The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Towards Automatic Clustering Analysis Using Traces of Information Gain: The InfoGuide Method

Paulo Rocha,[1] **Diego Pinheiro,**[2] **Martin Cadeiras,**[2] **Carmelo Bastos-Filho**[1]
[1] Department of Computer Engineering, University of Pernambuco, Brazil
{phar, carmelofilho}@poli.br
[2] Department of Internal Medicine, University of California, Davis, US
{pinsilva, mcadeiras}@ucdavis.edu

Abstract

Clustering analysis has become a ubiquitous information retrieval tool in a wide range of domains, but a more automatic framework is still lacking. Though internal metrics are the key players towards a successful retrieval of clusters, their effectiveness on real-world datasets remains not fully understood, mainly because of their unrealistic assumptions underlying datasets. We hypothesized that capturing *traces of information gain* between increasingly complex clustering retrievals—*InfoGuide*—enables an automatic clustering analysis with improved clustering retrievals. We validated the *InfoGuide* hypothesis by capturing the traces of information gain using the Kolmogorov-Smirnov statistic and comparing the clusters retrieved by *InfoGuide* against those retrieved by other commonly used internal metrics in artificially-generated, benchmarks, and real-world datasets. Our results suggested that *InfoGuide* can enable a more automatic clustering analysis and may be more suitable for retrieving clusters in real-world datasets displaying nontrivial statistical properties.

Introduction

Clustering analysis has become ubiquitous in the retrieval of clusters from a plethora of datasets arising from a wide range of domains (Adolfsson, Ackerman, and Brownstein 2019), supporting the characterization of the development of personalized medical therapies (Bakir et al. 2018), the understanding of intricate social-economic factors (Mirowsky et al. 2017), and the development of healthcare ranking systems (Wallace et al. 2019). Since the creation of the first clustering algorithm in 1948, by the botanist and evolutionary biologist Thorvald Sørensen while studying biological taxonomy (Sørensen 1948), novel algorithms for clustering retrieval have been proposed (Xu and Tian 2015). However, a framework for automatic clustering analysis is still lacking and even determining the optimal number of clusters to be retrieved remains a major methodological issue (Tibshirani, Walther, and Hastie 2001).

Given that ground truth labels are inherently absent, the clustering retrieval largely relies on internal metrics of clustering quality (Arbelaitz et al. 2013). These metrics are idealized aspects of clustering quality defined a priori and often involve unrealistic assumptions about the datasets (Tibshirani, Walther, and Hastie 2001; Rousseeuw 1987; T. Caliński 1974). Nevertheless, clustering algorithms not only, directly or indirectly, retrieve clusters according to these internal metrics (MacQueen 1967; Jr. 1963), but also have their clustering retrieval subsequently evaluated according to these same metrics. As a result, different internal metrics often disagree with each other regarding the quality of a specific clustering retrieval (Tibshirani, Walther, and Hastie 2001; Xu and Tian 2015). Another crucial issue relies on the fact that most of the metrics use distances and, sometimes, the distances in different attributes of the problem may have different meanings.

We hypothesized that capturing the *traces of information gain* between increasingly complex clustering retrievals—the *InfoGuide* method—can enable a more automatic clustering analysis. We validated the *InfoGuide* hypothesis by capturing the traces of information using the well-known Kolmogorov–Smirnov statistic and comparing the clusters retrieved by *InfoGuide* against those retrieved by other commonly used internal metrics over artificially-generated, benchmarks, and real-world datasets. Our results suggest that *InfoGuide* may be more suitable to retrieve clusters in real-world datasets displaying nontrivial statistical properties.

Related Work

The application of a clustering algorithm g over a dataset $\mathcal{X}$, with N data points and a set of F features, to retrieve k groups, a clustering retrieval $C^{(k)}$, can be generally defined as a mapping $g_k : \mathcal{X} \rightarrow C^{(k)}$ such that each data point $x \in \mathcal{X}$ is assigned to one of the k clusters $c_i^k \in C^{(k)}$. Each $c_i^{(k)}$ in $C^{(k)}$ represents a subgroup of the dataset $\mathcal{X}$ as following:

$$C^{(k)} = \{c_1^{(k)}, c_2^{(k)}, \ldots, c_k^{(k)}\} \ , \tag{1}$$

in which N_i is the number of data points in $c_i^{(k)}$.

Clustering algorithms can be classified into different categories according to their assumptions about clustering retrieval, namely, partitions, hierarchy, density, distribution,

subspace, to name but a few (Rodriguez et al. 2019; Xu and Tian 2015). Despite the differences among categories, the main idea underlying clustering retrieval is that data points belonging to the same cluster should be similar to each other and dissimilar from data points belonging to other clusters (Sørensen 1948).

In general, the similarity between two data points depends on their distance (Sørensen 1948). The chosen distance (e.g., Euclidean, Manhattan, Mahalanobis) applied to the g_k can generate a bias in the shape of the groups, resulting in different clustering retrievals even with the same $\mathcal{X}$ and k. To proper evaluate the quality of the mapping $g_k : \mathcal{X} \rightarrow C^{(k)}$, different internal metrics $m_k : C^{(k)} \rightarrow q$ have been proposed, in which q represents a comparable scalar enabling the comparison between different numbers of k as well as the possibility of finding the optimal number of clusters $\hat{k}$ for $g_k \in [k_{min}, k_{max}]$.

In the extensive study of Arbelaitz et al, 30 internal metrics were evaluated in a wide range of datasets, demonstrating that most of the metrics simply determine the quality of clustering retrievals by applying primarily two criteria: the distance between points in the same cluster, described as *cohesion*, and the distance between different groups, described as *separation*. These metrics have a bias that better evaluates a set of groups with both high cohesion and separation, and can thus be defined as distance-based internal metrics (Arbelaitz et al. 2013).

Conversely, a information theoretic measure of cluster separability was developed by Gokcay and Principe as a cost function to guide an optimization clustering algorithm (Gokcay and Principe 2002). The authors used both artificially generated and image segmentation datasets. Similarly, Faivishevsky and Goldberger proposed the clusters mutual information as the maximization objective to be used by a clustering algorithm (Faivishevsky and Goldberger 2010). The authors demonstrated that an entropy-based approach may be more suitable than a distance-based approach for clustering analysis over both artificially-generated and benchmark datasets.

In this paper, a information theoretic approach for clustering analysis was moved forward given the main challenges faced by distance-based metrics especially in real-world datasets with nontrivial distributions, in which concepts such as averages and distance-based similarities become unrealistic (Gokcay and Principe 2002). In this sense, we proposed the *InfoGuide* method for automatic clustering analysis in which an optimal clustering retrieval is based on the information gained between increasingly complex clustering retrievals.

Methods

Clustering analysis involves the following elements: a dataset, a set of clustering algorithms, as well as internal and external metrics of clustering retrieval (Figure 1, A). In this work, we proposed the *InfoGuide* method for automatic clustering analysis using traces of information gain (Figure 1, B).

InfoGuide—An Automatic Retrieval of Clusters using Traces of Information Gain

The challenge in clustering analysis is retrieving the highest number $\hat{k}$ of *meaningful* clusters as close to the optimal number k^* of clusters as possible, avoiding both underfitting $\hat{k} < k^*$ and overfitting $\hat{k} > k^*$. The definition of a *meaningful* cluster not only depends on the specific internal metric used but also is affected by the specific clustering algorithm employed.

Let $C^{(k)}$ and $C^{(k+1)}$ be the set of k and $k+1$ increasingly complex clustering retrievals, respectively, the *InfoGuide* method retrieves the smallest number of clusters $\hat{k}$ as long as an increased information gain can be obtained between increasingly complex clustering retrievals as following:

$$\hat{k} = \text{smallest k such that } C^{(k+1)} \stackrel{d}{=} C^{(k)} \ , \tag{2}$$

in which the clustering retrieval $C^{(k+1)}$ is equivalent to $C^{(k)}$ according to the pairwise equivalencies between their individual clusters as following:

$$\begin{aligned} & C^{(k+1)} \stackrel{d}{=} C^{(k)} \iff \\ & \forall c_i^{(k+1)} \ \exists \ c_j^{(k)} (c_i^{(k+1)}, c_j^{(k)} \in C^{(k+1)} \times C^{(k)}) \\ & c_i^{(k+1)} \stackrel{d}{=} c_j^{(k)} \ , \end{aligned} \tag{3}$$

in which individual clusters $c_i^{(k+1)}$ and $c_j^{(k)}$ are as following:

$$c_i^{(k+1)} \stackrel{d}{=} c_j^{(k)} \iff (\forall f \in F) \ f_i \stackrel{d}{=} f_j \ , \tag{4}$$

in which the feature f in $c_i^{(k+1)}$ and $c_i^{(k)}$ are equivalent in distribution. Therefore, the *InfoGuide* method only considers that the clustering retrieval $C^{(k+1)}$ increases the information gain relative to $C^{(k)}$ when it retrieves novel clusters not already contained in $C_{(k)}$. Otherwise, retrieving a higher number of clusters only results in a more complex model without information gain.

In this work, the Kolmogorov-Smirnov KS statistic was used to quantify the equivalency in distribution between features such that $f_i \stackrel{d}{=} f_j \equiv KS(f_i, f_j)$. Information gain is thus the statistical evidence that both features may not come from the same statistical distribution whenever the p-value of the KS test is lower than the statistical significance α after using the Bonferroni correction for the $F \times (k+1) \times k$ multiple comparisons. The optimal $\hat{k} \in [k_{min}, k_{max}]$ is the highest $\hat{k}$ that can be obtained for a range of $\alpha_u \in (0, \alpha]$.

Metrics

The *InfoGuide* method was compared with three commonly used internal metrics that embrace the two main ideas underlying clustering analysis, namely, cohesion and separation. Let $\mathcal{X}^N$ be a dataset with N data points and $C^{(k)}$ a clustering retrieval, these internal metrics use the following basic calculations of distance: between two data points, $(x_i - x_j)$, between a data point and the estimated value of a group,

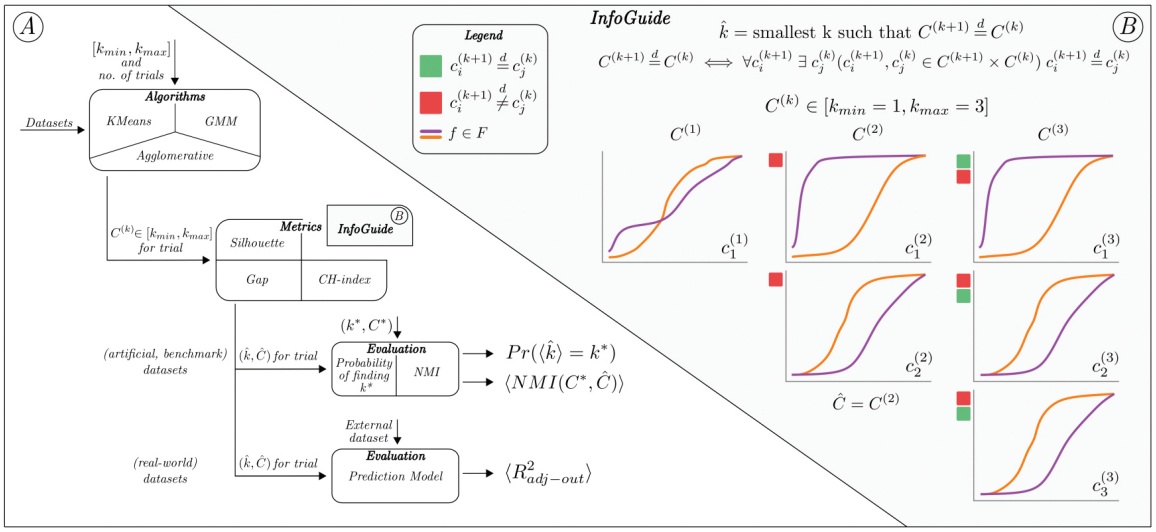

Figure 1: (A) Clustering Analysis. For each dataset, a clustering retrieval is obtained by each clustering algorithm. Retrievals are subsequently evaluated by internal metrics, and an optimal clustering retrieval $\hat{C}$) is determined by each metric. When ground-truth is available, the probability of finding the true number of clusters k^* and the normalized mutual information $NMI(C^*, \hat{C})$ are evaluated. In the absence of ground-truth, the goodness of fit of a prediction model (e.g., the out-sample adjusted R^2, $R^2_{adj-out}$) is evaluated when the clustering retrieval $\hat{C}$ is included as an additional predictor. (B) Illustration of the *InfoGuide* method for $k_{min} = 1$, $k_{max} = 3$, and $k^* = 2$. The *InfoGuide* evaluates the equivalency between increasingly complex clustering retrievals $C^{(k)}$ and $C^{(k+1)}$. There is no information gain, for instance, when comparing $C^{(2)}$ and $C^{(3)}$ because for each cluster in $C^{(3)}$ there is at least one equivalent cluster in $C^{(2)}$. As a result, the optimal number of clusters is $\hat{k} = 2$.

$(x_i - \langle c_i^{(k)} \rangle)$, and between the estimated values of a group and a dataset, $(\langle c_i^{(k)} \rangle - \langle C^{(k)} \rangle)$.

The *Silhouette* chooses the optimal $\hat{k}$ by maximizing the average difference between the separation and cohesion as following (Rousseeuw 1987):

$$SI = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max(b_i, a_i)} \;, \tag{5}$$

in which a_i measures the cohesion of a data point i as following:

$$a_i = \frac{1}{N_i - 1} \sum_{j=1, j \neq i}^{N_i} (x_i - x_j) \;, \tag{6}$$

and b_i measures the separation of a data point i to the other points belonging to nearest cluster as following:

$$b_i = \min_{1 \leq l \leq k, x_i \notin c_i^{(k)}} \left(\frac{1}{N_l} \sum_{j=1}^{N_l} (x_i - x_j) \right) \;. \tag{7}$$

The *Calinsk-Harabasz (CH) Index* chooses the optimal $\hat{k}$ by maximizing the ratio between the Sum of Squares Within (SSW) and the Sum of Squares Between (SSB) as following (T. Caliński 1974):

$$CH = \frac{N - k}{k - 1} \cdot \frac{SSB}{SSW} \;, \tag{8}$$

in which $SSW = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (x_j - \langle c_i^{(k)} \rangle)^2$ is a measure of cohesion and $SSB = \sum_{i=1}^{k} N_i \cdot (\langle c_i^{(k)} \rangle - \langle C^{(k)} \rangle)^2$ is a measure of separation. It is normalized by the number of data points N and the number of groups k to ensure a similar scale when comparing different numbers of groups.

The *Gap Statistic* maximizes the SSW of a clustering retrieval from the actual dataset relative to what would be expected by a clustering retrieval from an uniformly distributed dataset SSW_{random} as following (Tibshirani, Walther, and Hastie 2001):

$$Gap = \mathbb{E}(\log(SSW_{random})) - \log(SSW) \;, \tag{9}$$

such that greater the difference between random and actual cohesions, the higher the quality of the clustering retrieval. The optimal $\hat{k}$ is chosen as the smallest k where $Gap(k) \geq Gap(k+1) - S_{k+1}$, and S_{k+1} is the standard deviation of $\log SSW_{random}$.

Experimental Setup

The *InfoGuide* method was validated by comparing the quality of its clustering retrievals with those of other internal metrics over artificially-generated, benchmarks, and real-world datasets. Three commonly used clustering algorithms with distinct underlying approaches were used:

K-Means (MacQueen 1967), Gaussian Mixture Model (GMM) (Rasmussen 2000) and the Agglomerative Ward which is a Hierarchical Agglomerative with Ward's linkage (Jr. 1963). For each algorithm, $k \in [k_{min}, k_{max}]$ clusters were repeatedly retrieved 30 times with $k_{min} = 1$ and $k_{max} = 11$. The optimal clustering retrieval $\hat{C}$ was obtained according to the *InfoGuide* as well as to the Silhouette, Calinsk-Harabasz Index, and Gap Statistic. A total of $7,920$ clustering retrievals $C^{(k)}$ were considered using 8 datasets $\times$ 30 trials $\times$ 3 algorithms $\times$ $|[k_{min}, k_{max}]| = 11$ number of clusters.

For artificially-generated and benchmark datasets, for which the ground truth C^* are available, two evaluations were performed: the probability of finding the true k^*, $Pr(\langle\hat{k}\rangle = k^*)$, and the Normalized Mutual Information ($NMI(C^*, \hat{C})$) between the clusters retrieved $\hat{C}$ and ground-truth C^*. The probability of finding k^* was quantified using the Wilson Score, which estimates the population proportion of a binomial distribution in which a success is encoded as $\hat{k} = k^*$. The NMI quantifies the decrease in the entropy of $\hat{C}$ by knowing C^*.

For real-world datasets, an external evaluation was performed by quantifying the goodness of fit of a prediction model when the optimal clustering retrieval $\hat{C}$ is included as an additional predictor. In this work, a Linear Regression was used, and the goal was to compare different metrics instead of obtaining the best prediction model. To control for model complexity and avoids overfitting, the adjusted R^2 out-sample, $R^2_{adj-out}$ was used. All of the code, datasets, and analysis are available on the Open Science Framework (OSF) repository of this project at https://doi.org/10.17605/OSF.IO/ZQYNC.

Data

Artificially-generated, benchmark, and real-world datasets were used (Table 1). The artificial datasets were reproduced from the previous work of Tibshirani et al on Gap Statistic (Tibshirani, Walther, and Hastie 2001), in which 5 datasets were artificially generated according to normally distributed features. For this work, the first dataset was excluded to ensure a fair comparison among the other internal metrics. This dataset arbitrarily assumes that only one group exists and internal metrics such as Silhouette and the CH index are not intrinsically able to retrieve $\hat{k}$ as one. Benchmark datasets have been extracted from the UCI repository (Dua and Graff 2017), which contains, unlike artificially-generated data, datasets with non-normal statistical distributions, often displaying, for instance, a high skewness. In this work, a real-world dataset of containing socioeconomic variables at the county-level was obtained from the American Community Survey (ACS 2018). It includes race, education, and income for each county in the United States. A goodness of fit measurement of a prediction model was used in which the number of heart-failure deaths is predicted based on the following associated predictors: the total population size, the number of population with diabetes and obesity as well as the percentage of the population with age greater than 65 years. The dataset was obtained from the Centers for Disease Control and Prevention (CDC 2018).

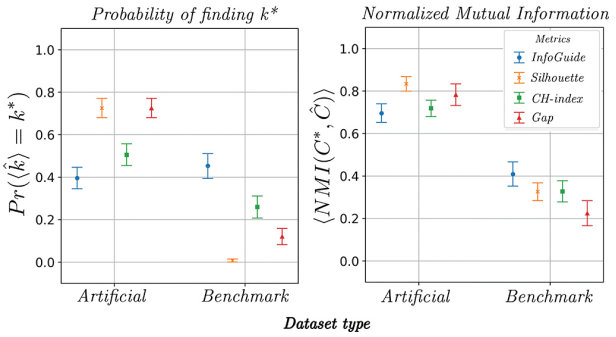

Figure 2: Comparison of clusters retrieved from artificially-generated and benchmark datasets according to (left) the probability of finding the true number of clusters k^* and (right) the normalized mutual information between the retrieved $\hat{C}$ and true C^* clusters.

type	dataset	N	F	k^*
artificial	b	1000	10	3
	c	1000	10	4
	d	1000	10	4
	e	1000	10	2
benchmark	Iris	150	4	3
	Wine	178	13	3
	Wine quality	1599	11	6
real-world	ACS county	3142	21	-

Table 1: The characteristics of the datasets.

Results

Comparison of Clustering Retrieval among Dataset Types

Quality measures of clustering retrieval quantifies to extent to which the retrieved clusters resemble idealized clustering aspects that are often unrealistic when considered the statistical properties underlying the generating process of real-world datasets. Not surprisingly, these measures are largely evaluated over artificially-generated datasets. The clustering retrieval of *InfoGuide* was compared against other approaches using both artificially-generated (Figure 2, left) and benchmark data sets (Figure 2, right).

Overall, the correct number of clusters is more likely retrieved and a higher information gain is typical obtained in the artificially-generated datasets than in the benchmark datasets. *InfoGuide* not only displays the highest information gain in the benchmark datasets but also it displays the smallest decrease in information gain from artificial to benchmark datasets. Though the Silhouette and Gap appears

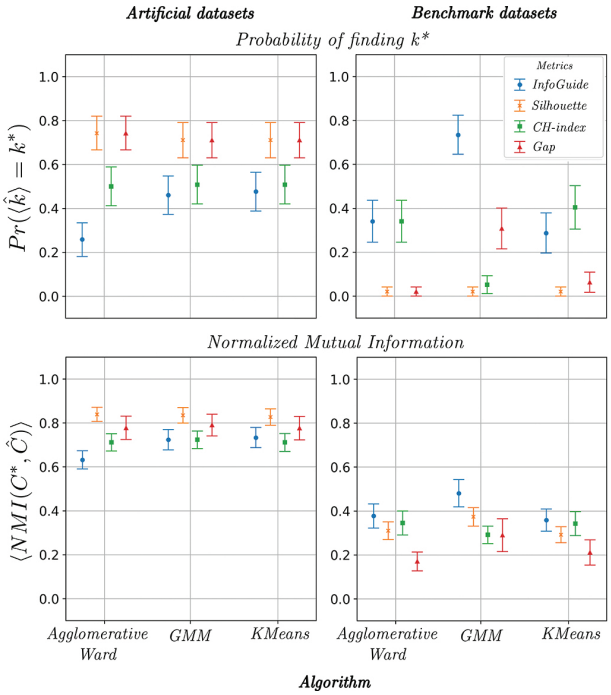

Figure 3: Comparison of clusters retrieved from artificially-generated (left) and (right) benchmark datasets according to (top) the probability of finding the true number of clusters k^* and (bottom) the normalized mutual information between the retrieved $\hat{\mathcal{C}}$ and true $\mathcal{C}^*$ clusters.

to retrieve superior clusters in the artificial datasets, they retrieve the worst clusters in the benchmark datasets.

Comparison of Clustering Retrieval among Algorithms

Generally, each clustering algorithm attempts to retrieve clusters that resemble its idealized aspects of clustering quality defined a priori. Therefore, the clustering retrieval of each algorithm was separately compared according $Pr(\langle k \rangle = k^*)$ and $NMI(C^*, \hat{C})$ using both artificially-generated (Figure 3, left) and benchmark (Figure 3, right) datasets.

When the Agglomerative-Ward is used, both Gap and Silhouette retrieved the best clusters in the artificial datasets but the worst clusters in the benchmark datasets. Even *InfoGuide* has retrieved the worst clusters when the Agglomerative-Ward is used. Interestingly, Agglomerative-Ward is the only deterministic algorithm and its results may suggest that stochastic components may aid algorithms navigating complex datasets.

When the algorithms GMM and KMeans were used, each metric retrieved comparable clusters from the artificial datasets according to either $Pr(\langle k \rangle = k^*)$ and NMI. Using the benchmark datasets, however, *InfoGuide* retrieved the best clusters when the algorithm GMM was used such that when compared to the second-best metric, Gap, *InfoGuide* were two times more likely to retrieve the correct number of clusters and also obtained almost two times more information gain.

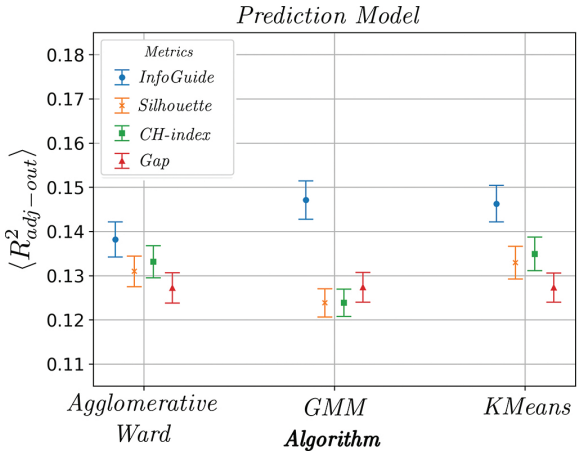

Figure 4: Results of the Linear regression model for $R^2_{adjusted}$ (out-of-sample) metric. The prediction was the rate of heart failure by county at USA, the clusters found by each algorithm, guided by each metric, using information about race, income and education (also by county), were used to help the prediction. The clusters found by *InfoGuide* added more information to the model in comparison to the other metrics for the GMM and K-Means algorithms.

Comparison of Clustering Retrievals in Real-World Datasets

Clustering analysis has been used to find groups in real-world datasets lacking ground truth. To circumvent the absence of ground truth, external validation is commonly used by independently choosing an external dataset of interest that contains metadata associated to all data points within each cluster.

Overall, the clusters retrieved by $InfoGuide$ obtained the highest adjusted coefficient of determination out-sample $R^2_{adj-out}$ when compared to the other metrics (Figure 4). The clusters retrieved by *InfoGuide* were able to explain roughly 3% more variation of heart failure deaths. Though it is a modest improvement, it can correspond to a total of 100 thousand heart failure deaths incorrectly predicted among the 2.3 million total heart failure deaths in the US.

Conclusions

After half-century since the inception of the first clustering algorithm, however, clustering analysis still lacks a more automatic framework for clustering retrieval that is based on internal metrics with less unrealistic assumptions (e.g., normal distributions). In this work, we proposed the *InfoGuide* method that uses traces of information gain for automatic clustering analysis.

The results demonstrated that *InfoGuide* may be more suitable for retrieving clusters in real-world datasets displaying nontrivial statistical properties. In benchmark and real-world datasets, GMM, which is the algorithm with less strict

assumptions, was capable of obtaining the best clustering retrieval. Future works should include a more diverse set of clustering algorithms and datasets from other domains, and a comparative performance analysis of the method and other internal metrics. Despite additional validations, the *InfoGuide* method and the idea of using traces of information gain may become a suitable method for automatic clustering analysis.

References

ACS. 2018. American community survey. https://factfinder.census.gov/. Accessed: 2019-11-18.

Adolfsson, A.; Ackerman, M.; and Brownstein, N. C. 2019. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition* 88:13–26.

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

Bakir, M.; Jackson, N. J.; Han, S. X.; Bui, A.; Chang, E.; Liem, D. A.; Ardehali, A.; Ardehali, R.; Baas, A. S.; Press, M. C.; Cruz, D.; Deng, M. C.; DePasquale, E. C.; Fonarow, G. C.; Khuu, T.; Kwon, M. H.; Kubak, B. M.; Nsair, A.; Phung, J. L.; Reed, E. F.; Schaenman, J. M.; Shemin, R. J.; Zhang, Q. J.; Tseng, C.-H.; and Cadeiras, M. 2018. Clinical phenomapping and outcomes after heart transplantation. *The Journal of Heart and Lung Transplantation* 37(8):956–966.

CDC. 2018. Centers for disease control and prevention. https://nccd.cdc.gov. Accessed: 2019-11-18.

Dua, D., and Graff, C. 2017. UCI machine learning repository.

Faivishevsky, L., and Goldberger, J. 2010. Nonparametric information theoretic clustering algorithm. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 351–358.

Gokcay, E., and Principe, J. C. 2002. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(2):158–171.

Jr., J. H. W. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58(301):236–244.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–297. Berkeley, Calif.: University of California Press.

Mirowsky, J. E.; Devlin, R. B.; Diaz-Sanchez, D.; Cascio, W.; Grabich, S. C.; Haynes, C.; Blach, C.; Hauser, E. R.; Shah, S.; Kraus, W.; Olden, K.; and Neas, L. 2017. A novel approach for measuring residential socioeconomic factors associated with cardiovascular and metabolic health. *Journal of Exposure Science and Environmental Epidemiology* 27(3):281–289.

Rasmussen, C. E. 2000. The infinite gaussian mixture model. In *Advances in neural information processing systems*, 554–560.

Rodriguez, M. Z.; Comin, C. H.; Casanova, D.; Bruno, O. M.; Amancio, D. R.; Costa, L. d. F.; and Rodrigues, F. A. 2019. Clustering algorithms: A comparative approach. *PloS one* 14(1):e0210236.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53 – 65.

Sørensen, T. J. 1948. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. I kommission hos E. Munksgaard.

T. Caliński, J. H. 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3:1–27.

Tibshirani, R.; Walther, G.; and Hastie, T. 2001. 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.

Wallace, M.; Sharfstein, J. M.; Kaminsky, J.; and Lessler, J. 2019. Comparison of US County-Level Public Health Performance Rankings With County Cluster and National Rankings. *JAMA Network Open* 2(1):e186816–11.

Xu, D., and Tian, Y. 2015. A comprehensive survey of clustering algorithms. *Annals of Data Science* 2(2):165–193.