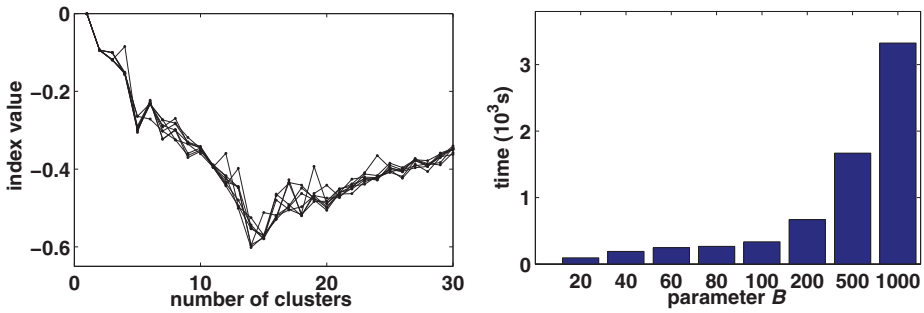

Figure 1. The settings of parameter *B* make little difference on the performance of the proposed index (left) while the processing time increases with the increment of *B* value (right).

algorithm is run on the sub-sampled data X_s, and P_x is the result. We compute an index value I_x of the defined external index between a reference partition G and P_x. Here, G is built according to the intuition about the clustering structure of the data set that the data set is not randomly collected. Let $c = \lfloor N/M \rfloor$, the reference partition $G_{N\times M}$ is generated by:

$$[G]_{ij} = \begin{cases} 1, if \quad i > (j-1)\times c \quad \& \quad i < j \times c + 1 \\ 0, otherwise \end{cases} \tag{8}$$

Next, B synthetic data sets X_b are generated in the area of the sub-sampled data (for each dimension independently) by a uniform distribution. The same clustering algorithm is run on these data sets, and let P_{ub} be the resulting clustering. We compute index values I_{ub} of the same external index between the reference partition G and P_{ub}. Index values I_{ub} are the approximation of the probability density function of the defined external index. We define $I^* = \sum_{b=1}^{B} I_{ub}/B$ as a reference of I_x.

Input: $X = \{x_1, x_2, ..., x_n\}$, K_{max}
Output: K_{opt}

```
Xs = subsampling(X) ;
for k = 2 : K_max do
    Set reference labels G = [g_ij]_{N×M} ;
    P_x = CLUSTER(Xs) ;
    I_x = ExternalIndex(P_x, G) ;
    for b = 1 : B do
        Generate reference data X_b uniformly ;
        P_ub = CLUSTER(X_b) ;
        I_ub = ExternalIndex(P_ub, G) ;
    end
    index(k) = I* − I_x ;
end
K_opt = min(index) ;
return K_opt
```

Algorithm 1: Pseudocode of the proposed method

Finally, I_x and I_{ub} are obtained for different number of clusters within the range $k \in [2, K_{max}]$, where a rule of thumb of K_{max} is $K_{max} \sim (N/2)^{1/2}$ [17]. Thus, $index(k) = I^* - I_x$ is calculated under different k, and $K_{opt} = argmin_k\{index(k)\}$.

Table I
DESCRIPTION OF THE DATA SETS, D IS DIMENSIONALITY, N IS DATA SIZE AND M IS NUMBER OF CLUSTERS.

Name	D	N	M	Generated
Touching	2	73	2	artificial
rdata3	2	300	3	artificial
S1-S4	2	5000	15	artificial
Iris	3	150	3	real
wine	13	178	3	real(Normalized)
wdbc	30	569	2	real(Normalized)
Zernike	47	2000	10	real(Normalized)
image	3	116*261	NA	real

IV. EXPERIMENTS

Experiments were performed on several real and synthetic data sets (Table I). The data set S1-S4 consists of 5000 vectors and 15 Gaussian clusters with different degree of cluster overlapping. The rdata3 is generated under Gaussian distribution with three smaller groups of data points. Touching contains two connecting clusters. The real data sets are obtained from UCI Machine Learning Repository [18]. All real data sets instead of Iris are normalized by statistical normalization. The image in (Fig. 4) in YUV color space is used for image segmentation. We test the proposed method on K-means (KM), EM and Fuzzy C-means (FCM) clustering algorithms for hard and soft clustering.

For the setting of parameter *B*, we run the proposed method with increasing *B* values. As shown in Fig. 1, the increment of *B* value increases the processing time while it brings little effect on the index value. Thus *B* in Algorithm 1 is set to 20 to get less processing time.

Spearman's rank correlation [19] is a non-parametric measure of statistical dependence between two variables. To decide the setting of α, we calculate the Spearman's rank correlation among ARI for hard partitions and ARI for soft

partitions in different α settings in Table II. As shown in the table, it has very high correlation among the ARI values on $\alpha = 10$, $\alpha = 15$ and $\alpha = 20$. However, it has very low correlation among the values when $\alpha = 1$ and the others. The correlation of the ARI values on hard partitions and soft partitions is the highest when $\alpha = 5$. Thus, we set $\alpha = 5$ in this paper.

Table II
SPEARMAN'S RANK CORRELATION AMONG ARI FOR HARD PARTITIONS AND ARI FOR SOFT PARTITIONS IN DIFFERENT α SETTINGS.

	hard	α=1	α=5	α=10	α=15	α=20
hard	1	0.65	**0.87**	0.79	0.78	0.76
α=1	0.65	1	0.64	0.57	0.53	0.51
α=5	0.87	0.64	1	0.95	0.93	0.91
α=10	0.79	0.57	0.95	1	0.99	0.98
α=15	0.78	0.53	0.93	0.99	1	1
α=20	0.76	0.51	0.91	0.98	1	1

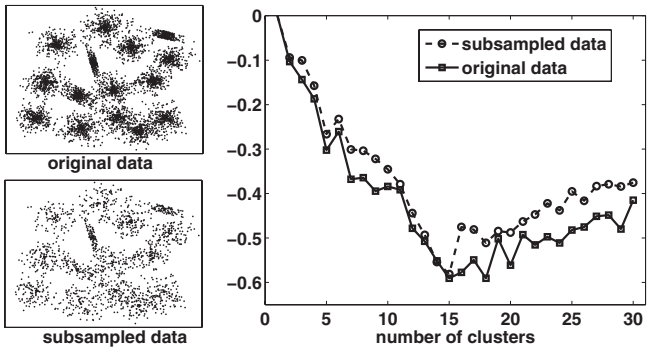

Figure 2. Original and sub-sampled data distribution (left) and the results from the proposed method on both data sets (right).

A. Sub-sampling algorithm

First, we need to verify if the sub-sampling algorithm affects the final result. As shown in Fig. 2, the original data size is sampled from 5000 to 1724 data points so that the data structure is preserved while the density is reduced. The result of the proposed method on the sub-sampled data has similar trend as that on the original data in Fig. 2. It indicates that the sub-sampled data works well in our method, although variation exists.

Table III
THE PROCESSING TIME WITHOUT AND WITH SUB-SAMPLING ON DIFFERENT PARTS IN THE PROPOSED METHOD. THE RUNNING TIME FOR THE SUB-SAMPLING PROCEDURE IS 0.09 SECONDS.

	without	with	reduced
External measures	0.13s	0.08s	38%
K-means	8.41s	1.82s	78%
EM	28.43s	8.42s	70%
FCM	42.12s	14.22s	66%

The time costs in Table III are for external measures and clustering algorithms in Algorithm 1 for data set S2 with 15 clusters. The sub-sampling method reduces time cost 38%-78% on different parts in the proposed method according to Table III. Compared to the running time for sub-sampling procedure, which is 0.09s, the reduced time is much more than that. The sub-sampling method can reduce remarkable running time in resampling method since the clustering algorithms employed in the method need to be repeated multiple times.

B. Determining the number of clusters

We tested the proposed method on the data sets in Table I. The hard partitions are resulted from K-means and the soft partitions of FCM and EM algorithm by taking the cluster with the maximal membership value.

An example on data S2 is shown in Fig. 3, where the data distribution with partitioning from FCM is displayed. The running time of the proposed method varies from different clustering algorithms, where K-means is the fastest and EM is the slowest. Thus, the running time depends highly on the choice of the clustering algorithm. Compare hard and soft clusterings, for example, hard and soft partitions from FCM and EM, the computation time have little difference as Fig. 3 indicates.

The index values of the proposed method with the increasing number of clusters are plotted, where the minimal values of the curve indicate the number of clusters. For data set S2, the proposed method reveals the structure of the data set on hard partitions from different clustering algorithms. Results from soft partitions work similar as those from hard partitions with higher variance. The main reason is that external measures on hard partitions are more robust than that of soft partitions.

To validate the proposed method, we listed the determined number of clusters for the data presented in Table I by the proposed method and two internal measures [20] in Table IV. Calinski-Harabsz (CH) index is popular as an internal measure, which is based on within and between cluster variance. Xie and Beni proposed a validity index (XB) for fuzzy clustering, which considered the data set, geometric distance measure, distance between cluster centroids and more importantly on the fuzzy partition generated by any fuzzy algorithm used. We employed K-means results for Calinski-Harabsz index and FCM results for Xie-Beni index. The bold-faced numbers in Table IV represents the correctly determined number of clusters.

For determining the number of clusters, the proposed method works well on real data sets and small Gaussian-distributed data sets. For higher overlapped data S3 and S4, the proposed method works well. In general, it has better performance on hard partitions than soft ones. Internal measures have less accurate result on real data sets, but Xie-Beni index works well on artificial data sets.

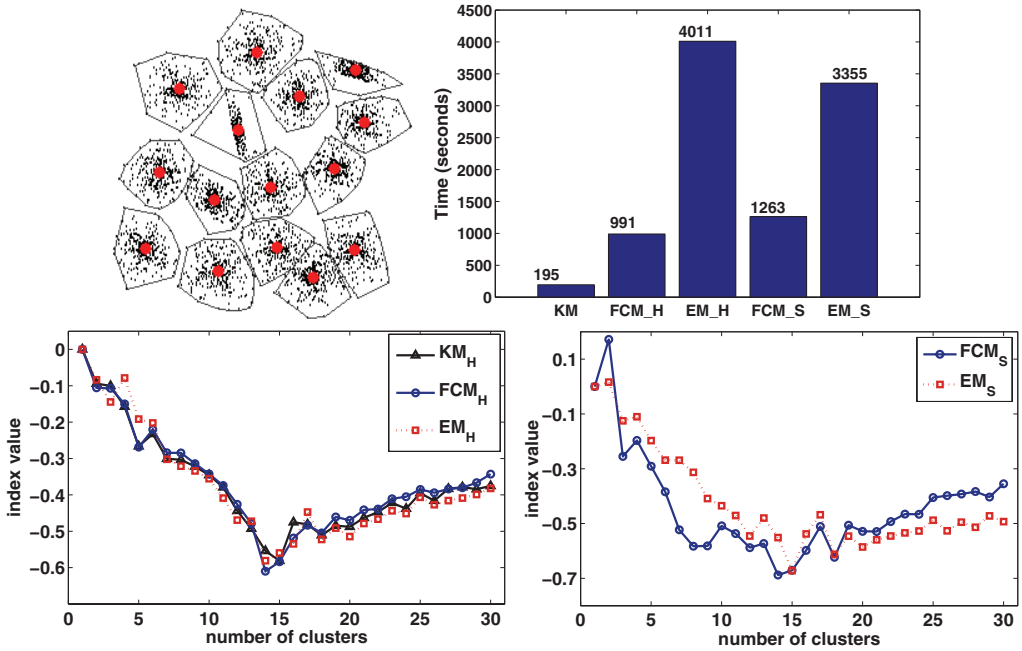

Figure 3. (a) a clustering on data set S2 from FCM; (b) a comparison on the running time of the proposed method on different soft and hard clusterings; (c) and (d) the index value of the proposed method (hard and soft clustering respectively) on the increasing number of clusters.

Table IV
THE NUMBER OF CLUSTERS DETERMINED BY THE PROPOSED METHOD FOR HARD AND SOFT PARTITIONS.

Data	KM_H	FCM_H	EM_H	FCM_S	EM_S	XB	CH
Iris	**3**	**3**	**3**	**3**	2	2	2
wine	**3**	**3**	**3**	**3**	**3**	2	2
wdbc	**2**	**2**	**2**	6	8	**2**	**2**
Zernike	**10**	5	**10**	4	**10**	2	2
image	3	3	3	5	5	2	2
rdata3	**3**	**3**	**3**	**3**	2	**3**	10
Touching	**2**	**2**	**2**	3	6	**2**	**2**
S1	17	14	17	13	18	**15**	**15**
S2	**15**	14	14	14	**15**	**15**	20
S3	**15**	**15**	**15**	**15**	**15**	4	6
S4	**15**	**15**	**15**	**15**	**15**	**15**	16

C. Unsupervised evaluation of image segmentation

For image segmentation, K-means, FCM and EM algorithms are conducted. For simplification, the segmentation results are represented by the clustering labels. We use the proposed method to determine the number of clusters for images. The proposed index and BIC (Bayesian Information Criterion) are used to evaluate the segmentation results, see Fig. 4, where the image in YUV color space is segmented into 3 clusters by different algorithms. The evaluation under different numbers of clusters (from 2 to 10) are shown in Fig. 5. There is a strong indication on three clusters by the proposed method, whereas BIC on EM algorithm has no clear suggestion.

Figure 4. An image in YUV color space and image segmentations of three clusters by KM, EM and FCM.

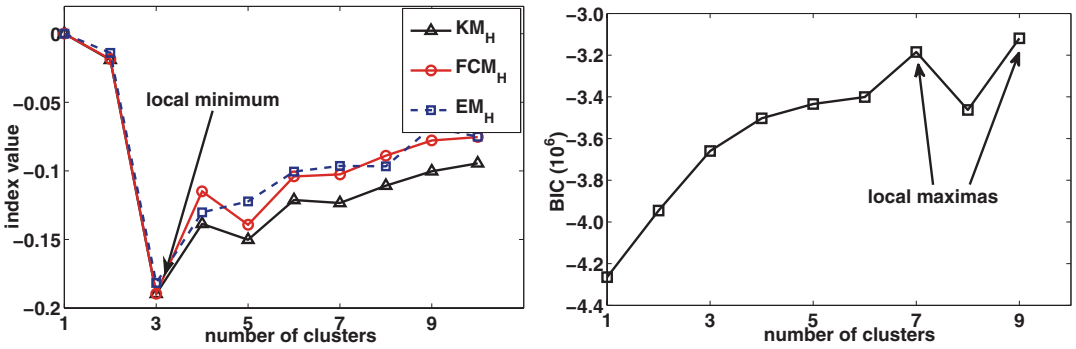

Figure 5. Evaluation results of the proposed method (left) and BIC (right) for the image.

V. Conclusion

We extended the external measures for both hard and soft partitions in clustering problems when no prior information is available. To the computational efficiency of resampling method, a state-of-art sub-sampling method is implemented and applied instead. Experimental results indicate that the proposed method are very effective in solving the number of clusters in practice, for example, the unsupervised evaluation of image segmentation. The proposed method can be envisioned as a general approach that can be applicable to other clustering algorithms and external measures.

References

[1] B. E. Dom, "An information-theoretic external cluster-validity measure," *Research Report RJ 10219, IBM*, 2001.

[2] W.M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.

[3] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clustering comparison: Is a correction for chance necessary?," *ICML'09*, pp. 1073–1080, 2009.

[4] E.B. Fowlkes and C.L. Mallows, "A method for comparing two clusterings," *Journal of the American Statistical Association*, vol. 75, pp. 553–569, 1983.

[5] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," *KDD'09*, pp. 877–886, 2009.

[6] R.J.G.B. Campello, "A fuzzy extension of the rand index and other related indexes for clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833–841, 2007.

[7] C. Michele and M. Antonio, "A fuzzy extension of some classical concordance measures and an efficient algorithm for their computation," *Int'l Conf. on Knowledge-Based Intelligent Information and Engineering Systems*, pp. 755–763, 2008.

[8] L.I. Kuncheva and D.P. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE TPAMI*, vol. 28, no. 11, pp. 1798–1808, 2006.

[9] M. Falasconi, A. Gutierrez, M. Pardo, G. Sberveglieri, and S. Marco, "A stability based validity method for fuzzy clustering," *Pattern Recognition*, vol. 43, no. 4, pp. 1292–1305, 2010.

[10] C. Carson, S. Belongie, H. Greenspan, and J.Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE TPAMI*, vol. 24, no. 8, pp. 1026–1038, 2002.

[11] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE TPAMI*, vol. 29, pp. 929–944, 2007.

[12] M.H.C. Law and A.K. Jain, "Cluster validity by bootstrapping partitions," *Technical Report MSU-CSE-03-5, Dept. of Computer Science and Engineering, MSU, Michigan, USA*, 2003.

[13] E. Levine and E. Domany, "Resampling method for unsupervised estimation of cluster validity," *Neural Computation*, vol. 13, pp. 129–137, 2001.

[14] W.M. Abd-Elhafiez A. Farag M. El-Melegy, E.A. Zanaty, "on cluster validity indexes in fuzzy and hard clustering algorithms for image segmentation," *ICIP'07*, pp. 5–8, 2007.

[15] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society*, vol. 63, pp. 411–423, 2001.

[16] T. Hasan, Y. Lei, A. Chandrasekaran, and J.H.L. Hansen, "A novel feature sub-sampling method for efficient universal background model training in speaker verification," *ICASSP'10*, pp. 4494–4497, 2010.

[17] K. V. Mardia, J. T. Kent, and J. M. Bibby, "Multivariate analysis," *Academic Press*, 1979.

[18] A. Asuncion and D.J. Newman, "UCI machine learning repository," http://archive.ics.uci.edu/ml/, 2007.

[19] Jerome L. Myers and Arnold D. Well, "Research design and statistical analysis (second edition ed.)," *Lawrence Erlbaum*, p. 508, 2003.

[20] Q. Zhao, M. Xu, and P. Fränti, "Sum-of-square based cluster validity index and significance analysis," *ICANNGA*, pp. 313–322, 2009.