ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Comparison of Cluster Validation Indices with Missing Data

Marko Niemelä[1,2]*, Sami Äyrämö[1] and Tommi Kärkkäinen[1]†

1- Faculty of Information Technology, University of Jyvaskylä
PO Box 35, FI-40014 Jyväskylä, Finland

2- Niilo Mäki Institute
PO Box 35, FI-40014 Jyväskylä, Finland

Abstract.
Clustering is an unsupervised machine learning technique, which aims to divide a given set of data into subsets. The number of hidden groups in cluster analysis is not always obvious and, for this purpose, various cluster validation indices have been suggested. Recently some studies reviewing validation indices have been provided, but any experiments against missing data are not yet available. In this paper, performance of ten well-known indices on ten synthetic data sets with various ratios of missing values is measured using squared euclidean and city block distances based clustering. The original indices are modified for a city block distance in a novel way. Experiments illustrate the different degree of stability for the indices with respect to the missing data.

1 Introduction

In clustering, a given set of data is divided into subsets, clusters, such that observations in a cluster are similar to each other and dissimilar to observations in the other clusters. Even though the principle is simple, there exist multiple clustering approaches [1] of which the main groups are prototype-based and hierarchical clustering. Prototype-based algorithms, such as K-means [2], utilize error functions based on within-cluster distances, which then provide data partition with location estimates, e.g., the sample mean, as the cluster prototypes. K-medians is a robust variant of K-means algorithm, which does not assume spherically symmetric, normally distributed cluster shapes, but instead the variables can consist of discrete values with uniform quantization error [3]. Further, another property of K-medians is robustness against outliers since the breakdown point of the median is 50 %.

Prototype-based clustering typically requires the number of clusters, denoted by K, as an input parameter. Determining the correct number of clusters is a difficult task, because there are often more than one possible solutions to a clustering problem. The existing methods to estimate the number of clusters are based on, e.g., visual evaluation of clustering error [4], stability of the solution

*The work has been supported by the Academy of Finland from the project 311737 (DysGeBra)

†The work has been supported by the Academy of Finland from the projects 311877 (Demo) and 315550 (HNP-AI)

[5], and multiobjective evolutionary algorithms [6]. Cluster validation indices analyze the quality of clustering models by assessing compactness and separability of clusters with different values of K.

Internal cluster validation indices have been compared in recent studies. In [7], `kCE-index` was found to be the best performing index over 43 indices, being the only index able to validate successfully the single cluster data set, in which the other indices recommended higher numbers. In [8], `Wemmert-Gançarski` outperformed other indices when three distance measures and clustering approaches with 56 synthetic and 6 real world data sets were used. The study summarized different results for different indices. For some indices, the performances varied between different distances. In [9], `Silhouette` index was generally the best of 30 indices through a large number of experiments, including demanding data sets with high dimensionalities, noise, and overlapping clusters.

Despite the extensive comparisons of indices in the previous studies, none of them considered data sets with missing values. However, missing values are common in the real-world data. There could be a variety of reasons to explain missingness of variables, including measurement error, device malfunction, unanswered question, etc. Many clustering approaches are based on the assumption of complete data sets, therefore, such methods cannot be applied directly if some of the data values are missing.

In this work, the previous work especially in [7, 8] was continued by selecting the best performing indices to the comparison. The original indices based on euclidean distance were extended also for city block distance. The selected clustering methods and indices, presented in Section 2, were developed to be tolerant for missing values. Numerical results demonstrating the quality of indices are given and the main findings are discussed in Section 3.

2 Methods

The prototype-based clustering methods consist of an initialization step, in which an initial partition of the data is decided, and a local refinement step, in which the quality of the initial partition is improved by an iterative local search algorithm. Hence, in a general case, the following clustering error is minimized during the local search:

$$\mathcal{J}(\{\mathbf{c}_k\}) = \sum_{i=1}^{N} \min_{k=1,\dots,K} \|\mathbf{x}_i - \mathbf{c}_k\|_p^q = \sum_{k=1}^{K} \mathcal{J}_{p,k}^q = \mathcal{J}_p^q, \tag{1}$$

where $\{\mathbf{x}_i\}_{i=1}^N, \mathbf{x}_i \in \mathbb{R}^n$, is the given set of n-dimensional observations, N is the number of observations, and $\{\mathbf{c}_k\}_{k=1}^K$ are the obtained prototype vectors. l_p-norms to q-th power are utilized for different location estimates. The within-cluster error in cluster C_k, is denoted by $\mathcal{J}_{p,k}^q$ and the total residual error of a local minimizer of Eq. 1 is denoted by $\mathcal{J}_p^q$. By choosing $p = q = 1$ or $p = q = 2$, the error function for K-medians or K-means, respectively, are obtained. Note that if $q = 1$ it can be omitted from the notation.

In this study, a partial distance strategy for calculating distances is adopted from [10] since the data vectors may consist missing values. The idea is that the sum of differences of the known components are used and scaled to the missing components. The original method was developed for the l_2-norm, but a modified version for the l_1-norm is offered in the current study. Distances based on l_1 and l_2 norms read as $\hat{d_1}(\mathbf{x}, \mathbf{y}) = \frac{n}{\hat{n}} \sum_{j=1}^{\hat{n}} |(\mathbf{x})_j - (\mathbf{y})_j|$ and $\hat{d_2}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{n}{\hat{n}} \sum_{j=1}^{\hat{n}} ((\mathbf{x})_j - (\mathbf{y})_j)^2}$, respectively. $\hat{n}$ indicates the number of components that exist in both of the compared vectors. We assume that $\hat{n} > 0$. The modified version of Eq. 1 is required due to missing data. The new estimated clustering error, based on the partial distance strategy, is defined as $\hat{\mathcal{J}}_p^q = \sum_{i=1}^N \min_{k=1,\ldots,K} \hat{d}_p^q(\mathbf{x}_i, \mathbf{c}_k)$.

Internal cluster validation indices prefer both high within clusters similarity and between clusters separability. In this work, the measured within-cluster similarity is referred to as *Intra* and between-cluster separability as *Inter*. Low values are better for *Intra* and high values for *Inter*. The optimal solution is obtained by minimizing or maximizing the ratio of Intra and Inter measures.

The eight best performing incides from [7] in addition to WB-index (WB) [8] and Davies-Bouldin* [9] were compared in this study. All the indices, except Silhouette, are defined in Table 1. We presented general forms of reduced formulas, where constant terms or monotone functions have been omitted. The formulas are attempted to be minimized since *Intra* is divided by *Inter*. The clustering error is often used as *Intra*. Further, many indices tend to define *Inter* as the minimum distance between cluster prototypes. Distances between cluster prototypes and the whole data prototype are also commonly applied as *Inter* value. In addition, WB, Calinski-Harabasz, and kCE-index utilize penalization terms for a high number of clusters that were originally defined in the context of the squared euclidean distance. Initial experiments showed that these terms penalized too much while non-squared counterparts were used, therefore, square roots over terms were taken in these cases.

In Silhouette index, *Intra* is the average dissimilarity of $\mathbf{x}_i$ to all other points in the same cluster and *Inter* is the minimum average dissimilarity of $\mathbf{x}_i$ to all points in a different cluster. Silhouette index is defined as $\sum_{i=1}^{N} \frac{Inter(\mathbf{x}_i) - Intra(\mathbf{x}_i)}{\max(Intra(\mathbf{x}_i), Inter(\mathbf{x}_i))}$. Contrary to indices that use full prototypes for calculating an index value with missing data, Silhouette calculates distances between observations that are sometimes incomplete. Hence, the adopted distance calculation technique, presented in this study, is especially beneficial for Silhouette since there is always a higher risk that at least one of pairwise components is missing.

Ten synthetic data sets were used in the study. Four S[1] sets and two D[1] sets were selected from [11]. *Sim2D2*[2] and *Sim5D2*[2] data sets were selected from [7]. New similar *O200*[2] and *O2000*[2] data sets with a different number of observations were created for this study. Both O data sets consist of five clusters in total, one Gaussian and four Laplace distributed clusters. In addition, 10 % of uniformly

[1] http://cs.uef.fi/sipu/datasets/

[2] http://users.jyu.fi/~mapeniem/CVI/Data/

distributed noise was added to new data sets. D sets are 32 and 256 dimensional and the other presented data sets are two dimensional.

Table 1: Formulas of cluster validation indices.

Name	Intra	Inter	Formula
Calinski-Harabasz (CH)	$\hat{\mathcal{J}}_p^p$	$\sum_{k=1}^{K} n_k \lVert \mathbf{c}_k - \mathbf{m} \rVert_p^p$	$(\frac{K-1}{N-K})^{\frac{1}{3-p}} \times \frac{Intra}{Inter}$
Davies-Bouldin (DB)	$\frac{\hat{\mathcal{J}}_{p,k}}{n_k} + \frac{\hat{\mathcal{J}}_{p,k'}}{n_{k'}}$	$\lVert \mathbf{c}_k - \mathbf{c}_{k^*} \rVert_p$	$\frac{1}{K} \sum_{k=1}^{K} \max_{k \neq k'} \frac{Intra(k,k')}{Inter(k,k')}$
Davies-Bouldin* (DB*)	$\frac{\hat{\mathcal{J}}_{p,k}}{n_k} + \frac{\hat{\mathcal{J}}_{p,k'}}{n_{k'}}$	$\lVert \mathbf{c}_k - \mathbf{c}_{k^*} \rVert_p$	$\frac{1}{K} \sum_{k=1}^{K} \frac{\max_{k \neq k'} Intra(k,k')}{\min_{k \neq k^*} Inter(k,k^*)}$
Generalized Dunn (GD)	$\max \frac{\hat{\mathcal{J}}_{p,k}}{n_k}$	$\min_{k \neq k'} \lVert \mathbf{c}_k - \mathbf{c}_{k'} \rVert_p$	$\frac{Intra}{Inter}$
kCE-index (KCE)	$\hat{\mathcal{J}}_p^p$	1	$K^{\frac{1}{3-p}} \times Intra$
Pakhira-Bandyopadhyay-Maulik (PBM)	$\hat{\mathcal{J}}_p$	$\max_{k \neq k'} \lVert \mathbf{c}_k - \mathbf{c}_{k'} \rVert_p,$	$K \times \frac{Intra}{Inter}$
Ray-Turi (RT)	$\hat{\mathcal{J}}_p^p$	$\min_{k \neq k'} \lVert \mathbf{c}_k - \mathbf{c}_{k'} \rVert_p^p$	$\frac{Intra}{Inter}$
WB-index (WB)	$\hat{\mathcal{J}}_p^p$	$\sum_{k=1}^{K} n_k \lVert \mathbf{c}_k - \mathbf{m} \rVert_p^p$	$K^{\frac{1}{3-p}} \times \frac{Intra}{Inter}$
Wemmert-Gançarski (WG)	$\hat{d}_p(\mathbf{x}_i, \mathbf{c}_k)$	$\min_{k \neq k'} \hat{d}_p(\mathbf{x}_i, \mathbf{c}_{k'})$	$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \frac{Intra(\mathbf{x}_i)}{Inter(\mathbf{x}_i)}$

3 Experimental results and conclusion

Experiments were performed using MATLAB (R2015B, 64-BIT). Data sets were min-max scaled to a range of [-1, 1] before clustering and index value calculations. Incomplete data sets with varying numbers of missing values were created by removing data values completely at random from the existing test data sets. The clustering was repeated 100 times from random initial conditions of prototypes and the solution of the lowest local minima was selected as the final solution. The initialization was performed in an iterative manner such that K ranged from 2 to 20. More specifically, the obtained prototypes were saved for each K and these previously saved prototypes were utilized during the next initialization. The generalized version of K-means++ algorithm (see [8] for details) was used and therefore the next prototype was selected based on the calculated distances to the closest already selected prototypes such that the most distant point had the highest probability of being selected.

Table 2 shows the obtained results. Clearly, WG and Silhouette were generally the two best performing indices suggesting 64 and 63 correct solutions in total, respectively. Further, WG, KCE, and CH were the three best performing indices for the euclidean distance, giving 36, 33, and 33 correct solutions, respectively. In addition, Silhouette and WG were the two best ones for the city block distance, proposing 31 and 28 correct solutions, respectively. Regarding the stability of indices, WG showed to be the most stable, giving always nine correct solutions over ten data sets for the euclidean distance while the proportion of missing data was gradually increased from 0 % to 20 %. CH was the stable index for the city block. However, it only offered six correct solutions for each level of missing values. For the most of indices, especially for euclidean distance based indices, the high number of missing values has negative impact on the performance. As shown in Table 2, the whole clustering algorithm did not cause instability to the index results since only in four cases the correct number of clusters was not found after clustering with random initial prototypes, but only after using the known centers, given by the authors of the data sets, as initial prototypes in clustering.

This section provided results which were obtained when cluster validation indices were compared. Previous studies [7, 8] were continued by extending clustering methods and indices to city block distances and to handle missing values. Similarly to the previous studies, `WG`, `Silhouette`, and `KCE` were nominated to be the best performing indices in this study. All indices performed better with the euclidean distance compared to the city block distance. The used data sets are all continuous valued which may explain the better results with the euclidean distance. `Silhouette` produced almost identical results for these two distances and was the best index for the city block. Different stability patterns for the indices were shown in the study. `WG` was the most stable index, recommending nearly always the same numbers for clusters over the different levels of missing values. Future research direction is to use real-world data in experiments. Further testing is also needed with multidimensional data since all the indices offered always correct answers for *D32* and *D256* data sets.

References

[1] C. C. Aggarwal and C. K. Reddy. *Data clustering: algorithms and applications.* CRC press, 2013.

[2] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[3] M. Saarela and T. Kärkkäinen. Analysing student performance using sparse data of core bachelor courses. *JEDM-Journal of Educational Data Mining*, 7(1):3–32, 2015.

[4] R. L. Thorndike. Who belongs in the family. *Psychometrika*, pages 267–276, 1953.

[5] L. I. Kuncheva and D. P. Vetrov. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1798–1808, 2006.

[6] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. C. P. L. F. de Carvalho. A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(2):133–155, 2009.

[7] S. Jauhiainen and T. Kärkkäinen. A simple cluster validation index with maximal coverage. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2017*, pages 293–298, 2017.

[8] J. Hämäläinen, S. Jauhiainen, and T. Kärkkäinen. Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms*, 10(3), 2017.

[9] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[10] J. Gower. A general coefficient of similarity and some of its properties. 27:857–871, 1971.

[11] Q. Zhao and P. Fränti. Wb-index: A sum-of-squares based index for cluster validity. *Data & Knowledge Engineering*, 92:77–89, 2014.

Euc	***CH***	***DB***	***DB****	***GD***	***KCE***
Cit					
S1	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
	15 15 15 15	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
S2	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
	15 15 15 15	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
S3	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	4 4 **15** 4	**15 15 15 15**
	15 15 15 15	7 14 14 14	4 4 4 4	4 4 4 4	**15 15 15** 16
S4	**15 15 15 15**	14 14 14 17	13 13 13 13	4 4 4 4	**15 15 15 15**
	15 15 15 15	17 17 17 17	4 4 4 4	4 4 4 4	**15 15 15** 16
D32	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
	16 16 16 16	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
D256	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
	16 16 16 16	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
Sim2D2	**2 2 2 2**	**2 2** 20 19	**2 2 2 2**	**2 2 2 2**	**2 2 2 2**
	4 4 4 20	13 13 20 18	**2 2 2 2**	**2 2 2 2**	**2 2 2** 20
Sim5D2	3 3 3 3	3 3 3 3	3 3 3 3	3 3 3 3	**5 5 5 5**
	3 3 3 3	3 3 3 3	3 3 3 3	3 3 3 3	3 3 3 3
O200	**5** 20 20 20	**5 5 5** 20	**5 5 5 5**	4 4 4 4	20 20 20 20
	20 20 20 20	8 8 20 20	8 **5** 4 **5**	**5 5 5 5**	20 20 20 20
O2000	**5 5 5 5**	**5 5 5 5**	**5 5 5 5**	**5 5 5 5**	**5** 6 6 7
	6 12 13 20	6 6 6 **5**	4 4 4 **5**	4 4 4 **5**	1 6 6 14
Total	**9 8 8 8**	**8 8 7 6**	**8 8 8 8**	**6 6 7 6**	**9 8 8 8**
	6 6 6 6	**4 4 4 5**	**5 6 5 7**	**6 6 6 7**	**7 7 7 4**
	PBM	***RT***	***SIL***	***WB***	***WG***
S1	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
	15 15 15 15	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
S2	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
	15 15 15 15	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**	**15 15 15 15**
S3	5 5 4 4	4 4 4 4	**15 15** 2 2	**15 15 15** 16	**15 15 15 15**
	4 4 4 4	4 4 **15 15**	**15 15 15** 2	**15 15 15** 16	**15 15 15 15**
S4	4 4 4 4	13 13 10 10	**15 15 15** 3	**15 15 15** 20	**15 15 15 15**
	5 5 5 5	17 17 4 14	**15 15** 14 14	**15 15 15** 16	16 16 16 16
D32	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
	16 16 16 16	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
D256	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
	16 16 16 16	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**	**16 16 16 16**
Sim2D2	**2 2 2 2**	**2 2 2 2**	**2 2 2 2**	12 12 20 20	**2 2 2 2**
	2 2 2 2	**2 2 2 2**	**2 2 2 2**	4 20 20 20	**2 2 2 2**
Sim5D2	**5 5 5 5**	3 3 3 3	3 3 3 3	**5 5 5 5**	3 3 3 3
	5 5 4^+ 4^+	3 3 3 3	3 3 3 3	4 7 7 17	3 3 3 3
O200	**5 5 5 5**	**5 5 5 5**	**5 5 5 5**	20 20 20 20	**5 5 5 5**
	3 3 3 3	**5 5 5 5**	**5 5 5 5**	20 20 20 20	**5 5** 20 20
O2000	**5** 4 4 4	**5 5 5** 4	**5 5 5** 6	6 7 7 20	**5 5 5 5**
	3 3 3 3	4 4 4 **5**	**5 5** 6^+ 6^+	14 13 20 20	**5 5** 6 2
Total	**8 7 7 7**	**7 7 7 6**	**9 9 8 6**	**7 7 7 5**	**9 9 9 9**
	6 6 5 5	**6 6 7 8**	**9 9 7 6**	**6 6 6 4**	**8 8 6 6**

$^+$ Result can be corrected using the known centers as initial prototypes

Table 2: The determined number of clusters by cluster validation indices. The correct numbers are bolded. The results are given in four columns, one column for each percentage (0, 5, 10, and 20 %) of missing values.