ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

A Simple Cluster Validation Index with Maximal Coverage

Susanne Jauhiainen and Tommi Kärkkäinen

Department of Mathematical Information Technology
University of Jyvaskyla, Finland

Abstract. Clustering is an unsupervised technique to detect general, distinct profiles from a given dataset. Similarly to the existence of various different clustering methods and algorithms, there exists many cluster validation methods and indices to suggest the number of clusters. The purpose of this paper is, firstly, to propose a new, simple internal cluster validation index. The index has a maximal coverage: also one cluster, i.e., lack of division of a dataset into disjoint subsets, can be detected. Secondly, the proposed index is compared to the available indices from five different packages implemented in R or Matlab to assess its utilizability. The comparison also suggests many interesting findings in the available implementations of the existing indices. The experiments and the comparison support the viability of the proposed cluster validation index.

1 Introduction

Clustering is one of the most popular unsupervised techniques in machine learning and data mining to profile a given dataset. The purpose of clustering is to divide data into groups, clusters, such that members of a cluster are similar to each other and dissimilar to members in other clusters [5]. There exists various clustering approaches, like density-based clustering, probabilistic clustering, grid-based clustering, and spectral clustering [1], but the classical division of clustering methods is to distinguish hierarchical and prototype-based methods [2]. The methods are realized in different clustering algorithms, like k-means, kernel k-means, or agglomerative hierarchical clustering [3]. Details of the algorithms, e.g., how to initialize an iterative clustering procedure, vary as well.

Usually the creation of clusters and determination of the number of clusters are treated as separate problems. Analysis of the quality of clustering is referred as cluster validation and measures that can be used for indicating number of clusters are called Cluster VAlidation Indices (CVAI). For the prototype-based methods or with a statistical estimate to represent a subtree of the dendrogram in hierarchical clustering, one can further distinguish external and internal CVAIs. For an external index, the number of clusters is given by an oracle, for example, one can create the same number of clusters as the number of classes for a classification data, to study the geometric separability of different classes (especially, when initializing cluster prototypes as class means).

Here we consider internal CVAIs, which propose measures to determine the number of clusters in an unsupervised manner, using only information from the clustering solution. Moreover, we restrict ourselves to the most common prototype-based clustering algorithm, namely the k-means [4].

2 Methods

K-means and other iterative relocation algorithms accomplish the general purpose of clustering in two faces, where the initialization explores the data by locating the initial prototypes, and the refinement of prototypes, during the search phase, performs local exploitation. Hence, for N observations, one minimizes locally the following clustering error criterion

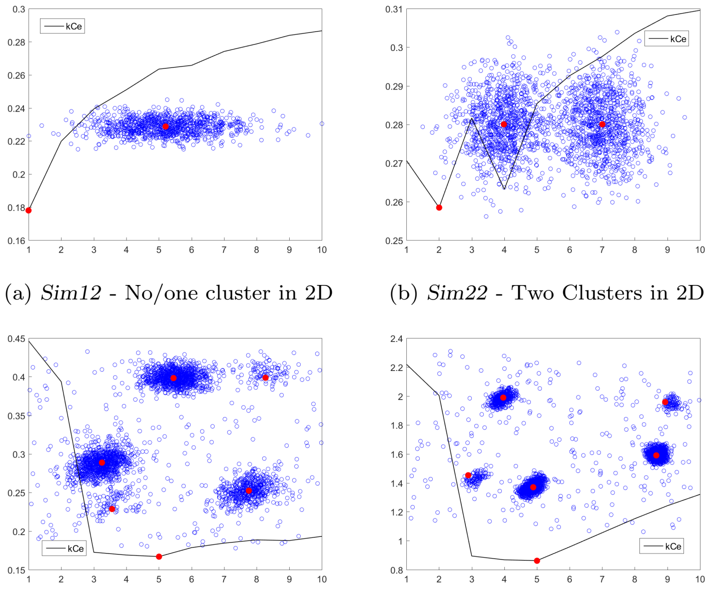
$$\mathcal{J}\{\mathbf{c}^k\} = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{C}_k} \|\mathbf{c}_k - \mathbf{x}_i\|^2 = \sum_{k=1}^{K} \mathcal{J}_k, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^n$ denotes a given data vector and $\mathbf{c}_k \in \mathbb{R}^n$ is the prototype, i.e., mean of the subset of data $\{\mathbf{x}_i \in \mathbf{C}_k\}$ in the kth cluster of size $|C_k|$, closest to the prototype. By $\mathcal{J}_k^*$ we refer to the kth *within-cluster* error and by $\mathcal{J}^*$ the whole residual of a local minimizer of (1).

The internal CVAIs analyse and compare the accuracy of the data representation with the prototypes and the dissimilarity of the prototypes. Hence, we measure and estimate the within-cluster and between-cluster separability, so according to the general purpose of clustering, we try to minimize the former and maximize the latter (or, equivalently, minimize the reciprocal). In this most common case, the minimum of CVAI indicates the number of clusters, although many existing indices also indicate this as their argument maximum (equivalently, as minimum of the reciprocal). Let us generally refer the within-cluster error as *Intra* and between-cluster error as *Inter*, so that a prototypical CVAI to be minimized reads as *Intra/Inter*. In the following, we will not provide precise depictions of the constructions of numerous CVAIs, but try to describe some common examples. Derivations, forms, and references to the original articles of the indices are introduced and covered in the descriptions of the tested packages (see Section 3).

`Ball-Hall` is an example of CVAI that only takes into account *Intra*, in the form of mean of the mean within-cluster errors $\mathcal{J}_k^*/|C_k|$. `Ray-Turi (RT)` and `BR` by Ristevski et al. (2008) are examples of CVAIs taking into account both *Intra* and *Inter*. In `RT`, *Intra* is computed as the mean residual $\mathcal{J}^*/N$, whereas in `BR` the sum of normalized errors over the clusters $\mathcal{J}_k^*/K$ is used. Both then compute *Inter* as the minimum of $c_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\|^2, i \neq j$. For `Davies-Bouldin (DB)`, we let $R_{ij} = (\mathcal{J}_i^*/|C_i| + \mathcal{J}_j^*/|C_j|)/c_{ij}$ ($c_{ij} \neq 0$), define $R_i = \max_{j \neq i} R_{ij}$, and set $\texttt{DB} = \frac{1}{K}\sum_{k=1}^{K} R_k$. Two similar CVAIs being maximized are `Calinski-Harabasz (CH)` and `generalized Dunn (GD)`, whose reciprocals read as follows: for 1/`CH` *Intra* $= (K-1)\mathcal{J}^*$ and *Inter* $= \sum_k |C_k|\|\mathbf{c}_k - \mathbf{m}\|^2$, where $\mathbf{m}$ refers to the whole data mean. For 1/`GD`, *Intra* is taken as the maximum of the mean $\mathcal{J}_k^*$s and *Inter* is defined similarly to `RT`.

Desired properties of CVAIs were given in [6], where it was required that no extra parameters should be needed, hierarchical data sets should be tolerated, relatively large data sets should be handled efficiently, and arbitrary dimensions should be supported. The original `DB` as described above is an example of a nontrivial index attempting to satisfy all these requirements.



(a) *Sim12* - No/one cluster in 2D

(b) *Sim22* - Two Clusters in 2D

(c) *Sim52* - Five clusters in 2D

(d) *Sim510* - Five clusters in 10D

Fig. 1: Illustrating the new cluster validation index `kCE`

2.1 The novel internal cluster validation index

We propose to use $\texttt{kCE} = k \times \mathcal{J}^*$ (k times the clustering error) as an internal cluster validation index. This suggestion is motivated and illustrated by the simulated datasets in Fig. 1, where (a)–(c) depict 2D datas and (d) two main scores of 10D data; along with the values of `kCE` (y-axis) for k =1–10 (x-axis). The intuitive argument is as follows: For k-means type of algorithm, the data mean is the prototype (marked in red in Fig. 1) for $k = 1$ with the corresponding value `1CE`. Therefore the proposed index has a maximal coverage, being able to propose one/no cluster as in (a). Then, when computing and testing the clustering accuracy for $k = 2, 3, \ldots$, the base case is that two prototypes should reduce the clustering error by 2 compared to the data mean, three prototypes by three etc. If adding a prototype does not pay off, then `kCE` is increasing as shown in Fig. 1. However, if some of the tested cases include k spherical subsets like in Figs. 1 (b)–(d), then the more rapid decrease of the clustering error indicates the number of clusters.

Hence, the suggested novel CVAI is closest to the `Ball-Hall` index, because we only need the clustering error for its computation. This makes it computationally maximally efficient, because no extra quantities like in typical CVAIs as described above need to be computed. Similarly, also enlargements of the proposed index to both robust and weighted [7, 8] prototype-based clustering algorithms are straightforward.

3 Experimental results

Next we present results from a comparison of CVAIs. Our goal here is twofold: (i) a large set of available implementations of CVAIs on two popular platforms, R (ver3.3.0) and Matlab (R2014a), are tested, and (ii) the viability of the proposed index kCE is assessed.

Many of the propositions of new CVAIs include an experimental evalution of multiple indices, typically concluding the proposed index as the best one. Moreover, eight CVAIs were compared in [9]. Most suggested the correct number of clusters with 5% additional noise, different densities, and skewed distributions, but only three were able to recognize closed subclusters. Sdbw was the only CVAI that suggested the right number of clusters for all data sets. Often no single CVAI dominates in experiments. This was the conclusion in [10], where a comparison of 30 different indices with 720 synthetic and 20 real datasets was made. In this study, Silhouette was nominated as the best general index.

Here the comparison is based on 12 synthetic data sets. We used the four 2D *S* sets with 15 centers and the four higher dimensional *Dim* sets with 16 centers from http://cs.uef.fi/sipu/datasets/. Other simulated data sets included *Sim12*, *Sim22*, *Sim52*, and *Sim510* as illustrated in Fig. 1. Especially the latter self-generated noisy datasets contained clusters and subclusters of different densities and sizes, similarly to the difficult sets in [9]. More precisely, *Sim5D2* and *Sim5D10* have five clusters, of which two are harder to detect being smaller, more sparse and situated next to a bigger cluster, with 10% noise.

The three R packages, NbClust[1] (P1), cclust[2] (P2) and clusterCrit[3] (P3) were tested with 30, 15, and 27 implemented CVAIs, respectively. In Matlab's function evalclusters (P4) we applied three CVAIs (Davies-Bouldin, Calinski-Harabasz, and Silhouette), while 10 CVAIs from the Cluster Validity Analysis Platform, CVAP (P5), downloaded from Matlab's file exchange center, were tested. Apart from NbClust package, clustering was done in Matlab using the k-means algorithm. The k-means algorithm was repeated 1000 times by selecting the solution with smallest clustering error as the final result. We applied min-max scaled into $[-1, 1]$ before clustering and index computations, which were tested for k =2–20 and additionally for $k = 1$ for kCE. We noticed that the clustering in Matlab improved the performance of the CVAIs in the two R-packages P2 and P3 compared to R's own implementation.

Altogether 43 different CVAIs were tested, many with several implementations in different packages. The results were highly varying, even between the different implementations of the same CVAI. The results for the 17 best performing CVAIs in all five packages are presented in Table 1. Hyphen in the table means that there was no implementation of that CVAI in the package or that the calculation failed (producing NaN, inf etc.). Row "Correct" measured the difficulty of a data by counting the number of correctly determined number of

[1] https://cran.r-project.org/web/packages/NbClust/NbClust.pdf

[2] https://cran.r-project.org/web/packages/cclust/cclust.pdf

[3] https://cran.r-project.org/web/packages/clusterCrit/clusterCrit.pdf

clusters with at least one implementation. The number of correct propositions from an implementation of an index is given in column "Correct", where next to the last shell provides the median of correct propositions over all packages. The last row in Table 1 contains results of kCE.

4 Conclusions

NbClust was distinctly the worst package, probably because the clustering was done in R with fewer repetitions. The CVAIs in P4 were clearly the best performing with the median of correct propositions being nine out of 12 datasets. This is no suprise, since P4 is a commercial package. Rest of the packages worked quite poorly, proving mostly nonconsistent suggestions. Of course, this might be because of the construction of the CVAI or due to a poor implementation.

Four available indices, namely Calinski-Harabasz, Silhouette, Pakhira-Bandyopadhyay-Maulik (PBM), and Wemmert-Gançarski, were recognized as the best ones, proposing correct number of clusters for nine out of the 12 datasets, with at least one implementation. Sdbw worked only for similar datasets as in [9]. The new index kCE was the only one that worked for all datasets, being able to recognize the case of only one/no cluster where other indices suggested large numbers. kCE also worked for the challenging *Sim52* and *Sim510* datasets. As the main limitation, all the tested clusters were almost spherical, so the results might not apply to more complex data.

References

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

[2] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[3] Mohammed J Zaki and Wagner Meira Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms.* Cambridge University Press, 2014.

[4] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[5] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[6] David L Davies and Donald W Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

[7] M. Saarela and T. Kärkkäinen. Weighted Clustering of Sparse Educational Data. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2015*, pages 337–342, 2015.

[8] M. Saarela and T. Kärkkäinen. Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data. In *Proceedings of the 8th International Conference on Educational Data Mining*, pages 156–163, 2015.

[9] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916. IEEE, 2010.

[10] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, JesúS M PéRez, and IñIgo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Table 1: Numbers of clusters suggested by the CVAIs

P1,P2,P3 P4,P5	*S1*	*S2*	*S3*	*S4*	*D32*	*D64*	*D128*	*D256*	*Sim52*	*Sim510*	*Sim22*	*Sim12*	Correct
Davies Bouldin	**15**,**15**,17 **15**,**15**	14,**15**,16 **15**,14	10,**15**,**15** **15**,**15**	14,17,14 14,**15**	15,**16**,**16** **16**,17	20,**16**,**16** **16**,19	18,**16**,**16** **16**,18	18,**16**,**16** **16**,18	3,3,4 3,3	3,3,3 3,2	**2**,16,7 **2**,14	18,16,18 20,15	2,7,5 **8**,3
Calinski Harabasz	16,**15**,**15** **15**,**15**	14,**15**,**15** **15**,**15**	19,**15**,**15** **15**,**15**	**15**,**15**,**15** **15**,**15**	17,**16**,**16** **16**,**16**	19,**16**,**16** **16**,**16**	18,**16**,**16** **16**,**16**	17,**16**,**16** **16**,**16**	3,3,3 3,3	3,3,3 3,4	**2**,**2**,**2** **2**,**2**	6,6,6 6,6	2,**9**,**9** **9**,**9**
Silhouet	**15**,-,**15** **15**,**15**	14,-,**15** **15**,**15**	**15**,-,**15** **15**,**15**	**15**,-,14 **15**,**15**	15,-,- **16**,**16**	19,-,- **16**,**16**	18,-,- **16**,**16**	17,-,- **16**,**16**	3,-,3 3,3	3,-,3 3,4	**2**,-,**2** **2**,**2**	17,-,18 18,18	4,-,4 **9**,**9**
Hartigan	7,20,**15** -,2	14,20,4 -,2	4,20,4 -,2	3,20,3 -,2	15,**16**,**16** -,2	18,**16**,**16** -,2	18,**16**,**16** -,2	17,**16**,**16** -,2	3,20,3 -,2	3,2,3 -,3	4,16,4 -,**2**	3,16,3 -,2	0,4,**5** -,1
Dunn	9,-,**15** -,**15**	9,-,**15** -,**15**	5,-,20 -,4	10,-,**15** -,4	3,-,**16** -,**16**	5,-,**16** -,**16**	5,-,**16** -,**16**	8,-,**16** -,**16**	3,-,3 -,3	19,-,19 -,4	18,-,20 -,**2**	19,-,13 -,3	0,-,**7** -,**7**
Cidx	17,-,**15** -,20	**15**,-,**15** -,20	20,-,20 -,20	16,-,20 -,20	15,-,**16** -,-	19,-,**16** -,20	18,-,14 -,20	17,-15 -,-	17,-,3 -,20	4,-,7 -,2	3,-,20 -,20	17,-,20 -,20	1,-,**4** -,0
Rubin	**15**,**15**,**15**	14,**15**,**15**	19,**15**,**15**	14,**15**,**15**	15,3,**16**	19,3,13	18,3,19	17,3,19	3,19,19	3,17,17	4,12,12	17,15,15	1,4,**5**
Ray Turi	-,-,**15**	-,-,**15**	-,-,4	-,-,13	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,3	-,-,3	-,-,**2**	-,-,6	-,-,**7**
Sdbw	**15**,-,**15**	19,-,**15**	20,-,2	19,-,2	20,-,-	20,-,-	20,-,-	20,-,-	**5**,-,3	20,-,-	20,-,**2**	20,-,2	2,-,**3**
GenDunn	-,-,**15**	-,-,**15**	-,-,20	-,-,**15**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,2	-,-,19	-,-,20	-,-,13	-,-,**7**
Gamma	-,-,**15**	-,-,**15**	-,-,20	-,-,20	-,-,**16**	-,-,**16**	-,-,14	-,-,15	-,-,**5**	-,-,7	-,-,20	-,-,20	-,-,**5**
G+	-,-,**15**	-,-,**15**	-,-,20	-,-,20	-,-,**16**	-,-,**16**	-,-,14	-,-,15	-,-,**5**	-,-,7	-,-,20	-,-,20	-,-,**5**
PBM	-,-,**15**	-,-,**15**	-,-,5	-,-,4	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**5**	-,-,**5**	-,-,**2**	-,-,4	-,-,**9**
WemGan	-,-,**15**	-,-,**15**	-,-,**15**	-,-,**15**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,3	-,-,3	-,-,**2**	-,-,20	-,-,**9**
Xie Beni	-,-,**15**	-,-,**15**	-,-,20	-,-,16	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,3	-,-,11	-,-,20	-,-,3	-,-,**6**
CXu	-,**15**,-	-,**15**,-	-,16,-	-,16,-	-,2,-	-,2,-	-,2,-	-,2,-	-,**5**,-	-,19,-	-,**2**,-	-,3,-	-,-,**4**
Ssi	-,13,-	-,**15**,-	-,6,-	-,4,-	-,**16**,-	-,**16**,-	-,**16**,-	-,**16**,-	-,3,-	-,3,-	-,13,-	-,3,-	-,**5**,-
Correct	16/17	16/17	5/17	7/17	15/17	14/17	11/17	11/17	5/17	1/17	10/17	0/17	1.5,5,5 9,5

kCE	**15**	**15**	**15**	**15**	**16**	**16**	**16**	**16**	**5**	**5**	**2**	**1**	**12**