Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)

doi:10.3233/FAIA210140

Towards Expert-Inspired Automatic Criterion to Cut a Dendrogram for Real-Industrial Applications

Shikha SUMAN [a,1], Ashutosh KARNA [a] and Karina GIBERT [a]

[a] *Knowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Centre, Universitat Politècnica de Catalunya*

Abstract.
Hierarchical clustering is one of the most preferred choices to understand the underlying structure of a dataset and defining typologies, with multiple applications in real life. Among the existing clustering algorithms, the hierarchical family is one of the most popular, as it permits to understand the inner structure of the dataset and find the number of clusters as an output, unlike popular methods, like k-means. One can adjust the granularity of final clustering to the goals of the analysis themselves. The number of clusters in a hierarchical method relies on the analysis of the resulting dendrogram itself. Experts have criteria to visually inspect the dendrogram and determine the number of clusters. Finding automatic criteria to imitate experts in this task is still an open problem. But, dependence on the expert to cut the tree represents a limitation in real applications like the fields industry 4.0 and additive manufacturing. This paper analyses several cluster validity indexes in the context of determining the suitable number of clusters in hierarchical clustering. A new Cluster Validity Index (CVI) is proposed such that it properly catches the implicit criteria used by experts when analyzing dendrograms. The proposal has been applied on a range of datasets and validated against experts ground-truth overcoming the results obtained by the State of the Art and also significantly reduces the computational cost.

Keywords. Hierarchical Clustering, Cluster Validity Indices, Calinski-Harabasz index, Dendrogram

1. Introduction

Hierarchical clustering is a powerful technique that very well addresses the challenge of discovering the underlying structure by creating a hierarchy of data partitions into smaller object groups from top to bottom. This is represented in a tree diagram, called *dendrogram*. The dendrogram provides a visual trace of the whole clustering process that the clustering experts can inspect manually and decide the number of clusters in the dataset. There are two clear advantages of this approach. Firstly, the expert takes an

[1] Corresponding Author: Shikha Suman, Knowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Centre, Universitat Politècnica de Catalunya, Catalonia, Spain; E-mail: shikha.suman@estudiantat.upc.edu

informed decision after inspecting the dendrogram closely. Secondly, it avoids running multiple runs of CVI's, such as *silhouette* [1], *gap-statistic* [2] as in the case of flat clustering methods which aims to find the best number of clusters corresponding to the local maxima.

The motivation behind this research comes from a wider research project of developing an *Intelligent Decision Support System* for *Industry 4.0* applications. The proposed system is expected to monitor the performance of 3D printers in real-time through sensor data. In [3], the authors describe a hierarchical clustering-based method to profile sensor data from 3D printing. A key challenge in this system is to find appropriate print profiles from the sensor data with no prior knowledge about the cluster formation or the number of clusters itself. Hence, the research discussed in this paper directly helps in building an automated solution to find the number of clusters similar to what the human experts would find using a dendrogram manually.

The previous work by Karna *et al.* [4], however, shows that the original *Calinski-Harabasz* index has several instances where the number of clusters disagrees with what human experts suggested. Although proposed $\Delta_{K_{cond}}$ criterion [4] improves performance but still requires further fine-tuning to correctly match with experts' criterion.

Hence, the goal of this paper is to assess the reasoning behind the disagreement between the human-criterion and the algorithmic method. To contribute to the research, a complete methodology is discussed with a new CVI for detecting the number of clusters automatically in hierarchical clustering. Its performance on 100 samples from a real-life dataset is also discussed in brief. The rest of the paper is thus structured as follows. A brief survey of related literature is presented in Section §2, followed by a formal definition of the research problem in Section §3. Two new CVI's are proposed in Section §4 and the corresponding methodology is discussed in detail in Section §5. A summary of the experimentation results is discussed in Section §6. The authors conclude the paper in Section §7 and discuss the future lines of research and use cases.

2. Literature Survey

Some strategies to determine the number of clusters in a hierarchical clustering are cross-validation, resampling, and finding the *knee* or *elbow* of an error curve. The cross-validation approach estimates the best number of clusters by partitioning the data into v parts and iteratively evaluate a cluster validity criterion developed on $v-1$ parts on the v^{th} part. However, the approach requires extensive computation which limits its applicability when the data becomes huge and results are expected quickly. In [5], Overall and Magee presented a replication-based stopping rule in which a replication defined by higher-order clustering helps identify the distinct underlying populations (clusters) in a multidimensional space. The resampling-based methods require drawing several bootstrapping samples from the parent distribution and this becomes infeasible as the size of the dataset grows and is not suitable when the size of the data is huge and time complexity is really important. Finding the number of clusters by optimizing a CVI curve and identifying the local maxima (or minima) at the *knee* is also used in a variety of situations. Sevilla *et al.* in [6] reviews several CVI's and how they associate with the data topology. *Gap-statistic* proposed by Tibshirani *et al.* [2] is another CVI. It tests the hypothesis that the model has a single cluster ($K = 1$) and tries to reject it with an alternative hypothesis that ($K > 1$). In [7], the authors used *clustering-gain* as a metric for finding an optimal number of clusters in hierarchical clustering. The *clustering gain curve* is designed to discover the distinct clusters when the intra-cluster similarity is the maximum and the inter-cluster similarity is minimum. In [8], Zhou *et al.* proposes a new CVI, called *CSP* (compact-separation-proportion) based on the idea of nearest neighbour. The optimal number of clusters is estimated corresponding to the maximum average value of the *CSP* index. The *Calinski-Harabasz* index [9] is one of the most common and often regarded as the best CVI to determine the number of clusters in hierarchical clustering. Milligan in [10] conducted an extensive experiment on thirty different CVI's and concluded the *Calinski-Harabsz* to be the most consistent one. In the recent work by Karna *et al.* in [4], the authors performed empirical analysis on several real-life datasets and presented an improved CVI, called, $\Delta_{K_{cond}}$, that maximizes the difference of successive *Calinski-Harabasz* indices over a range of K clusters ($k = 1, 2, ...K$) However, many instances were seen where this proposed criteria did not comply with the experts' determined number of clusters using dendrogram.

In [11] a proposal to find the right number of clusters in a big data environment is provided by Luna *et al.* that consists of two clustering validity indices handling a large amount of data in low computational time. The idea of reasoning over the heights of the nodes of the dendrogram has been explored by some authors, but none provides a simple and computationally cheap criterion that can suitably match what experts do in real practice.

To the extent of interpreting cluster patterns, several works relevant for this research are studied. In [12], the authors present an approach to interpret cluster patterns in real datasets. Gibert *et al.* [13, 14] the distance for clustering complex datasets with messy data is presented. In [15] this is generalized to include semantics variables. These metrics will be introduced when prior knowledge on the 3D printing problem enters into the system. In [16] the dynamics of the system is introduced to see how clusters evolve along time.

3. Research Problem

Let us consider a multivariate numerical dataset, with the information about a set I of N *k-dimensional* objects as $i_1, i_2, ...i_N$. Thus the goal of a hierarchical clustering is to partition I into a sequence of nested partitions P_k (k=2, 3, ...K= *N-1*).

$$P_k = \{C_{k_1}, C_{k_2}, ...C_{k_k}\}; k = 1, 2, ...N-1 \tag{1}$$

where C_{k_k} represents the k^{th} cluster of the P_k partition of I.

The successive P_k are composed of disjoint clusters covering I. Thus, the dendrogram maps into the sequence $P_1, P_2, ..., P_{N-1}$ and $\forall k \in (2, 3, ...N-1), P_k$ is nested in P_{k-1} so that one of the clusters of the P_{k-1} subdivides in two in the P_k.

The objective of this paper is to develop an automatic criterion to identify the most appropriate P_k partition that divides the dataset I into k clusters such that the outcome is closest to what the human experts would achieve using the visual method. The main optimization criterion in this method is to find the value of k that optimizes the difference between the homogeneity of clusters and distinguishability among them. The more homogeneous and distinguishable the clusters are, the better is the partition.

4. Research Proposal

In theory, the number of clusters obtained by using *Calinski-Harabasz* method should match with the number of clusters deduced by the experts looking at the dendrogram. In [4], 5 different criteria based on *Calinski-Harabasz* index are proposed and evaluated to this purpose. In this research, all of them are evaluated over several datasets to see if they approach sufficiently well the criteria used by experts in visual inspection. In practice, a human expert usually decides the best cut of the dendrogram where the branches have longer gaps between nodes (regarding height), and each branch below the horizontal cut of the dendrogram results in a separate cluster. Underperformance has been detected on all these criteria and will be discussed in the application section. Experts choose the height representing the biggest disruption on the distinguishability. Following this intuition, two new criteria are presented in this paper to find the number of clusters based on the height of different nodes of a dendrogram.

Let $h_k, k \in 1:N-1$ be the height of node k in a given dendrogram built over I. The values of h keep the value of the distance between clusters merged at each node of the dendrogram. These values directly depend on the linkage method used in the hierarchical process that generated the dendrogram.

I Δ_H **criterion**: The Δ_H criterion maximizes the gap of linkage height between two consecutive nodes of the dendrogram, starting from the root of the tree. Mathematically, the criterion can be defined as in eq 2.

$$K^*_{\Delta_H} = \underset{2 \le k \le K}{\operatorname{argmax}}(\Delta_{H_k}); k \in (2,3,..K-1) \tag{2}$$

, where

$$\Delta_{H_k} = h_k - h_{k+1}; k \in (2,3,..K-1) \tag{3}$$

As it will be seen later in this paper, experimental results elicited that Δ_H criterion underperforms where the best cut of the tree is 2 clusters, as experts apply a heuristic in these cases. For this reason, a knowledge-based heuristic is introduced and a second criterion is proposed.

II $\Delta_{H_{cond}}$ **criterion**: The $\Delta_{H_{cond}}$ incorporates the heuristic that in some cases the experts skip a best cut in 2 clusters. This does not happen when the second-best cut is much closer to the bottom of the tree. This notion is represented through a ratio between the heights of the two highest nodes of the dendrogram, represented by h_{root} and h_2 respectively. Mathematically, this can be defined as eq 4.

$$K^*_{\Delta_{H_{cond}}} = \begin{cases} K^*_{2\Delta_H} & \text{if } K^*_{\Delta_H} = 2 \text{ and } (h_2/h_{root}) > 1/3 \\ K^*_{\Delta_H} & \text{otherwise} \end{cases} \tag{4}$$

where $K^*_{\Delta_H}$ and $K^*_{2\Delta_H}$ are the maximum and second maximum of Δ_K criterion. This introduces the flexibility to even consider 2 as the best clustering solution but only when the height of the root of the dendrogram (that results in two clusters) is at least thrice highest than the height of the second-highest node of the tree. Fig 1 and fig 2 help understand the distinction between the two scenarios. Details of these figures and the underlying logic will be clarified in section §6.

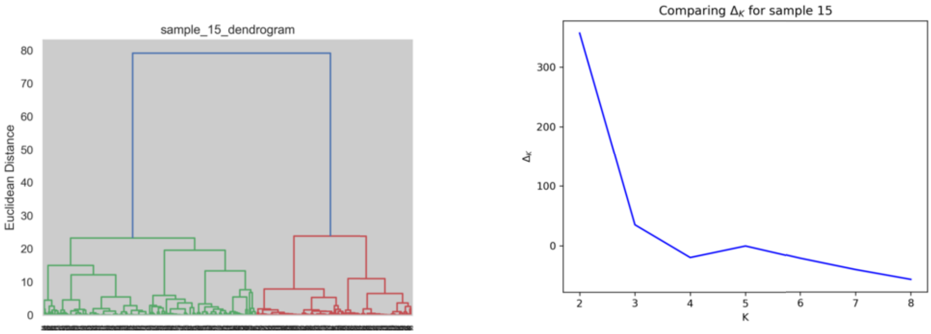

Figure 1. (a) Dendrogram for sample 15; (b) Δ_K curve for sample 15

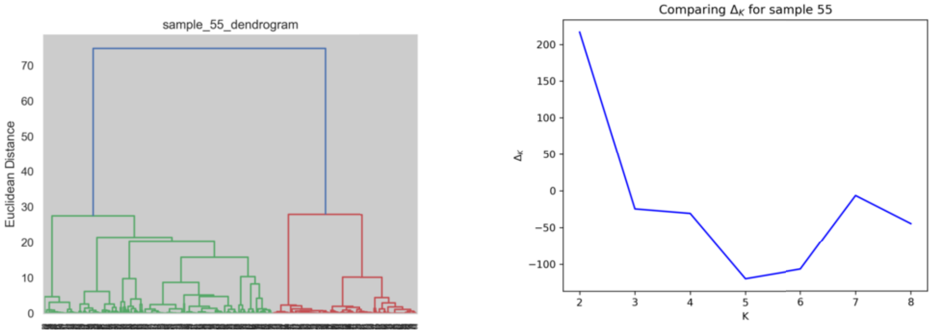

Figure 2. (a) Dendrogram for sample 55; (b) Δ_K curve for sample 55

The procedure to compute Δ_H and $\Delta_{H_{cond}}$ criteria are summarised as follows:

i For a dataset I containing N objects, compute the pairwise distance matrix.

ii Perform the hierarchical clustering I and build the corresponding dendrogram τ.

iii Obtain a list L of the nodes of τ sorted by descending height of nodes in descending order such that $H = h_{root} > h_2 > h_3 > .....h_{n-1}$.

iv Fix the maximum depth of nodes ($K \in 2 : N-1$) in the dendrogram to be considered to determine the number of clusters. Since, the bottom part of the dendrogram consists of nodes extremely close to each other, the optimal line of cut is placed in the top portion of the tree. Hence, in general, $K < N/2$ and this helps save half of the computations. The parameter K can heuristically be chosen (say, $K < N/2$). Return the first K nodes in L with their corresponding heights, namely (L', h'), in order to make a decision.

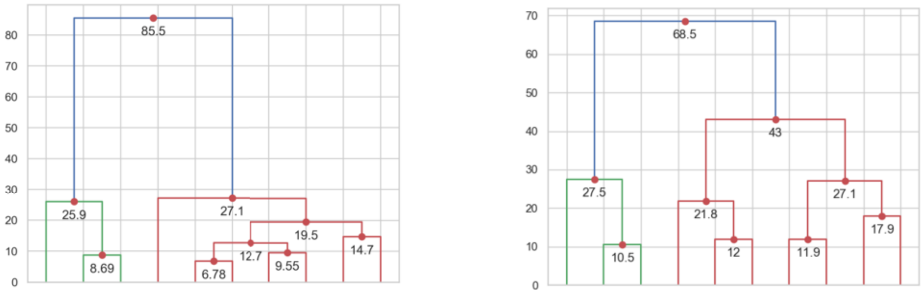

Figure 3. (a) Dendrogram of sample 10; (b) Dendrogram of sample 31

v Apply the criterion to the pair (L', h').

The approach can be verified by visualizing an annotated top part of τ with the height of top K nodes from the root of the tree. This step is to be carried out only to compare the values recommended by the criteria against the human-assigned one. An illustrative example can be seen in fig 3.

5. Research Methodology

This research evaluates 5 different CVI including the original *Calinski-Harabasz* index (denoted as M_K),its two variants (Δ_K, $\Delta_{K_{cond}}$) (as proposed in [4]) and the two new proposed criteria based on the inner morphology of the dendrogram, namely, Δ_H and $\Delta_{H_{cond}}$. These criteria are all applied to S datasets ($s \in (1,2,..S)$), all with a same number of objects N. For each dataset the dendrogram is obtained by using any hierarchical clustering method. The first K cuts of the dendrogram are obtained $K, k \in (2,3,...N/2)$. Given a CVI, f, $f \in (M_K, \Delta_K, \Delta_{K_{cond}}, \Delta_H, \Delta_{H_{cond}})$, a matrix $\chi_{f_s,k}$ is created where $\chi_{f_s,k}$ is the value of f for P_k of sample s.

Our main goal is to develop a criterion that approaches the number of clusters given by human experts. The proposed strategy is as follows:

I Let us consider a total of S dendrograms and the corresponding reference data, all of the same size.

II Subject each sample to hierarchical clustering (in this work with *Euclidean* distance and *Ward's* method), and obtain the dendrogram ($\tau_s, s \in (1,2,...S)$).

III For each τ_s, obtain P_k ($k \in 2...K = 9$) and compute the following for each k:

i Compute the matrix χ for the *Calinski-Harabasz* index ($\chi_{f_s,k} = M_{s,k}$)

ii Compute

$$\Delta_{s,k} = M_{s,k} - M_{s,k+1}, s \in 1,2,..S, k \in 2,3,..K-1 \qquad (5)$$

IV Let $K_{s,f}, f \in (M_k, \Delta_k, \Delta_{k_{cond}}, \Delta_H, \Delta_{H_{cond}}), s \in (1,2,..S), k \in (2,3,...9)$ be the returning number of clusters by each of the criteria f respectively.

V Get experts' assistance in providing the number of clusters from the dendrogram ($\tau_s, s \in (1,2,...S)$). Let $E_s, s \in (1,2,..S)$ be the number of clusters provided by the human experts for case $s \in (1,2,...S)$, by visual inspection of τ_s.

VI Compare the algorithmically determined number of clusters ($K_{s,f}$ against E_s) and assess the quality of f for each case $s_{\in}(1,2,...S)$

$$\varepsilon_{s,f} = |K_{s,f} - E_s| \tag{6}$$

VII Build a table of frequencies of $\varepsilon_{s,f}$ and the associated bar-charts and analyze the cases of largest mismatches.

VIII Take a representative case with $\varepsilon_{s,f} > 0$, visualize the dendrogram, get the CVI-proposed number of clusters and the one proposed by the experts, and try to understand the reason for the discrepancy by analyzing the dendrogram. Use as many cases as required until a comprehension of the failure of the criterion emerges. Use the results of this analysis to perform a modification on the CVI and evaluate the impact on the performance.

IX Compute the accuracy for a criteria f over S cases, denoted as A_f as follows:

$$A_f = \frac{card\{|K^*_{f,s} - E| = 0\}}{S}; s \in 1,2,..S \tag{7}$$

where $K^*_{f,s}$ denotes the optimal value of number of clusters using f^{th} criterion over s samples.

X Compare accuracy A_f over all of the candidate criteria. The winner criterion is the one that maximizes the accuracy.

6. Application

The research proposal has been evaluated on $S = 100$ real-life data samples obtained from an *Industry 4.0* process. Each dataset is of size N=500 and 10 numerical variables are used. *Euclidean* distance and *Ward* method have been used to cluster the samples. An ample mix of different types of morphological structures in their dendrograms is provided. In particular, it might be interesting to draw S random samples of a single reference dataset I, all of the same size, without replacement.

Following the methodology presented in 5, the dataset χ is built with 100 *Calinski-Harabasz* index curves for a range of $K = 8$ clusters (each curve representing a dendrogram with a different topology as shown in fig 4.

It can be seen that different morphologies of dendrograms correspond to different patterns of curves. Fig. 1 and fig. 2 show how the pattern of the Δ_K curve seems to be associated with the good or bad performance of the criterion, and also with different morphology of the reference dendrogram itself. Fig. 1 shows Δ_K curve monotonically decreasing with the maxima occurring at 2 ($K^*_\Delta = 2$), and the same is also evident from the dendrogram as well ($E_{15} = 2$), resulting in accurate match ($\varepsilon_{15,\Delta_K} = 0$). On the contrary, in the case of fig.2, while the local maximum occurs at 2 ($K^*_\Delta = 2$), the experts skip 2 as the best solution and rather suggests 7 clusters ($E_{55} = 7$) and results in a big mismatch ($\varepsilon_{55,\Delta_K} = 5$) which in fact, is the second local maximum.

Hierarchical clustering was performed on the dataset of Δ_K curves of all 100 samples in order to find groups of similar dendrogram morphologies together for detailed analysis. Seven distinct classes are identified with the clustering exercise and samples falling in different classes are analyzed independently. It is observed that the practice of skipping the maxima when (K=2) and switching to the second optimal value of k is quite common among human experts. This is due to an implicit exercise that cutting the tree into 2 clusters leads to a dichotomous solution which is rarely useful enough for further decision making. And thus, an implicit clustering rule often tends to shift to second-best solution where more than 2 clusters exist. This reasoning can also be seen when the authors in [4], proposed $\Delta_{K_{cond}}$ criterion. In a particular case of sample-55 (fig 2), $\Delta_{K_{cond}}$ does provide 7 as the number of clusters matching correctly with the experts. However, this is not always true and while analyzing the classes from the previous clustering exercise, several instances can be seen where this condition disagrees with the experts. Hence, it can be concluded that the *Calinski-Harabasz* criteria either in its original form or the variants as proposed in [4], do not truly capture the structure of the dataset, whereas the experts make a decision based on the vertical gaps in the dendrogram. This leads us to our proposal of using the height of nodes in a dendrogram to decide the best number of clusters.

The error distribution after applying the Δ_H criterion on all 100 samples can be seen in Table 1. It is clear that in most of the cases, $K^*_{\Delta_H} = 2$, that implies that in 85% of the cases, the algorithm determined 2 as the best cluster while the expert determined otherwise. This is similar to the case of *Calinski-Harabasz* index (M_K) which differs from the experts as guessed. This is aligned with the experts' judgment as discussed in the case of sample-15 and sample-55.

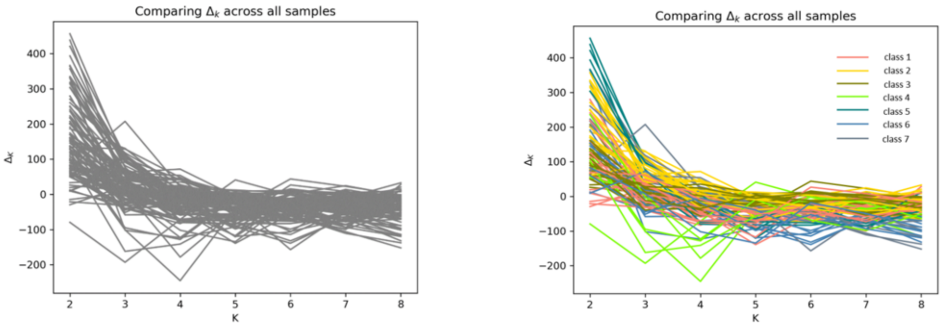

Figure 4. (a) CH curves for all samples; (b) CH curves for all samples after clustering

Following the proposal in section 4, the $\Delta_{H_{cond}}$ can visually be expressed with the help of an annotated dendrogram of sample-10 and sample-31. From fig 3, one can compute the height-factor as ($h_2/h_{root} = 27.1/85.5 = 0.316$) and also $K^*_{\Delta_H} = 2$. Hence, following the criterion in eq 4, the $K^*_{\Delta_{H_{cond}}} = 2$ as the ratio of height is lesser than $1/3$.

Applying the same criterion to sample-31, the height-factor becomes ($h_2/h_{root} = 43.0/68.5 = 0.627$) and with $K^*_{\Delta_H} = 2$ and $K^*_{2\Delta_H} = 3$. Therefore, the second maxima is to considered and thus $K^*_{\Delta_{H_{cond}}} = 3$ is chosen as the best value. Thus, the proposed criterion $\Delta_{H_{cond}}$ correctly matches with experts' number under different morphological structure of the dendrogram.

The error distribution of the $\Delta_{H_{cond}}$ criterion along with other candidate criteria, is provided in Table 1. This method correctly matches 93% of all experts' numbers and performs significantly greater than all other criteria including the proposals in [4].

Table 1. Error distribution of the different CVI analyzed

	Values of error									
CVI	$\varepsilon = 0$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$	$\varepsilon = 6$	$\varepsilon = 7$	**Error rate**	**Acc.**
M_K	15	23	9	5	15	13	19	1	0.85	0.15
Δ_K	23	38	16	15	5	2	1	0	0.77	0.23
$\Delta_{K_{cond}}$	55	25	10	6	2	2	0	0	0.45	0.55
Δ_H	15	44	17	15	5	3	1	0	0.85	0.15
$\Delta_{H_{cond}}$	93	2	2	3	0	0	0	0	0.07	0.93

7. Conclusion

The research reveals that though in theory, the *Calinski-Harabasz* index, being the ratio of between-cluster and within-cluster variance, works similar to how a dendrogram is built when choosing *Ward's* criteria, it does not fully unveil the properties of the data. As a result, the number of clusters determined using dendrogram varies to a large extent when the same is being done through the M_K method. A deeper analysis of this has been discussed in section §6, which leads to developing a novel approach of using dendrogram-height based index ($\Delta_{H_{cond}}$) that emerges to be far superior to any of the criteria listed here. A key advantage of using $\Delta_{H_{cond}}$ lies in the direct application of scaling the hierarchical clustering into real-life applications where humans are replaced with intelligent automated systems while still retaining their inherent heuristic in a mathematical form (as defined in eq 4).

This research directly fits into a bigger project that aims at developing an intelligent decision support system for Industry 4.0 processes in which a specific module deals with the automatic clustering part. Considering the immediate application of this approach, certain preprocessing steps have been ignored in this research such as missing-value-treatment and outlier-removal, however, they may be included in the future lines of this research. Also, the research methodology has been applied on real-life datasets with unknown clusters and rather experts' judgment and is taken as ground truth. This would be extended further in the future by applying the proposal on synthetic as well as a pre-labeled dataset to assess the performance of this criterion. It is also to be noted that the main intent of this research is to identify the number of clusters closest to a human, however, a deeper investigation is also to be done in the future with respect to assessing the quality of the clusters. A key advantage of this research lies in reducing the CPU-time that is consumed in finding the *elbow* of the CVI curve. The proposed criterion identifies the right number of clusters from the initial linkage matrix and no multiple runs of hierarchical clustering are needed.

References

[1] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[2] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[3] Ashutosh Karna and Karina Gibert. Using hierarchical clustering to understand behavior of 3d printer sensors. In *International Workshop on Self-Organizing Maps*, pages 150–159. Springer, 2019.

[4] Ashutosh Karna and Karina Gibert. Automatic identification of the number of clusters in hierarchical clustering. *Neural Computing and Applications*, pages 1–16, 2021.

[5] John E Overall and Kevin N Magee. Replication as a rule for determining the number of clusters in hierarchial cluster analysis. *Applied Psychological Measurement*, 16(2):119–128, 1992.

[6] Beatriz Sevilla-Villanueva, Karina Gibert, and Miquel Sànchez-Marrè. Using cvi for understanding class topology in unsupervised scenarios. In *Conference of the Spanish Association for Artificial Intelligence*, pages 135–149. Springer, 2016.

[7] Yunjae Jung, Haesun Park, Ding-Zhu Du, and Barry L Drake. A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization*, 25(1): 91–111, 2003.

[8] Shibing Zhou, Zhenyuan Xu, and Fei Liu. Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. *IEEE transactions on neural networks and learning systems*, 28(12):3007–3017, 2016.

[9] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[10] Glenn W Milligan. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.

[11] José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, and José C Riquelme Santos. An approach to validity indices for clustering techniques in big data. *Progress in Artificial Intelligence*, 7(2):81–94, 2018.

[12] Beatriz Sevilla-Villanueva, Karina Gibert, and Miquel Sànchez-Marrè. A methodology to discover and understand complex patterns: Interpreted integrative multiview clustering (i2mc). *Pattern Recognition Letters*, 93:85–94, 2017.

[13] Karina Gibert and Claudio Ulises Cortés García. Weighting quantitative and qualitative variables in clustering methods. *Mathware & soft computing. 1997 Vol. 4 Núm. 3*, 1997.

[14] Karina Gibert and Ramon Nonell. Impact of mixed metrics on clustering. In *Iberoamerican Congress on Pattern Recognition*, pages 464–471. Springer, 2003.

[15] Karina Gibert, Aïda Valls, and Montserrat Batet. Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. *Knowledge and information systems*, 40(3):559–593, 2014.

[16] Karina Gibert, Gustavo Rodríguez-Silva, and Ignasi Rodríguez-Roda. Knowledge discovery with clustering based on rules by states: A water treatment application. *Environmental Modelling & Software*, 25(6):712–723, 2010.