An Ensemble Method for Validation of Cluster Analysis

Sunghae Jun

Department of Bioinformactics and Statistics, Cheongju University
Cheongju, Chungbuk 360-764, Korea

Abstract

Clustering is more subjective work than classification and regression. Though classification and regression have many general validation measures, clustering has few validation measures. Also, it is difficult to develop general measure of cluster validation. So, many evaluation measures have been published for cluster validation. In this paper, we propose an ensemble method of validation for cluster analysis. We use voting approach to some validation measures of cluster analysis. To verify our improved performance, we make experiments by some objective data sets from UCI machine learning repository.

Keywords: *Cluster Analysis, Cluster Validation, Ensemble Method, Voting, Internal measures, Stability measures.*

1. Introduction

Clustering is a rudimentary and exploratory approach to start of multivariate data analysis [1]. Also, clustering is grouping data points into clusters and then the points within a cluster are highly similar to each other [2]. That is, the variance of the points within same cluster is small and the variance of the points in other clusters is large. Recently, clustering was used in diverse fields of bio data analysis such as bioinformatics [3-10]. But, unlike classification and regression, there is no general validation measure in the clustering. Many validation measures of the clustering results have been published in diverse fields [3],[11-15]. They gave improved performance on case by case. That is, we should select a cluster validation measure for given data set. So, in this paper, we propose an ensemble method as a general evaluation for cluster validation. We use many measures such as internal and stability validations in our proposed method. All measures used in our method will be voted to validate the clustering result. We will make experiment to verify our improved performance. Three data sets from UCI machine learning repository will be used in our experiment.

2. Cluster Validation using Ensemble Method

We found various measures for cluster validation [16-18]. These measures were based on internal and external properties of clustering results [19-22]. In this paper, we divide validation measures to two types, internal and stable. First, internal validation measures reflect the compactness, connectedness, and separation of the cluster partitions [3]. From previous researches, we knew that it is difficult to evaluate clustering result by a validation measure. In this paper, we propose an ensemble method of validation for cluster analysis. We vote the results of some validation measures for efficient cluster validation. In our study, we use three internal validation measures which are connectivity, Silhouette width, and Dunn index. Let N and P denote the numbers of observations and variables respectively. We can define the connectivity as follow [3].

$$Connect(C) = \sum_{i=1}^{N}\sum_{j=1}^{R} x_{i,n_{ij}} \quad (1)$$

Where n_{ij} is defined as the jth nearest neighbor of observation i. C has k disjoint clusters, $(C_1, \ldots, C_k)$. R is used as the number of nearest neighbor. Also, we can validate the clustering result as the following.

$$x_{i,n_{ij}} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ are in same cluster} \\ \dfrac{1}{j} & \text{otherwise} \end{cases} \quad (2)$$

This value is ranges between 0 and ∞. The cluster validation is better by minimizing the connectivity value. Next, Silhouette width is defined as the average value of all observations' Silhouette values [3],[11]. This has the value between -1 and 1. The cluster validation is good when the value is close to 1 and is bad when the value is close to -1. The Silhouette width of observation i is defined as follow [11].

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \quad (3)$$

Where $a(i)$ is the distance measure as follow.

$$a(i) = d(i, A) \quad (4)$$

This is defined as the average distance between i and all other observations in cluster A. Also, $b(i)$ is the average

distance of i to the observations in the nearest neighbor cluster. In this paper, we use Dunn index as third internal measure. This index is a ratio as follow [3].

$$Dunn_{index} = \frac{\min(O)}{\max(I)} \quad (5)$$

Where O is the distance between observations not in the same cluster and I is the intra-cluster distance.
Next stability measures validate the clustering results using all instances of given data columns [3]. These measures are average proportion of non-overlap (APN), average distance (AD), average distance between means (ADM), and figure of merit (FOM) [3],[12-13]. So, we validate the results from clustering using internal and stability measures. Next table shows the criteria of all validation measures in this paper.

Table 1: Criteria of validation measures

Validation measures		***Range of value***	***Criteria of well clustering***
Internal	Connectivity	(0, ∞)	Minimized
	Silhouette	(-1,1)	Near 1
	Dunn index	(0, ∞)	Maximized
Stability	APN	(0, 1)	Close to 0
	AD	(0, ∞)	Minimized
	ADM	(0, ∞)	Minimized
	FOM	(0, ∞)	Minimized

Using these criteria for the clustering results, we can evaluate the clustering results as following figure.

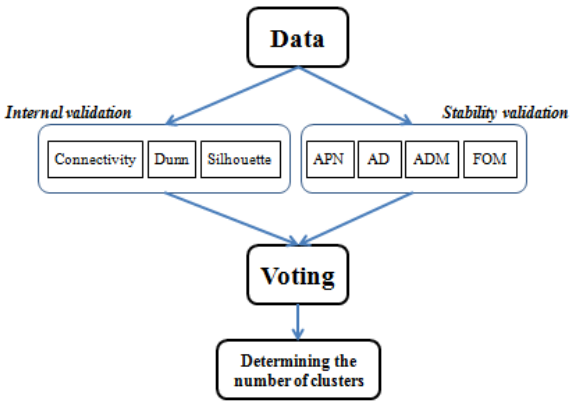

Fig. 1 Proposed ensemble method of cluster validation.

In this paper, we vote the clustering results from internal and stability validations.

3. Experimental Result

To verify improved performance of our research, we made experiments using objective data sets from UCI machine learning repository [23]. We used 'Abalone', 'Glass identification', and 'Yeast' in our experiment. Next table shows the numbers of instances and variables of the data sets.

Table 2: Data sets from UCI machine learning repository

Data set	***Number of instances***	***Number of variables***
Abalone	4177	8
Glass	214	9
Yeast	1484	8

These were used for evaluating performance of the clustering results by internal and stability validations. Next table shows the internal validation result of Abalone data set.

Table 3: Internal validation: Abalone

Clustering algorithm	***Internal validation***	***Number of clusters***	***Validation value***
hierarchical	Connectivity	2	3.8286
	Dunn	9	0.2233
	Silhouette	2	0.6268
K-means	Connectivity	2	1.0552
	Dunn	9	0.3994
	Silhouette	2	0.5537
PAM	Connectivity	2	0.0000
	Dunn	9	0.1648
	Silhouette	3	0.5288
CLARA	Connectivity	2	0.0000
	Dunn	9	0.0203
	Silhouette	10	0.5179

The occurred ratio of the number of clusters was 2 with 50%. So, using internal validation, we decided the number of clusters to 2in Abalone data set. Another clustering result of Abalone data set is shown in the following table.

Table 4: Stability validation: Abalone

Clustering algorithm	***Stability validation***	***Number of clusters***	***Validation value***
hierarchical	APN	2	0.0054
	AD	10	1.0658

	ADM	10	0.2531
	FOM	10	0.4465
K-means	APN	2	0.0707
	AD	10	1.0526
	ADM	4	0.2529
	FOM	10	0.4402
PAM	APN	2	0.0486
	AD	10	0.8956
	ADM	2	0.2176
	FOM	7	0.4353
CLARA	APN	2	0.0488
	AD	10	0.9560
	ADM	2	0.2220
	FOM	10	0.4331

This result was based on stability validation. We selected 10 as a proper number of clusters of Abalone data set. This result was different from the result of internal validation. So, we voted two clustering results as follow.

Table 5: Ensemble result of Abalone data

# of clusters	*Frequency*	*Ratio (%)*
2	12	42.86
3	1	3.57
4	1	3.57
7	1	3.57
9	4	14.29
10	9	32.14

From the ensemble result of internal and stability validation, we found that 2 was the number of clusters for Abalone data set. Next we made another experiment using Glass identification data for verifying our research.

Table 6: Internal validation: Glass identification

Clustering algorithm	*Internal validation*	*Number of clusters*	*Validation value*
hierarchical	Connectivity	2	3.8579
	Dunn	2	0.4934
	Silhouette	2	0.6343
K-means	Connectivity	2	8.8643
	Dunn	2	0.1904
	Silhouette	4	0.5879
PAM	Connectivity	2	21.2698
	Dunn	3	0.1244
	Silhouette	3	0.5822
CLARA	Connectivity	2	17.8778
	Dunn	4	0.1633
	Silhouette	4	0.5883

We decided 2 as the number of clusters for Glass identification using internal validation. The frequency ratio of 2 was 58.33%. Next table shows the clustering result of stability validation for Glass identification.

Table 7: Stability validation: Glass identification

Clustering algorithm	*Stability validation*	*Number of clusters*	*Validation value*
hierarchical	APN	2	0.0010
	AD	10	1.6310
	ADM	2	0.0186
	FOM	10	0.4963
K-means	APN	4	0.0368
	AD	10	1.3887
	ADM	4	0.2868
	FOM	10	0.4763
PAM	APN	2	0.0523
	AD	10	1.2717
	ADM	4	0.1987
	FOM	10	0.4187
CLARA	APN	3	0.0351
	AD	10	1.3614
	ADM	3	0.1555
	FOM	10	0.4284

In the result of stability validation, we selected 10 as the proper number of clusters. We got the ensemble result from the internal and stability validations in the following table.

Table 8: Ensemble result of Glass identification data

# of clusters	*Frequency*	*Ratio (%)*
2	10	35.71
3	4	14.29
4	6	21.43
10	8	28.57

So, we found that the optimal number of clusters was 2 with the frequency ratio of 35.71%. Last experimental data

set was Yeast in the paper. Next table shows the clustering result based on the internal validation measures.

Table 9: Internal validation: Yeast

Clustering algorithm	*Internal validation*	*Number of clusters*	*Validation value*
hierarchical	Connectivity	2	2.9290
	Dunn	2	0.5038
	Silhouette	2	0.6229
K-means	Connectivity	2	82.6579
	Dunn	7	0.0782
	Silhouette	4	0.2864
PAM	Connectivity	2	60.1119
	Dunn	6	0.0465
	Silhouette	2	0.3068
CLARA	Connectivity	2	60.3103
	Dunn	3	0.0560
	Silhouette	2	0.2971

We decided the number of clusters to 2 with 66.67% using internal measures. Next table shows the clustering result of Yeast data set by stability validation measures.

Table 10: Stability validation: Yeast

Clustering algorithm	*Stability validation*	*Number of clusters*	*Validation value*
hierarchical	APN	2	0.0000
	AD	10	0.3101
	ADM	2	0.0000
	FOM	10	0.0865
K-means	APN	6	0.0723
	AD	10	0.2495
	ADM	2	0.0320
	FOM	10	0.0850
PAM	APN	2	0.1443
	AD	10	0.2466
	ADM	3	0.0416
	FOM	10	0.0858
CLARA	APN	2	0.1256
	AD	8	0.2598
	ADM	2	0.0427
	FOM	10	0.0862

The number of clusters was decided to 10 with the frequency ratio of 43.75% using stability validation measures. So, we voted the results for our ensemble method. Next table shows the ensemble result for determining the number of clusters for Yeast data set.

Table 11: Ensemble result of Yeast data

# of clusters	*Frequency*	*Ratio (%)*
2	14	50.00
3	2	7.14
4	1	3.57
6	2	7.14
7	1	3.57
8	1	3.57
10	7	25.00

From the ensemble result, we decided that the number of clusters for Yeast data set was 2 with the frequency ratio of 50.00%.

4. Conclusions

We proposed an ensemble method for cluster validation. In this paper, we used the internal and stability measures for validating the cluster result. Also, we voted the clustering results from internal and stability validations. To verify the performance of proposed method, we made experiment using some objective data sets from UCI machine learning repository. This paper tried to find the number of clusters objectively using the ensemble approach by voting the clustering results. But, we need more general ensemble method for cluster validation. This is our future work.

References

[1] R. A. Johnson, and D. W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, 1992.
[2] J. Han, and M. Kamber, Data Mining Concept and Techniques, Morgan Kaufmann, 2001.
[3] G. Brock, V. Pihur, S. Datta, and S. Datta, “clValid: An R Package for Cluster Validation”, Journal of Statistical Software, Vol. 25, Iss. 4, 2008, pp. 1-20.
[4] J. L. DeRisi, V. R. Iyer, and P. O. Brown, “Exploring the metabolic and genetic control of gene expression on a genomic scale”, Science, Vol. 278, No., 5338, 1997, pp. 680-686.
[5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns”, Proc. Natl. Acad. Sci. USA, Vol. 95, No. 25, 1998, pp. 14863-14868.
[6] V. Bhattacherjee, P. Mukhopadhyay, S. Singh, C. Johnson, J. T. Philipose, C. P. Warner, R. M. Greene, and M. M. Pisano, “Neural crest and mesoderm lineage-dependent gene expression in orofacial development”, Differentiation, 2007.

[7] G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data", Bioinformatics, Vol. 18, No. 3, 2002, pp. 413-422.
[8] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data", Bioinformatics, Vol. 19, No. 8, 2003, pp. 973-980.
[9] L. Fu and E. Medico. FLAME, "A novel fuzzy clustering method for the analysis of DNA microarray data", BMC Bioinformatics, Vol. 8, No. 3, 2007, pp. 1-15.
[10] S. Sarmah, and D. K. Bhattacharyya, "An Effective Technique for Clustering Incremental Gene Expression data", International Journal of Computer Science Issues, Vol. 7, Iss. 3, No. 3, 2010, pp. 31-41.
[11] P. J. Rousseeuw, "Silhouette: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, Vol. 20, 1987, pp. 53-65.
[12] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data", Bioinformatics, Vol. 19, No. 4, 2003, pp. 459-466.
[13] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data", Bioinformatics, Vol. 17, No. 4, 2001, pp. 309-318.
[14]Y. Abofathy, and B. Zarei, "A New Approach for Optimal Clustering of Distributed Programs Call Flow Graph", International Journal of Computer Science Issues, Vol. 7, Iss. 4, No. 3, 2010, pp. 37-43.
[15] P. Agarwal, M. A. Alam, and R. Biswas, "Issues, Challenges and Tools of Clustering Algorithms", International Journal of Computer Science Issues, Vol. 8, Iss. 3, No. 2, 2011, pp. 523-528.
[16] M. K. Kerr and G. A. Churchill, "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments", Proc. Natl. Acad. Sci. USA, Vol. 98, No. 16, 2001, pp. 8961-8965.
[17] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data", Bioinformatics, Vol. 17, No. 4, 2001, pp. 309-318.
[18] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data", Bioinformatics, Vol. 19, No. 4, 2003, pp. 459-466.
[19] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation", Genome Res., Vol. 12, No. 10, 2002, pp. 1574-1581.
[20] I. Gat-Viks, R. Sharan, and R. Shamir, "Scoring clustering solutions by their biological relevance", Bioinformatics, Vol. 19, No. 18, 2003, pp. 2381-2389.
[21] N. Bolshakova, F. Azuaje, and P. Cunningham, "A knowledge-driven approach to cluster validity assessment", Bioinformatics, Vol. 21, No. 10, 2005, pp. 2546-2547.
[22] S. Datta and S. Datta, "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes", BMC Bioinformatics, Vol. 7, 2006, pp. 397.
[23] UCI machine learning repository, archive.ics.uci.edu/ml.

Sunghae Jun is associate professor in the department of statistics, Cheongju University, Korea. He received the B.S., M.S., and Ph.D. degrees in department of statistics from Inha University, Korea, in 1993, 1996. Also he took the doctor's degree in computer science and engineering from Sogang University, Korea, 2007. He worked in NCR as a data mining consultant from 2000 to 2001. He was a visiting scholar in department of statistics, Oklahoma State University, OK, USA from 2009 to 2010. His research fields are machine learning, evolutionary computing, and management of technology.