CLUSTERING VALIDITY BASED ON THE IMPROVED S_DBW* INDEX[1]

Tong Jianhua Tan Hongzhou

(*Department of Electronics and Communication Engineering, Sun Yat-Sen University, Guangzhou 510275, China*)

Abstract For many clustering algorithms, it is very important to determine an appropriate number of clusters, which is called cluster validity problem. In this paper, a new clustering validity assessment index is proposed based on a novel method to select the margin point between two clusters for inter-cluster similarity more accurately, and provides an improved scatter function for intra-cluster similarity. Simulation results show the effectiveness of the proposed index on the data sets under consideration regardless of the choice of a clustering algorithm.

Key words Clustering validity; Inter-cluster similarity; Intra-cluster similarity

CLC index TP391

DOI 10.1007/s11767-007-0151-8

I. Introduction

Clustering, perceived as an unsupervised process, analyzes data objects without class labels. Based on the principle of maximizing the intra-cluster similarity and minimizing the inter-cluster similarity, the objects are clustered or partitioned. The optimal partition should be compatible for the inherent classification characteristic of data set. How to evaluate a clustering result and choose the optimal partition of data is one of the most important issues in clustering analysis. In general terms, there are three kind of methods for evaluating clustering validity: clustering validity checking approaches based on internal, those on external criteria and those on relative criteria[1]. The major drawback of techniques based on internal or external criteria is their high computational demands. Because they use Monte Carlo simulation which scans the data sets multiple times resulting in exponential complexities. The fundamental idea of relative criteria is to choose the best clustering scheme of a set of defined schemes according to a pre-specified criterion. The clustering validity function is a kind of pre-specified criterions in common use, which is used to find the best number of clusters. There are many clustering validity functions proposed in the literature. They belong to three kinds approximately: the validity functions based on fuzzy partition[2], those on the geometrical structure[3] and those on statistic of data[4]. Currently, the research of clustering validity on the geometrical structure of data has been paid more attention.

In the paper, a new clustering validity index S_Dbw_{new} is proposed, which takes advantage of the concept underlying S_Dbw^* index[3], while resolving problems in S_Dbw^* index. S_Dbw_{new} index can deal with inter-cluster similarity and intra-cluster similarity more robustly. And it enables the selection of optimal number of clusters more effectively for a clustering algorithm. The rest of this paper is organized as follows. Section II presents the related work and Section III defines a new validity index, S_Dbw_{new}. Then, in Section IV, we show the simulation results of S_Dbw_{new} index compared with the results of S_Dbw^* index. Finally, in Section V, we conclude by briefly presenting our contributions and provide directions for further research.

II. Related Work

There are many indices for clustering validity assessment including Dunn and Dunnlike indices[6], DB (the Davies-Bouldin) index[7], RMSSDT (Root-Mean-Square Standard Deviation of the new cluster) index[8], SD index[9,10], the modified Hubert Γ statistic index[9–11] , S_Dbw (Scatter and Density

[1] Manuscript received date: August 21, 2007; revised date: January 18, 2008.

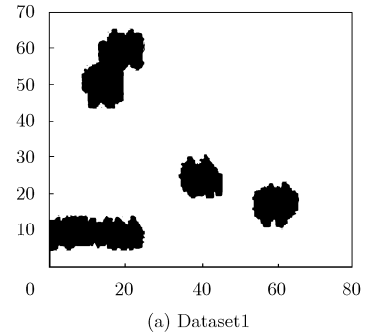
Communication author: Tan Hongzhou, born in 1965, male, Professor. Department of Electronics and Communication Engineering, Sun Yat-Sen Uuniversity, Guangzhou 510275, China.
Email: issthz@mail.sysu.edu.cn.

between clusters) index[5], *etc.* The definition of S_Dbw index indicates that both criteria of inter-cluster and intra-cluster similarity are properly combined, enabling reliable evaluation of clustering results. Recently research shows that S_Dbw index has better performance and effectiveness than other indices[4,5].

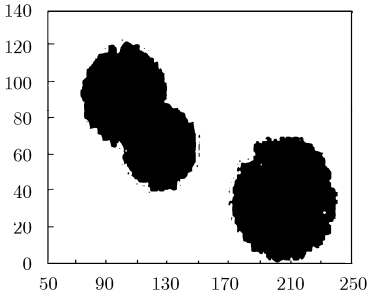
Fig.1 depicts five data sets used in Ref.[5]. Intuitively, Dataset1 has four clusters, Dataset2 has two or three clusters, Dataset3 has seven clusters, Dataset4 has three clusters, and Real data has three clusters, respectively. Tab.1 shows the optimal number of clusters proposed by RS, RMSSTD, DB, SD, S_Dbw indices for the data sets given in Fig.1. It indicates that S_Dbw index can find optimal clusters for all given data sets[5].

Tab.1 Optimal number of clusters proposed by different validity indices[5]

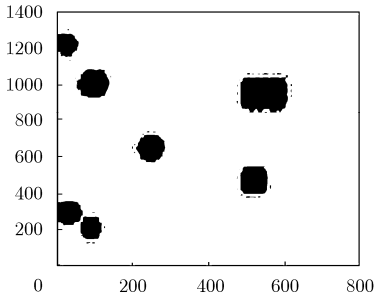
Index	Optimal number of clusters				
	Dataset1	Dataset2	Dataset3	Dataset4	Real data
RS, RMSSTD	3	2	5	4	3
DB	6	3	7	4	3
SD	4	3	6	3	3
S_Dbw	4	2	7	3	3



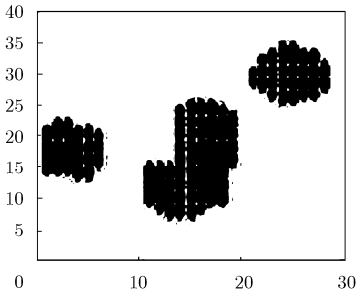
(a) Dataset1



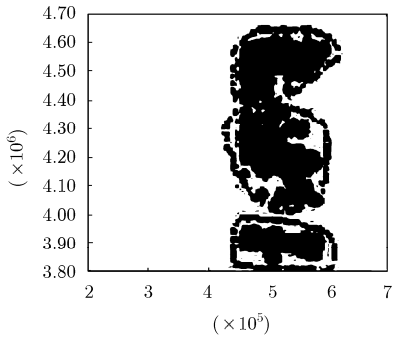
(b) Dataset 2



(c) Dataset 3



(d) Dataset 4



(e) Real data

Fig.1 Sample synthetic datasets and a real dataset[5]

Though S_Dbw index is effective in many cases, it has some problems. For example, it cannot handle properly the case of non-circular clusters. To settle above problem S_Dbw* index[3] is proposed. Firstly, it finds the neighborhood by using the confidence interval of each dimension rather than a hyper-sphere, so that the size and the shape of each cluster is reflected in the density function. Secondly, it measures the weighted average of scattering within clusters, so that the data ratio of each cluster is reflected in the scatter function. The simulation results implies that S_Dbw* index outperforms S_Dbw index, especially when the real partition is noncircular or sparse.

III. S_Dbw$_{new}$ Index

S_Dbw* index still has some shortcomings.

Firstly, it selects the middle point of the line segment defined by the centers of two clusters as the point which is used to compute the density of the margin region between the two clusters. But when the sizes and densities of the two clusters have large discrepancy, the middle point of the line segment can not represent the margin point between the two clusters. Fig.2 shows two close clusters whose sizes and densities are very differrent. The coordinate axises of Fig.2 provide the situation of every point. We can obviously find that the middle point (85,50) belongs to the larger cluster from Fig.2(a), and that the margin point (105,50) is closer to the smaller cluster from Fig.2(b). Secondly, though Scat^* of S_Dbw* index for intra-cluster similarity measures the scattering more accurately by calculating relative ratio between the average variance of clusters and the variance of the entire data set, it will be monotonically increasing with the number of clusters getting large which will destroy the precision of the S_Dbw* index.

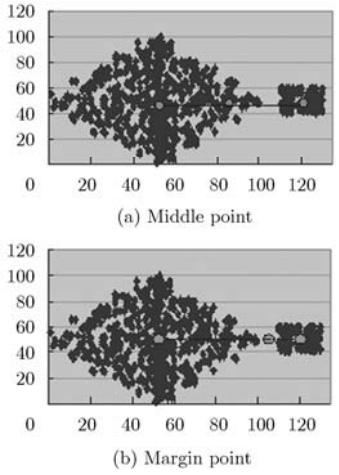


Fig.2 The middle point of the line segment and the margin point between two clusters

To overcome the problem of S_Dbw*, a new index S_Dbw$_{\mathrm{new}}$ is defined in the paper. S_Dbw$_{\mathrm{new}}$ is based on the same clustering evaluation concepts as S_Dbw*(inter-cluster, intra-cluster similarity). However, S_Dbw$_{\mathrm{new}}$ index is different from S_Dbw* index in the sense that it finds the more precise point which can represent the margin region between two clusters rather than the middle point of the line segment defined by the centers of two clusters. So that we can improve the measurement of the inter-cluster similarity. Another difference is that S_Dbw$_{\mathrm{new}}$ index improves the parameter for measuring the weighted average of scattering within clusters. So that the measurement of the intra-cluster similarity can be improved. Because it can maintain constant with the number of clusters increasing when other conditions are the same.

The inter-cluster similarity of S_Dbw$_{\mathrm{new}}$ index is defined as follows. Let S be a data set, and n the number of tuples in S. For a given data point m, the term density$^*(m)$ is defined as in Eq.(1):

$$\mathrm{density}^*(m)=\sum_{i=1}^{n} f^*(x_i,m) \tag{1}$$

The function $f^*(x,m)$ denotes whether the distance between two data points x and m is less or equal to the confidence interval for each dimension. Let l be the number of dimensions for the data set, then $f^*(x,m)$ is defined as:

$$f^*(x,m)=\begin{cases}1, \mathrm{CI}^p \le d(x^p,m^p) \le \mathrm{CI}^p, (\forall p, 1\le p\le l)\\ 0, \text{ otherwise}\end{cases} \tag{2}$$

where CI^p denotes the confidence intervals of p-th dimension. With the confidence interval of 95%, CI^p is defined as follows:

$$\mathrm{CI}^p = u^p \pm \left(1.96\times\frac{\sigma^p}{\sqrt{n}}\right) \tag{3}$$

where u^p and σ^p represent the average and the standard deviation of p-th dimension, respectively. Then, terms in Eq.(4) are substituted with $u_{ij}^p, \sigma_{ij}^p, n_{ij}$, where

$$u_{ij}^p=\frac{u_i^p+u_j^p}{2},\ \sigma_{ij}^p=\frac{\sigma_i^p+\sigma_j^p}{2},\ n_{ij}=n_i+n_j \tag{4}$$

Let m_{ij} be the selected point of the line segment defined by the centers of clusters v_i, v_j, which represents the margin region between the two clusters, where m_{ij}^p, v_i^p, v_j^p denote the p-th dimension of m_{ij}, v_i, v_j and n_i is the number of tuples in cluster c_i.

$$m_{ij}^p=\lambda\left(\frac{n_j v_i^p+n_i v_j^p}{n_i+n_j}\right) + (1-\lambda)\left(\frac{\mathrm{density}^*(v_i)v_i^p+\mathrm{density}^*(v_j)v_j^p}{\mathrm{density}^*(v_i)+\mathrm{density}^*(v_j)}\right) \tag{5}$$

It is clear that the point m_{ij} is closer to the cluster whose density is more and the number of tuples in which is less. Here λ is a positive constant between 0 and 1 called the intensity parameter, which can adjust the effect of the two factors. After many simulations we find that S_Dbw$_{\text{new}}$ index gets the best result when λ is equal to 0.7. From above definitions, Dens_bw$_{\text{new}}$ (c) is defined as in Eq.(6):

$$\text{Dens\_bw}_{\text{new}}(c) = \frac{1}{c(c-1)}\sum_{i=1}^{c}\left[\sum_{\substack{j=1\\ j\neq i}}^{c}\frac{\text{density}^*(m_{ij})}{\max\left\{\text{density}^*(v_i),\text{density}^*(v_j)\right\}}\right] \tag{6}$$

The basic idea of Dens_bw$_{\text{new}}$ is that it selects the proper margin point between two clusters based on both the densities of clusters and the numbers of tuples in clusters, resulting the selected point closer to the real position than that before. This implies that Dens_bw$_{\text{new}}$ is more flexible than Dens_bw* of S_Dbw* index in case that the sizes and the densities of clusters are very different, while the distance between clusters is very small.

For the intra-cluster similarity of S_Dbw* index, Scat*(c) is defined as:

$$\text{Scat}^*(c) = \frac{1}{c}\sum_{i=1}^{c}\frac{n-n_i}{n}\,||\,\sigma(v_i)^2\,||\,/\,||\sigma(S)^2\,|| \tag{7}$$

where $||x|| = (xx^{\mathrm{T}})^{1/2}$, and n is the total number of tuples in the data set S and n_i is the number of tuples in cluste c_i. It is obvious that $n = \sum_{i=1}^{c} n_i$. If two clusters have the same variance but different numbers of tuples, the variance of cluster with more tuples has more effect on the scatter function than the one with less tuples. By minimizing the influence of small clusters or noisy clusters, Scat* is very robust. But the problem of Scat* is that its value monotonically increasing with c increasing. For example, Letus assume $||\,\sigma(v_i)^2\,||\,/\,||\sigma(S)^2\,||$ maintains constant when c changes, Scat*(c) is equal to $(c-1)/c\,||\,\sigma(v_i)^2\,||\,/\,||\sigma(S)^2\,||$ which monotonically increasies when c increases. To overcome the shortcoming of Scat*(c), Scat$_{\text{new}}(c)$ is proposed in the paper which is defined as:

$$\text{Scat}_{\text{new}}(c) = \frac{\sum_{i=1}^{c}\frac{n-n_i}{n}\,||\,\sigma(v_i)^2\,||\,/\,||\,\sigma(S)^2\,||}{\sum_{i=1}^{c}\frac{n-n_i}{n}} = \frac{1}{c-1}\sum_{i=1}^{c}\frac{n-n_i}{n}\,||\,\sigma(v_i)^2\,||\,/\,||\,\sigma(S)^2\,|| \tag{8}$$

From Eq.(6) and (8), we have S_Dbw$_{\text{new}}(c)$ index, which is defined as follows:

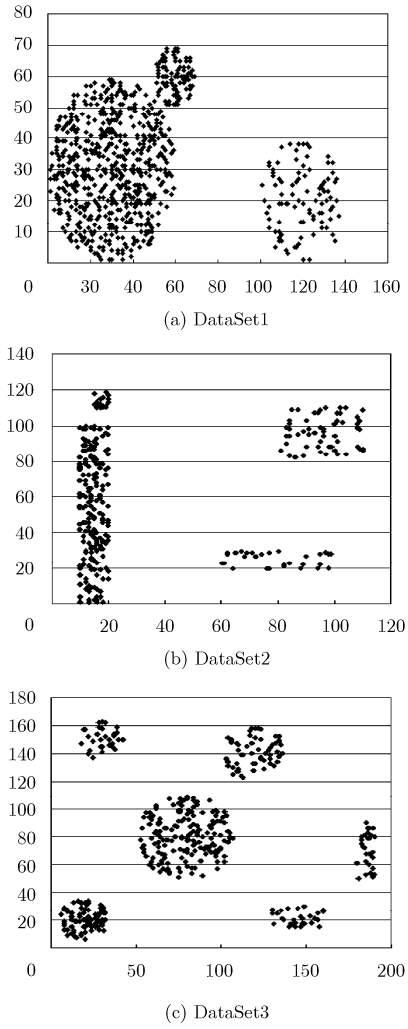
$$\text{S\_Dbw}_{\text{new}}(c) = \text{Dens\_bw}_{\text{new}}(c) + \text{Scat}_{\text{new}}(c) \tag{9}$$

The definition of S_Dbw$_{\text{new}}$ indicates that both criteria of "good" clustering (compactness and separation) are properly combined, enabling reliable evaluation of clustering results. Dens_bw$_{\text{new}}(c)$ indicates the average number of points between the c clusters (*i.e.*, an indication of inter-cluster density) in relation with density within clusters. A small Dens_bw$_{\text{new}}(c)$ value indicates well-separated clusters. Scat$_{\text{new}}(c)$ indicates the average scattering within c clusters. A small value of this term is an indication of compact clusters. The number of clusters, c, that minimizes the above index can be considered as an optimal value for the number of clusters present in the data set.

IV. Simulation

In this section, we experimentally evaluate the proposed validity index, S_Dbw$_{\text{new}}$, in comparison with S_Dbw* index. Fig.3 shows the simulation data sets consisting of three artificial data sets and one real data set. Intuitively, DataSet1 has three circular clusters, but two clusters' boundaries are very closely connected and their sizes are quite different. DataSet2 has four rectangle clusters, but two clusters consist of lined up points, one of which is very close to another small cluster. DataSet3 has six clusters. RealData is the birthday data set which gives the distribution of birthdays for births in the U.S. in 1978. Obviously it has two clusters whose shapes like two parallel wave lines[12].

S_Dbw$_{\text{new}}$ index is evaluated in comparison with S_Dbw* index experimentally by applying three representative clustering algorithms from different categories. Clustering algorithms used in the Simulation are K-Means (partition-based), DBSCAN (density-based), and SOM (Kohonen network-based). For each algorithm, the given number of clusters varies from two to eight. K-Means algorithms are applied to DataSet1 and DataSet3 to have circular clusters. We applied SOM and DBSCAN algorithms to DataSet2 and RealData, since K-Means algorithm is not appropriate to the case of non-circular sets.



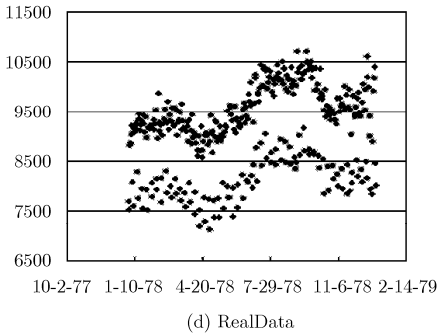


Fig.3 Samples synthetic (a) ∼ (c) datasets and a real dataset (d)

Tab.2 compares the index values evaluated by S_Dbw* and S_Dbw$_{new}$. For DataSet1, DataSet2 and RealData, two indices show the remarkable different results. For DataSet1, S_Dbw* index evaluates two clusters as the best one both inK-Means and SOM, while S_Dbw$_{new}$ index evaluates three clusters as the best one. For DataSet2, S_Dbw* index evaluates six clusters in DBSCAN and three clusters in SOM as the best one, while S_Dbw$_{new}$ evaluated four clusters as the best one in both algorithms. For RealData, S_Dbw* index evaluated four clusters as the best one and eight clusters as the worst one in SOM, while S_Dbw$_{new}$ finds two clusters correctly. For DataSet3, two indices find six clusters exactly as the best one both in K-Means and DBSCAN.

Tab.2 Number of cluster found by S_Dbw* and S_Dbw$_{new}$ for each algorithm and dataset

Dataset	Algorithms	Index	K						
			2	3	4	5	6	7	8
DataSet1	K-Means	S_Dbw*	0.357	0.466	0.418	0.561	0.526	0.482	0.514
		S_Dbw$_{new}$	0.442	0.227	0.453	0.580	0.587	0.540	0.568
	SOM	S_Dbw*	0.316	0.398	0.356	0.442	0.512	0.488	0.496
		S_Dbw$_{new}$	0.378	0.208	0.386	0.492	0.485	0.520	0.454
DataSet2	DBSCAN	S_Dbw*	0.458	0.327	0.336	0.362	0.325	-	-
		S_Dbw$_{new}$	0.504	0.462	0.302	0.412	0.403	-	-
	SOM	S_Dbw*	0.432	0.147	0.289	0.370	0.324	0.282	0.267
		S_Dbw$_{new}$	0.512	0.247	0.126	0.358	0.330	0.302	0.288
DataSet3	K-Means	S_Dbw*	0.181	0.168	0.165	0.027	0.009	0.092	0.112
		S_Dbw$_{new}$	0.244	0.201	0.189	0.038	0.016	0.114	0.136
	DBSCAN	S_Dbw*	0.181	0.168	0.165	0.032	0.009	0.088	0.117
		S_Dbw$_{new}$	0.244	0.201	0.189	0.036	0.017	0.128	0.156
RealData	SOM	S_Dbw*	0.135	0.195	0.116	0.166	0.183	0.166	0.202
		S_Dbw$_{new}$	0.112	0.178	0.196	0.182	0.163	0.204	0.197

From Fig.4 we can get the variety tendency of S_Dbw* and S_Dbw$_{new}$ value for different algorithm and dataset. When some clusters are close and have different shapes (such as DataSet1, DataSet2, RealData), the variety of two indices values are remarkable different. And the S_Dbw$_{new}$ index can evaluate the clusters more accurately than the former. Reversely, when all the clusters distribute averagely and have similar sizes (such as DataSet3), two indices have the similar variety tendency.

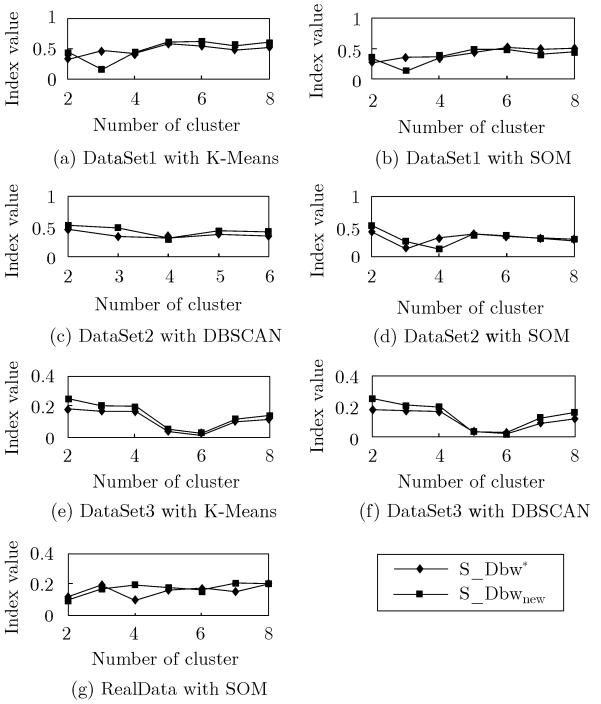


Fig.4 The variety tendency of S_Dbw* and S_Dbw$_{new}$ values for different algorithm and dataset

V. Conclusions

In this paper we address the important issue of assessing the validity of clustering algorithms' results. We define a new validity index, S_Dbw$_{new}$, for assessing the results of clustering algorithms. For inter-cluster similarity, S_Dbw$_{new}$ uses an apt formula to select the precise point which can represent the margin region between two clusters. And it is better than the middle point of the line segment defined by the centers of two clusters. While for intra-cluster similarity, S_Dbw$_{new}$ use the improved parameter for measuring the weighted average of scattering within clusters, so that the intra-cluster similarity will maintain constant with the number of clusters increasing when other conditions are the same. We evaluate the flexibility of S_Dbw$_{new}$ index experimentally on four data sets. The results imply that S_Dbw$_{new}$ index outperforms S_Dbw* index for all clustering algorithm used in the simulations, especially when the real partition does not distribute averagely and the sizes and densities of clusters are remarkable different.

The essence of clustering is not a totally resolved issue, and depending on the application domain. We may consider different aspects as more significant ones. Having an indication that a good partitioning can be proposed by the cluster validity index, the domain experts may analyze further the validation procedure.

References

[1] S. Theodoridis and K. Koutroubas. Pattern Recognition. New York, Academic Press, 1999, 79–132.

[2] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **13**(1991)4, 841–847.

[3] Youngok Kim and Soowon Lee. A clustering validity assessment Index. PAKDD'2003, Seoul, Korea, April 30–May 2, 2003, LNAI 2637, 602–608.

[4] Maria Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Intelligent Information Systems*, **17**(2001)2/3, 107–145.

[5] M. Halkidi and M. Varzirgiannis. Clustering validity assessment: finding the Optimal Partitioning of a data set. ICDM'2001, San Jose, California, USA, November 29–December 2, 2001, 187–194.

[6] J. C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics.* **4**(1974), 95–104.

[7] D. L. Davies and D. W. Douldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**(1979)2, 224–227.

[8] S. C. Sharma. Applied Multivariate Techniques. West Sussen, England, John Willy & Sons, 1996, 112–130.

[9] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part I. *ACM SIGMOD Record*, **31**(2002)2, 40–45.

[10] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part II. *ACM*

SIGMOD Record, **31**(2002)3, 19–27.

[11] Heng Zhao, Jimin Liang, and Haihong Hu. Clustering validity based on the improved Hubert Γ statistic and the separation of clusters. ICICIC'06, Beijing, China, August 30–September 1, 2006, 539–543.

[12] J. Laurie Snell Stat. Datasets – an unofficial collection. http://www.dartmouth.edu/~chance/teaching_aids/data.html, April 30, 2007.