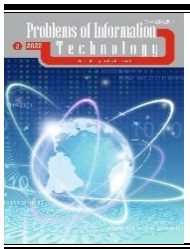

Available online at www.jpit.az

13 (2)
2022

COMPARATIVE ANALYSIS OF CLUSTER VALIDITY INDICES IN TERMS OF CONSISTENCY

Leyla R. Mammadova

Institute of Information Technology, Azerbaijan National Academy of Sciences, B. Vahabzade str., 9A, AZ1141 Baku, Azerbaijan

mammadova-1998@list.ru

ARTICLE INFO

http://doi.org/10.25045/jpit.v13.i2.03

Article history:
Received 18 January 2022
Received in revised form 28 March 2022
Accepted 30 May 2022

Keywords:
Cluster analysis
Clustering algorithms
Evaluation indices

ABSTRACT

Cluster analysis is one of the key issues in Data Mining, the most important stage of the Knowledge Discovery from Data (KDD) process, and is widely used. There are 3 main tasks of cluster analysis: determining the optimal number of clusters, clustering algorithms and evaluating the quality of clustering. One of the most important steps in cluster analysis is to evaluate the quality of clustering. A number of indices have been proposed to assess the outcome of clustering. The analysis shows that these indices, which are used to assess the quality of clustering, often show inconsistent results. Therefore, extensive research has recently been conducted on the study of indices and new indices are proposed. The article examines a number of internal and external evaluation indices. Different size data sets are taken and k-means, k-medoids, agglomerative hierarchical, BIRCH and OPTICS algorithms are applied to them. A number of internal and external evaluation indices are used to assess the results of the experiment, and the results are analyzed comparatively. Experiments show that Ac, Pr, Rc and F-m indices show similar results in in group determining in a given clustering structure.

1. Introduction

Clustering is an unsupervised method of machine learning. The main goal is to group similar objects in the same clusters and different objects in different clusters according to a certain similarity measure (Euclidean, etc.). (Alguliev, Aliguliyev, 2005). The application of cluster analysis is very wide: marketing, biology, medicine, oil and gas industry, smart city, information security, urban planning, earthquake research, etc.

The cluster analysis process is divided into the following steps (Aggarwal and Reddy, 2013):

- Problem formation;
- Selection of similarity measure (Euclidean distance, etc.);
- Selecting the appropriate algorithm for data clustering;
- Defining the number of clusters;
- Evaluation of clustering results.

One of the most important steps in cluster analysis is to assess the quality of clustering. Some indices have been proposed to assess the results of clustering. The analysis shows that these indices, which are used to assess the quality of clustering, often show conflicting results. Therefore, extensive research has recently been conducted on the study of indices, and new indices are proposed. For this purpose, some algorithms used during clustering, issues such as clustering quality assessment indices are considered in the article. Evaluation indices of clustering results are studied on different size data sets and analyzed comparatively.

2. Clustering algorithms and their classification

There are many algorithms for clustering. There are different clustering algorithms, but there is no perfect clustering algorithm (Mayra, et al, 2016). Each algorithm offers different approaches to the problem of group detection in data.

Clustering methods are divided into five classes: 1) partitioning methods; 2) hierarchical methods; 3) model-based methods; 4) density-based methods, and 5) grid-based methods.

Partitioning clustering is one of the most widely used clustering methods by analysts. Initially, all points are stored in the same cluster. Smaller clusters are created by dividing the cluster using sequential iteration (Aliguliyev, 2007). The iteration continues until the convergency criteria is completed (number of iterations, time, etc.). k-means, k-medoids are widely used partitioning clustering methods based on partitioning (Sajana, et al, 2016).

Hierarchical clustering is an unsupervised clustering method used to group data (Chana, & Arora, 2014; Chandra, & Anuradha, 2011). Hierarchical clustering is an algorithm that builds a hierarchy of clusters. The resulting clusters are displayed in a hierarchical tree called a dendrogram. This is a clear graphical representation of clusters that can be easily understood and interpreted (Mammadova, 2021).

There are two main types of hierarchical clustering methods (Han, Kamber, Pei, 2012):

The working principle of agglomerative algorithms is as follows (Turi, 2001):

1. It starts with a cluster equal to the number of points in the data set, treats each data as a cluster.
2. A distance matrix is constructed based on the similarity measure (Euclidean, Manhattan, etc.).
3. The two clusters located at the minimum distance are merged, and the distance matrix is updated.
4. Repeat step 3 until all data points are in a single cluster.

The divisional algorithm works in contrast to the agglomerative algorithm. The working principle of the divisional algorithm is as follows (Rafsanjani, et al, 2012):

1. The process begins with a cluster that includes all objects.
2. The cluster is divided into 2 clusters by the existing clustering algorithm (k-means, etc.).
3. SSE (sum of the squared errors) is calculated for each cluster, and the one with the highest value is selected, and the cluster is divided using the clustering algorithm.

$$SSE = \sum_{j=1}^{k} d(x, m_j)^2$$

where C_j is the cluster j, m_j is the center of the C_j, $d(x, m_j)$ is the Euclidean distance between point x, and the center m_j.

4. Repeat step 3 until the number of clusters is obtained.

The article contains a large number of data of different sizes, and many clustering algorithms have been applied to them. The result of clustering was assessed by various indices. The k-means, k-medoids, BIRCH, and OPTICS algorithms applied to the data set are explained in detail below:

2.1. k-means algorithm

k-means is a partitioning clustering algorithm. The working principle of the k-means algorithm is explained below (Ahmadov, 2021):

1. The number of clusters (k) is given.
2. Some cluster centers are randomly selected from a set of data.
3. The distance between the centers of the cluster and the points is calculated and the point is assigned to the nearest group.
4. The cluster center is recalculated for the created groups.
5. Repeat steps 3 and 4 until the convergence criterion is completed.

2.2. k-medoids algorithm

k-medoids is a partitioning clustering algorithm (Zhang, Couloigner, 2005). The calculation is more complicated than the k-means algorithm. Thus, while the time complexity of the k-means algorithm is $O(n \times m \times k \times t)$, the time complexity of the k-medoids algorithm is equal to $O(n^2 \times m \times k \times t)$. Here, n is the number of points, m is the number of attributes, k is the number of clusters, and t is the number of iterations.

The working principle of the k-medoids algorithm is explained in several steps (Hamerly, 2003):

1. The number of clusters (k) is given.

2. From the set of data, we select k points out of n points as the center of the cluster.
3. Each point is then connected to the nearest cluster center.
4. The cluster centers are reselected and step 3 is repeated.
5. Repeat steps 3 and 4 until the convergence criterion is completed.

2.3. BIRCH algorithm

BIRCH (BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies) algorithm (Zhang, Ramakrishnan, & Livny, 1997).

It is an algorithm used to perform hierarchical clustering, especially on big data. This algorithm uses a cluster feature tree for cluster hierarchy. A cluster feature is a three-dimensional vector: < S, XT,>

where

S – number of objects;

XT: $\sum_{i=1}^{S} T_i$;

KT: $\sum_{i=1}^{S} T_i^2$.

2.4. OPTICS algorithm

OPTICS **(**OPTICS - Ordering Points to Identify the Clustering Structure) **(**Ankerst, et al, 1999**)**. This algorithm searches for clusters by identifying the parts of the data set where the points are dense and points in low-density areas are marked as noise points (Alguliyev, Aliguliyev, Abdullayeva, 2019**)**.

The following are some parameters for explaining the working principle of the OPTICS algorithm:

ε (epsilon) – specified a distance around any point p.

minPts (minimum number of points) - this parameter is used with ε. It indicates the number of points that will be located at a distance ε of point p to create a cluster.

Core point - there are enough points around ε, including point p itself, it is considered a core point.

Core distance - is the smallest distance between point p and an object in its ε-neighborhood that p be the core point.

Reachability distance - any point is directly density-reachable from the core point. It continues to build clusters.

The algorithm starts with an arbitrary point, and its neighborhood information is taken from the parameter ϵ. Cluster formation begins, if there are minPts around ϵ. Otherwise, the point is marked as noise. This point can then be found in the neighborhood of another point ϵ and thus become part of the cluster. If a point is found to be the core point, then the points in the neighborhood ϵ are also part of the cluster. Thus, all points found around ϵ, if they are also core points, are added together with their ϵ neighborhood. This process continues until the density-related cluster is completely found. The process starts again with a new point, which can be part of a new cluster or marked as noise.

3. Indices for evaluation of the clustering quality

Various indices have been proposed to assess the quality of clustering. These indices are divided into two groups: internal and external evaluation indices (Aliguliyev, 2009; Arbelaitz, et al, 2013). If the classes of the data set are not known in advance, then internal indices are used to assess the quality of clustering. If the classes of the data set are known in advance, then external indices are used to assess the quality of clustering. These indices compare the calculated clusters with the reference classes. Various indices have been proposed by researchers to measure this closeness. This article will provide information on some of the most commonly used indices in cluster analysis.

First, let's look at external evaluation indices to assess the quality of clustering:

Accuracy:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

where,

TP – True Positives (points belonging to the same class and the same cluster)

FN – False Negatives (points belonging to the same class and different clusters)

FP – False Positives (points belonging to different classes and the same cluster)

TN – True Negatives (points belonging to different classes and different clusters)

F-measure: This index is calculated as a harmonic mean of precision and recall (Mammadova L., 2021).

Precision is calculated as follows:

$$Pr = \frac{TP}{TP + FP}$$

Recall is so calculated as:

$$Rc = \frac{TP}{TP + FN}$$

Then, the F-measure will be calculated as follows:

$$F\text{-}m = \frac{2}{\frac{1}{Pr}+\frac{1}{Rc}} = \frac{2*Pr*Rc}{Pr+Rc}$$

The best possible value of the F-measure is one, it shows perfect precision and recall, the lowest possible value of the F-measure is zero.

Rand index:

The Rand Index is an index used to measure the similarity between clusters. This index is calculated as follows (Rand, 1971):

$$\text{RI} = \frac{TP + TN}{N}$$

Adjusted Rand Index:

The main problem with the Rand Index is that the expected value of the Rand Index of two random partitions does not take a constant value. To solve the problem, the Adjusted Rand Index is introduced, where the generalized hyper-geometric distribution is adopted as a random model (Kvalseth, 1987).

$$\text{ARI} = \frac{(RI - expected(RI))}{(\max(RI) - expected(RI))}$$

Suppose that data set $\boldsymbol{D} = \{d_1, d_2, \dots, d_N\}$ divided into classes $\boldsymbol{A} = (A_1, A_2, \dots, A_q)$. After clustering the data set, let us mark the set of clusters obtained by $\boldsymbol{C} = (C_1, C_2, \dots, C_k)$. To calculate similarity with the clusters $\boldsymbol{C} = (C_1, C_2, \dots, C_k)$ of classes $\boldsymbol{A} = (A_1, A_2, \dots, A_q)$, various indices are determined as follows (Aliguliyev, 2009).

V-measure: This index is calculated as the harmonic mean of homogeneity and completeness.

Homogeneity: The cluster has points belonging to the same class. Homogeneity is calculated by the following formula (Rosenberg & Hirschberg, 2007):

$$Hom = 1 - \frac{B(q,k)}{B(q)}$$

where that

$$B(q,k) = -\sum_{p=1}^{k}\sum_{r=1}^{q}\frac{NM_{pr}}{N}\log\left(\frac{NM_{pr}}{\sum_{r=1}^{q}NM_{pr}}\right)$$

and

$$B(q) = -\sum_{r=1}^{q}\frac{\sum_{p=1}^{k}NM_{pr}}{q}\log\left(\frac{\sum_{p=1}^{k}NM_{pr}}{q}\right)$$

Here q is the number of different classes,
k is the number of clusters, NM_{pr} is the number of common points of class A_r with the cluster C_p, in other words, the number of points at the intersection of these two sets, $NM_{pr} = |C_p \cap A_r|$, N is the number of all points.

Completeness: is the collection of all points belonging to the same class in the same cluster. Completeness is calculated by the following formula (Rosenberg & Hirschberg, 2007):

$$Cm = 1 - \frac{B(k,q)}{B(k)}$$

where that,

$$B(k,q) = -\sum_{r=1}^{q}\sum_{p=1}^{k}\frac{NM_{pr}}{q}\log\left(\frac{NM_{pr}}{\sum_{p=1}^{k}NM_{pr}}\right)$$

and

$$B(k) = -\sum_{p=1}^{k}\frac{\sum_{r=1}^{q}NM_{pr}}{q}\log\left(\frac{\sum_{r=1}^{q}NM_{pr}}{q}\right)$$

In this case, the V-measure is calculated as follows:

$$V = \frac{2}{\frac{1}{Hom}+\frac{1}{Cm}} = \frac{2 \cdot Hom \cdot Cm}{Hom + Cm}$$

Purity: Initially, the Purity Index is calculated for the C_p cluster (Murtagh, 1984):

$$\text{Purity}(\boldsymbol{C}_p) = \frac{1}{N_p}\max_{r=1,\dots,q} NM_{pr}, \quad p = 1,2,\dots,k,$$

N_p is the number of points in the cluster C_p.

The following formula calculates the Purity Index for the whole set of clusters:

$$\text{Purity}(\boldsymbol{C}) = \sum_{p=1}^{k}\frac{N_p}{N}\text{Purity}(C_p),$$

The purity index is rated in the interval $\left[\frac{1}{q}, 1\right]$. The large value of this index is the best solution for clustering.

Mutual information. This index calculates the similarity of points between clusters. The mutual information index for all clusters is calculated as follows:

$$MI(A,C) = H(A) - H(A|C) = H(C) - H(C|A)$$

$$MI(A,C) = H(A) + H(C) - H(A,C)$$

where $\text{Mİ}(A, C)$ is the mutual information between A and C, H() is the entropy and is calculated as follows:

$$H(C) = \sum_{c \epsilon C} p(c) log \frac{1}{p(c)}$$

$$H(A, C) = \sum_{a \epsilon A, c \epsilon C} p(a, c) \log p(a, c)$$

Some internal evaluation indices are given on below:

Silhouette index: The silhouette index for point i is calculated as follows (Rousseeuw, 1987):

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Here, a(i): is the average distance between i and the other points in its group, b(i): i is the average distance between the points in the nearest group. The Silhouette index for all points is calculated as follows:

$$S = \frac{1}{N} \sum_{i=1}^{k} s(i)$$

N is the number of all points.
Silhouette index is rated in the range [-1;1]. -1 is the worst case, and 1 the best case.

Calinski & Harabasz index (CH): This index is calculated by the following formula (Calinski & Harabasz, 1974):

$$CH = \frac{SSB/(k-1)}{SSW/(N-k)}$$

SSW - a sum of squares of distances between points within a cluster, SSB - a sum of squares of distances between clusters.

When the clusters are compact and well separated, the value of this index is high, and the result is considered good.

Davies-Bouldin index (DB): With this index, which aims to minimize the center distances of the data within the cluster and maximize the distances between the clusters, the quality of clustering is calculated as follows (Davies & Bouldin, 1979):

$$DB = \frac{1}{k} \sum_{i=1}^{k} R_{ij}$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Here $i = 1,2, \dots, k$; $j = 1,2, \dots, k$; d_{ij} – distance between the centers of two clusters; s_i is the radius of the cluster i. The small value of this index is considered good for clustering.

4. Experiments

In this section, the results of the experiment for conducting a comparative analysis of the clustering quality assessment indices are given. Python programming language (Python3.8) is used to study clustering quality assessment indices. k-means, k-medoids, agglomerative hierarchical, BIRCH, OPTICS algorithms are applied to different sized data sets. Seven different sized data sets are used in experimentation. The number of clusters (k) for each set of data is known in advance. Experiments are performed for values of k from 2 to 8.

A number of external and internal evaluation indices are used to assess the results. The experiments are performed on "HTRU2", "Electrical Grid Stability Simulated Data", "Gender Gap in Spanish WP", "MAGIC Gamma Telescope", "Productivity Prediction of Garment Employees Data Set", "Shill Bidding DataSet", "HCV data" (UCI Machine Learning Repository).

The characteristics of the data are shown in Table 1.

Table 1. Database characteristics

Name	Number of points	Attributes number	Classes number
HCV	615	14	2
Product.	1197	15	2
Gender.	4746	21	3
Shill	6321	13	2
Electric.	10000	14	2
HTRU2	17898	9	2
Magic.	19020	11	2

The following tables (2-6) show the results obtained using Accuracy (Ac), Precision (Pr), Recall (Rc), F-measure (F-m), Adjusted Rand Index (ARI), Homogeneity (Hom), Completeness (Cm), Mutual Information (MI), Purity, Silhouette (Silho), Calinski & Harabasz (CH), Davies Bouldin (DB) of indices of k-means, k-medoids, agglomerative hierarchical, BIRCH, OPTICS algorithms. In the column named best, max - a large value of the index is the best solution for clustering, and min - a small value of the index is the best solution for clustering.

Table 2. Results of the agglomerative hierarchical algorithm

Method	İnd.	Best	Dataset	Number of clusters (*k*)						
				2	**3**	**4**	**5**	**6**	**7**	**8**
Agglomera. (Hierarch.)	Ac	max	HCV	**0.8766**	0.0259	0.0324	0.0454	0.0259	0.0389	0.8636
			Product.	0.2166	**0.3166**	0.3166	0.3033	0.2900	0.0266	0.0466
			Gender.	**0.4709**	0.3917	0.3167	0.2535	0.1929	0.1777	0.1954
			Shill	**0.6552**	0.3166	0.3166	0.3033	0.2900	0.0266	0.0466
			Electric.	0.3444	**0.4864**	0.1320	0.1968	0.1008	0.1940	0.1860
			HTRU2	0.0907	**0.8627**	0.0357	0.0464	0.6994	0.0404	0.0308
			Magic.	**0.6519**	0.2609	0.4548	0.0607	0.1036	0.2094	0.3554
	Pr	max	HCV	0.1776	**0.3677**	0.0485	0.3153	0.2095	0.2153	0.2641
			Product.	0.1279	**0.4200**	0.3137	0.2782	0.2325	0.0212	0.0301
			Gender.	0.3116	**0.3836**	0.2612	0.2197	0.1680	0.1535	0.1320
			Shill	**0.5004**	0.4200	0.3137	0.2782	0.2325	0.0212	0.0301
			Electric.	0.3710	**0.4579**	0.1435	0.1661	0.0889	0.1579	0.1565
			HTRU2	0.2756	**0.4358**	0.1690	0.0931	0.2684	0.1755	0.0579
			Magic.	**0.6054**	0.2574	0.3631	0.1238	0.1579	0.1393	0.1248
	Rc	max	HCV	**0.2000**	0.2000	0.0559	0.1344	0.1488	0.0603	0.1704
			Product.	0.3380	**0.3907**	0.2856	0.2219	0.1705	0.0149	0.0241
			Gender.	0.3287	**0.4416**	0.2661	0.2020	0.0950	0.0716	0.0582
			Shill	**0.5011**	0.3907	0.2856	0.2219	0.1705	0.0149	0.0241
			Electric.	**0.3743**	0.3272	0.0687	0.0897	0.0318	0.0606	0.0484
			HTRU2	0.1788	**0.3644**	0.0307	0.0269	0.1700	0.0429	0.0260
			Magic.	**0.5701**	0.2058	0.1954	0.0291	0.0338	0.0708	0.0739
	F-m	max	HCV	**0.1881**	0.0600	0.0134	0.0973	0.0996	0.0457	0.1752
			Product.	0.1854	**0.2829**	0.2100	0.1921	0.1554	0.0173	0.0268
			Gender.	0.3145	**0.3529**	0.2298	0.1683	0.1020	0.0834	0.0759
			Shill	**0.4680**	0.2829	0.2100	0.1921	0.1554	0.0173	0.0268
			Electric.	0.3440	**0.3758**	0.0857	0.1106	0.0457	0.0863	0.0720
			HTRU2	0.0890	**0.3904**	0.0196	0.0195	0.2055	0.0604	0.0186
			Magic.	**0.5629**	0.1950	0.2308	0.0460	0.0552	0.0752	0.0899
	ARI	max	HCV	0.1568	0.7096	0.7094	0.7262	0.7279	0.7311	**0.7325**
			Product.	0.2667	0.1326	0.1279	**0.2738**	0.1761	0.2058	0.2072
			Gender.	0.0041	**0.0247**	0.0106	0.0060	-0.0008	0.0068	0.0048
			Shill	0.2667	0.1326	0.1279	**0.2738**	0.1761	0.2058	0.2072
			Electric.	**0.0912**	0.0731	0.0809	0.0419	0.0522	0.0285	0.0296
			HTRU2	0.4646	**0.4737**	0.2389	0.2396	0.2434	0.1051	0.1057
			Magic.	**0.0636**	0.0434	0.0503	0.0374	0.0362	0.0375	0.0382
	Hom.	max	HCV	0.0779	0.3902	0.3902	0.4585	0.4795	0.5202	**0.5398**
			Product.	0.2822	0.2945	0.2946	0.4251	0.4265	0.4864	**0.4971**
			Gender.	0.0131	0.0222	0.0281	0.0230	0.0240	0.0248	**0.0249**
			Shill	0.2945	0.2945	0.2946	0.4251	0.4265	0.4864	**0.4971**
			Electric.	0.0485	0.0885	0.0915	0.0967	0.1135	0.1159	**0.1166**
			HTRU2	0.3161	0.3248	0.3664	0.3685	0.3714	0.3753	**0.4116**
			Magic.	0.0218	0.0237	0.0448	0.0450	0.0543	0.0602	**0.0670**
	Cm.	max	HCV	**0.6076**	0.5079	0.4971	0.4951	0.4976	0.5063	0.5075
			Product.	**0.4072**	0.3098	0.3039	0.3192	0.2793	0.2920	0.2957
			Gender.	0.0160	**0.0182**	0.0142	0.0126	0.0116	0.0111	0.0105
			Shill	**0.4072**	0.3098	0.3039	0.3192	0.2793	0.2920	0.2957
			Electric.	0.0494	**0.0534**	0.0451	0.0405	0.0423	0.0397	0.0376
			HTRU2	**0.2407**	0.2041	0.1157	0.1130	0.1072	0.0756	0.0819
			Magic.	**0.0294**	0.0158	0.0264	0.0213	0.0249	0.0267	0.0294
	MI	max	HCV	0.0690	0.3677	0.3652	0.4276	0.4466	**0.4716**	0.4712
			Product.	0.3307	0.2970	0.2918	0.3571	0.3284	0.3548	**0.3593**

			Gender.	0.0133	**0.0182**	0.0152	0.0135	0.0124	0.0116	0.0106
			Shill	0.3307	0.2970	0.2918	0.3571	0.3284	0.3548	**0.3593**
			Electric.	0.0487	**0.0662**	0.0599	0.0565	0.0609	0.0583	0.0559
			HTRU2	**0.2730**	0.2503	0.1754	0.1724	0.1657	0.1251	0.1359
			Magic.	0.0249	0.0187	0.0329	0.0285	0.0337	0.0364	**0.0402**
	Purity	max	HCV	0.8896	0.8896	0.8896	0.9090	0.9155	0.9285	**0.9350**
			Product.	0.6200	0.6200	0.6200	0.7300	0.7300	0.7533	**0.7566**
			Gender.	**0.6015**	0.6015	0.6015	0.6015	0.6015	0.6015	0.6015
			Shill	**0.8994**	0.6200	0.6200	0.7300	0.7300	0.7533	0.7566
			Electric.	0.6556	0.6556	0.6604	0.6604	0.6816	0.6816	**0.6816**
			HTRU2	0.9153	0.9231	0.9231	0.9231	0.9246	0.9246	**0.9327**
			Magic.	0.6519	0.6519	0.6799	0.6799	0.6799	0.6860	**0.6898**
	Silho.	max	HCV	**0.8936**	0.8837	0.5852	0.5864	0.5466	0.3457	0.3456
			Product.	0.6535	0.6674	**0.6754**	0.6625	0.6652	0.6667	0.5907
			Gender.	**0.4611**	0.4291	0.4072	0.4109	0.4054	0.4040	0.3878
			Shill	0.5406	0.6674	**0.6754**	0.6625	0.6652	0.6667	0.5907
			Electric.	0.2771	0.2643	0.2644	**0.2862**	0.2691	0.2531	0.2514
			HTRU2	**0.8423**	0.7445	0.7482	0.6071	0.6035	0.5987	0.4674
			Magic.	**0.5136**	0.3512	0.3610	0.3357	0.3336	0.2992	0.2852
	CH	max	HCV	233.98	276.66	**280.33**	251.48	237.31	211.41	207.50
			Product.	2152.40	1971.7	2713.2	3651.0	3686.1	3948.3	**3974.94**
			Gender.	5757.50	5777.3	5466.0	6207.4	6282.3	6242.5	**6379.60**
			Shill	**12156.4**	1971.7	2713.2	3651.0	3686.1	3948.3	3974.94
			Electric.	**1685.87**	1560.6	1400.6	1299.5	1142.2	1007.1	912.98
			HTRU2	**39372.07**	34091.02	33833.50	32562.09	30419.29	27191.08	25092.15
			Magic.	**7714.06**	6515.9	5811.1	5638.5	4994.5	4504.2	4108.15
	DB	min	HCV	1.0307	**0.6244**	1.1005	0.8707	1.1777	1.3371	1.2108
			Product.	0.5288	0.4712	0.3717	0.4227	0.3843	**0.3527**	0.4658
			Gender.	0.8064	0.8436	0.7871	0.7636	**0.7564**	0.8634	0.8639
			Shill	0.5655	0.4712	0.3717	0.4227	0.3843	**0.3527**	0.4658
			Electric.	2.3392	1.9447	2.1132	2.1950	**1.5137**	2.4179	2.4386
			HTRU2	**0.4693**	0.6871	0.7559	0.8635	1.0426	1.1475	1.3129
			Magic.	**1.2482**	1.3847	1.3172	1.3113	1.6276	1.6543	1.7524

Table 3. Results of k-means algorithm

Method	**İnd.**	**Best**	**Dataset**	**Number of clusters (*k*)**						
				2	**3**	**4**	**5**	**6**	**7**	**8**
k-means	Ac	max	HCV	**0.8766**	0.8571	0.8376	0.1103	0.0909	0.4220	0.7922
			Product.	0.2166	0.2200	**0.5766**	0.0200	0.3033	0.1200	0.3033
			Gender.	**0.4456**	0.3900	0.2030	0.2409	0.1406	0.1533	0.1600
			Shill	0.2166	0.2200	**0.5766**	0.0200	0.3033	0.1200	0.3033
			Electric.	**0.3968**	0.2584	0.3288	0.2536	0.2148	0.1996	0.0808
			HTRU2	**0.9068**	0.8545	0.8091	0.7106	0.0252	0.3110	0.3841
			Magic.	**0.6504**	0.4902	0.3125	0.2103	0.3200	0.3728	0.1888
	Pr	max	HCV	0.2441	0.3147	0.2153	**0.3571**	0.3438	0.2358	0.3094
			Product.	0.1293	**0.4633**	0.3598	0.2957	0.2735	0.1944	0.2321
			Gender.	0.3098	**0.3719**	0.2134	0.1932	0.1446	0.1460	0.1303
			Shill	0.1293	**0.4633**	0.3598	0.2957	0.2735	0.1944	0.2321
			Electric.	**0.3937**	0.2403	0.3240	0.2457	0.2085	0.1973	0.0757
			HTRU2	**0.7003**	0.4502	0.3424	0.2903	0.0765	0.1990	0.1413
			Magic.	**0.6038**	0.3377	0.2656	0.1855	0.2737	0.1463	0.1236
	Rc	max	HCV	0.2485	**0.3411**	0.2881	0.2305	0.1432	0.1380	0.1771
			Product.	0.3285	0.3400	**0.4100**	0.1854	0.2219	0.0501	0.1849
			Gender.	0.3357	**0.4318**	0.1244	0.1043	0.0779	0.0583	0.0781
			Shill	0.3285	0.3400	**0.4100**	0.1854	0.2219	0.0501	0.1849

		Electric.	**0.3864**	0.1625	0.1723	0.1086	0.0721	0.0560	0.0194
		HTRU2	**0.7036**	0.3766	0.2527	0.1998	0.0165	0.0666	0.0563
		Magic.	**0.5737**	0.3047	0.1391	0.0961	0.0917	0.0996	0.0503
F-m	max	HCV	0.2445	**0.2948**	0.2250	0.2215	0.1539	0.1666	0.2129
		Product.	0.1855	0.1914	**0.3472**	0.1282	0.1880	0.0764	0.1569
		Gender.	0.3060	**0.3487**	0.1468	0.1309	0.0866	0.0734	0.0758
		Shill	0.1855	0.1914	**0.3472**	0.1282	0.1880	0.0764	0.1569
		Electric.	**0.3848**	0.1925	0.2220	0.1485	0.1069	0.0867	0.0306
		HTRU2	**0.7019**	0.4058	0.2822	0.2352	0.0095	0.0998	0.0802
		Magic.	**0.5692**	0.3194	0.1812	0.1093	0.1233	0.1182	0.0680
ARI	max	HCV	0.1422	0.6222	**0.6456**	0.6304	0.6154	0.1998	0.5222
		Product.	**0.2925**	0.2865	0.2748	0.2430	0.2430	0.2430	0.2444
		Gender.	0.0110	**0.0145**	0.0055	0.0044	0.0073	0.0025	0.0034
		Shill	**0.2925**	0.2865	0.2748	0.2430	0.2430	0.2430	0.2444
		Electric.	**0.0422**	0.0309	0.0383	0.0336	0.0330	0.0339	0.0355
		HTRU2	0.3593	**0.4391**	0.3970	0.2541	0.2012	0.0923	0.0934
		Magic.	**0.0649**	0.0548	0.0361	0.0291	0.0302	0.0302	0.0312
Hom.	max	HCV	0.0526	0.3204	0.4218	0.5287	0.4743	0.5046	**0.5314**
		Product.	0.2952	0.2954	0.3675	0.3856	0.3856	0.3856	**0.3964**
		Gender.	0.0106	0.0452	0.0658	0.0789	0.0982	0.1107	**0.1294**
		Shill	0.2952	0.2954	0.3675	0.3856	0.3856	0.3856	**0.3964**
		Electric.	0.0369	0.0452	0.0658	0.0789	0.0982	0.1107	**0.1294**
		HTRU2	0.1687	0.3015	0.3353	0.3457	0.3527	0.3601	**0.4100**
		Magic.	0.0227	0.0652	0.0632	0.0632	0.0652	0.0639	**0.0720**
Cm.	max	HCV	0.2965	0.4042	**0.4375**	0.4366	0.4088	0.2321	0.3745
		Product.	**0.4243**	0.4127	0.3666	0.2935	0.2935	0.2935	0.2981
		Gender.	**0.0132**	0.0125	0.0106	0.0086	0.0108	0.0108	0.0109
		Shill	**0.4243**	0.4127	0.3666	0.2935	0.2935	0.2935	0.2981
		Electric.	0.0352	0.0272	0.0313	0.0323	0.0363	0.0376	**0.0411**
		HTRU2	0.1666	**0.1761**	0.1454	0.1060	0.0935	0.0701	0.0785
		Magic.	0.0291	**0.0464**	0.0355	0.0294	0.0288	0.0272	0.0262
MI	max	HCV	0.0412	0.2951	0.3939	**0.4059**	0.3743	0.1990	0.3378
		Product.	0.3456	0.3391	**0.3605**	0.3253	0.3253	0.3253	0.3305
		Gender.	0.0107	**0.0122**	0.0108	0.0085	0.0114	0.0114	0.0113
		Shill	0.3456	0.3391	**0.3605**	0.3253	0.3253	0.3253	0.3305
		Electric.	0.0358	0.0336	0.0419	0.0452	0.0522	0.0553	**0.0614**
		HTRU2	0.1673	**0.2219**	0.2023	0.1616	0.1472	0.1167	0.1311
		Magic.	0.0253	**0.0539**	0.0452	0.0398	0.0395	0.0376	0.0379
Purity	max	HCV	0.8766	0.8831	0.9155	0.9090	0.9220	0.9155	**0.9220**
		Product.	0.6300	0.6300	0.6800	0.7100	0.7100	0.7100	**0.7133**
		Gender.	0.6015	0.6015	0.6015	0.6015	0.6015	0.6015	**0.6015**
		Shill	0.6300	0.6300	0.6800	0.7100	0.7100	0.7100	**0.7133**
		Electric.	0.6276	0.6276	0.6816	0.6756	0.6752	0.6700	**0.6940**
		HTRU2	0.9153	0.9153	0.9153	0.9153	0.9195	0.9191	**0.9296**
		Magic.	0.6504	**0.6961**	0.6845	0.6841	0.6872	0.6906	0.6900
Silho.	max	HCV	**0.8870**	0.8321	0.5556	0.5922	0.4913	0.4475	0.1838
		Product.	0.6563	0.6697	0.6776	0.6765	0.6792	0.6839	**0.6878**
		Gender.	**0.5103**	0.4150	0.4306	0.4750	0.4544	0.4586	0.4433
		Shill	0.6563	0.6697	0.6776	0.6765	0.6792	0.6839	**0.6878**
		Electric.	0.3164	**0.3415**	0.3238	0.3069	0.3400	0.3368	0.3256
		HTRU2	**0.8408**	0.7798	0.7218	0.6294	0.6007	0.5562	0.5216
		Magic.	**0.5062**	0.3926	0.3951	0.3774	0.3347	0.3320	0.3416
CH	max	HCV	385.34	**596.67**	286.19	256.49	248.45	233.78	226.62
		Product.	2179.49	2010.8	3093.7	3960.3	4080.4	4427.3	**4467.75**
		Gender.	7351.18	6190.5	6498.1	7372.7	7136.5	**7585.1**	7571.86
		Shill	2179.49	2010.8	3093.7	3960.3	4080.4	4427.3	**4467.75**

			Electric.	5022.79	6456.5	6726.2	6475.3	6879.1	**6945.6**	6785.2
			HTRU2	51219.08	58019.42	63327.53	68959.33	76574.24	79552.79	**81278.50**
			Magic.	13400.30	14762.03	15461.15	**15883.87**	14892.49	14537.38	14397.53
	DB	min	HCV	0.8743	**0.5890**	1.0631	0.8377	0.9613	0.9165	1.0208
			Product.	0.5226	0.4200	0.3677	0.4017	0.3513	0.3559	**0.3462**
			Gender.	0.7355	0.9424	0.8629	0.7527	0.7510	**0.7052**	0.7596
			Shill	0.5226	0.4200	0.3677	0.4017	0.3513	0.3559	**0.3462**
			Electric.	1.2563	0.9177	0.9078	0.9672	**0.8141**	0.8188	0.8740
			HTRU2	**0.4305**	0.5365	0.5862	0.6079	0.6185	0.6318	0.6745
			Magic.	0.9058	0.8719	**0.8408**	0.8429	0.9066	0.9668	0.8935

Table 4. Results of BIRCH algorithm

Method	İnd.	Best	Dataset	Number of clusters (*k*)						
				2	3	4	5	6	7	8
BIRCH	Ac	max	HCV	**0.8831**	0.1233	0.1233	0.0259	0.0259	0.1298	0.0389
			Product.	0.2166	**0.2200**	0.2200	0.0266	0.1333	0.0233	0.0466
			Gender.	**0.4633**	0.3959	0.3016	0.2628	0.2181	0.1710	0.1499
			Shill	0.2166	**0.2200**	0.2200	0.0266	0.1333	0.0233	0.0466
			Electric.	**0.5364**	0.3100	0.3360	0.1476	0.2004	0.1576	0.1100
			HTRU2	**0.9117**	0.0534	0.0538	0.8721	0.0270	0.0337	0.0167
			Magic.	**0.8900**	0.2723	0.5066	0.2222	0.3943	0.1383	0.0910
	Pr	max	HCV	0.2776	0.2047	0.2113	0.0231	0.0231	**0.3525**	0.2136
			Product.	0.1286	**0.4620**	0.3549	0.2294	0.1987	0.0245	0.0384
			Gender.	0.3114	**0.3809**	0.2638	0.2293	0.1789	0.1273	0.1349
			Shill	0.1286	**0.4620**	0.3549	0.2294	0.1987	0.0245	0.0384
			Electric.	**0.5552**	0.2877	0.3152	0.1445	0.1539	0.1255	0.0896
			HTRU2	**0.6910**	0.3366	0.2827	0.3108	0.0877	0.0754	0.0189
			Magic.	**0.8810**	0.2533	0.2869	0.1676	0.1768	0.1534	0.1622
	Rc	max	HCV	**0.2500**	0.1251	0.1251	0.1357	0.2000	0.1281	0.1989
			Product.	0.3380	**0.3400**	0.2479	0.0228	0.0568	0.0180	0.0303
			Gender.	0.3282	**0.4362**	0.2592	0.1978	0.0876	0.0593	0.0510
			Shill	0.3380	**0.3400**	0.2479	0.0228	0.0568	0.0180	0.0303
			Electric.	**0.5579**	0.2131	0.1727	0.0566	0.0708	0.0497	0.0275
			HTRU2	**0.5890**	0.1367	0.0566	0.2240	0.0340	0.0302	0.0241
			Magic.	**0.9150**	0.2171	0.2186	0.1028	0.1124	0.0388	0.0219
	F-m	max	HCV	**0.2548**	0.1103	0.1108	0.0375	0.0407	0.1346	0.1226
			Product.	0.1861	**0.1900**	0.1455	0.0271	0.0842	0.0207	0.0338
			Gender.	0.3120	**0.3542**	0.2244	0.1720	0.1108	0.0755	0.0658
			Shill	0.1861	**0.1900**	0.1455	0.0271	0.0842	0.0207	0.0338
			Electric.	**0.4348**	0.2351	0.2148	0.0793	0.0939	0.0697	0.0419
			HTRU2	**0.6154**	0.1549	0.0897	0.2459	0.0284	0.0262	0.0185
			Magic.	**0.8861**	0.1956	0.2400	0.1080	0.1349	0.0617	0.0383
	ARI	max	HCV	0.1596	0.5002	0.5119	0.5293	0.5293	0.5294	**0.5443**
			Product.	**0.2752**	0.2692	0.1259	0.2436	0.2436	0.2436	0.2467
			Gender.	0.0009	**0.0217**	0.0102	0.0076	0.0018	0.0080	0.0031
			Shill	**0.2752**	0.2692	0.1259	0.2436	0.2436	0.2436	0.2467
			Electric.	0.0036	0.0162	0.0192	**0.0320**	0.0153	0.0193	0.0222
			HTRU2	0.2144	0.4191	0.4202	**0.4242**	0.2704	0.2713	0.2724
			Magic.	**0.6081**	0.0585	0.2400	0.1080	0.1349	0.0617	0.0383
	Hom.	max	HCV	0.0742	0.3927	0.4331	0.4841	0.4841	0.4903	**0.5283**
			Product.	0.2865	0.2866	0.2979	0.3890	0.3890	0.3890	**0.4157**
			Gender.	0.0111	0.0197	0.0216	0.0231	0.0232	0.0239	**0.0263**
			Shill	0.2865	0.2866	0.2979	0.3890	0.3890	0.3890	**0.4157**
			Electric.	0.0097	0.0113	0.0406	0.0606	0.0880	0.1029	**0.1069**

		HTRU2	0.0663	0.2344	0.2408	0.2607	0.2958	0.3315	**0.3589**
		Magic.	**0.6090**	0.0275	0.0548	0.0562	0.0571	0.0807	0.1044
Cm.	max	HCV	**0.5785**	0.4042	0.4853	0.4366	0.3991	0.3257	0.2505
		Product.	**0.4128**	0.3555	0.3598	0.3600	0.2791	0.3471	0.3373
		Gender.	0.0135	**0.0135**	0.0080	0.0098	0.0098	0.0106	0.0104
		Shill	**0.4128**	0.3555	0.3598	0.3600	0.2791	0.3471	0.3373
		Electric.	0.0094	0.0226	**0.0500**	0.0443	0.0338	0.0374	0.0341
		HTRU2	0.1202	**0.1216**	0.0737	0.0680	0.0631	0.0534	0.0522
		Magic.	**0.5725**	0.0037	0.0166	0.0180	0.0246	0.0182	0.0293
MI	max	HCV	0.1166	0.3307	**0.4378**	0.4033	0.4222	0.3433	0.3069
		Product.	0.3356	0.3604	0.3890	0.4030	0.3318	**0.4208**	0.4113
		Gender.	0.0111	**0.0133**	0.0076	0.0098	0.0100	0.0111	0.0107
		Shill	0.3356	0.3604	0.3890	0.4030	0.3318	**0.4208**	0.4113
		Electric.	0.0093	0.0278	**0.0672**	0.0621	0.0485	0.0550	0.0507
		HTRU2	0.0850	**0.1741**	0.1190	0.1127	0.1067	0.0919	0.0950
		Magic.	**0.5902**	0.0044	0.0221	0.0248	0.0351	0.0266	0.0438
Purity	max	HCV	0.8831	0.8831	0.8961	0.9090	0.9220	0.9090	**0.9285**
		Product.	0.6233	0.7133	0.7300	0.7566	0.7333	**0.7866**	0.7833
		Gender.	**0.6015**	0.6015	0.6015	0.6015	0.6015	0.6015	0.6015
		Shill	0.6233	0.7133	0.7300	0.7566	0.7333	**0.7866**	0.7833
		Electric.	0.6276	0.6276	0.6456	0.6620	0.6632	0.6708	**0.6752**
		HTRU2	**0.9153**	0.9153	0.9153	0.9153	0.9184	0.9153	0.9153
		Magic.	**0.8900**	0.6403	0.6403	0.6807	0.6851	0.6609	0.6874
Silho.	max	HCV	**0.8936**	0.2484	0.2354	0.2676	0.4913	0.2546	0.1911
		Product.	0.6535	**0.6593**	0.6508	0.3962	0.3086	0.4011	0.3634
		Gender.	**0.4611**	0.4543	0.4174	0.4556	0.4312	0.4036	0.3897
		Shill	0.6535	**0.6593**	0.6508	0.3962	0.3086	0.4011	0.3634
		Electric.	0.2871	**0.3415**	0.3245	0.3088	0.3316	0.3247	0.3089
		HTRU2	**0.8303**	0.6848	0.5650	0.5092	0.4957	0.4718	0.4637
		Magic.	**0.5994**	0.0044	0.0221	0.0248	0.0351	0.0266	0.0438
CH	max	HCV	**380.99**	129.29	98.30	119.30	248.45	85.62	77.39
		Product.	**2164.06**	1906.7	1653.0	755.46	686.13	589.52	509.26
		Gender.	5757.50	6026.2	5870.5	**7014.1**	6911.0	6182.6	6617.67
		Shill	**2164.06**	1906.7	1653.0	755.46	686.13	589.52	509.26
		Electric.	4496.09	6455.2	6714.5	6366.1	**6824.8**	6771.4	6530.39
		HTRU2	28927.97	47099.43	47238.35	46581.91	**55227.00**	53219.09	51733.36
		Magic.	**47954.73**	14278.3	12621.1	15536.7	9886.6	8811.7	12766.1
DB	min	HCV	**0.8778**	1.0465	1.0668	0.9867	0.9613	1.1640	1.1628
		Product.	0.5272	**0.4634**	0.5329	0.8037	0.8721	1.0470	1.0684
		Gender.	0.8064	0.8980	0.8044	**0.7402**	0.7659	0.9255	0.8360
		Shill	0.5272	**0.4634**	0.5329	0.8037	0.8721	1.0470	1.0684
		Electric.	1.3067	0.9187	0.9034	0.9006	**0.8126**	0.8557	0.8793
		HTRU2	**0.3288**	0.5851	0.5954	0.6033	0.6181	0.6222	0.6254
		Magic.	**0.5007**	0.8974	0.8949	0.8859	0.9441	0.9186	0.9021

Table 5. Results of the k-medoids algorithm

Method	İnd.	Best	Dataset	Number of clusters (*k*)						
				2	3	4	5	6	7	8
k-medoids	Ac	max	HCV	**0.8766**	0.8571	0.0519	0.0909	0.8246	0.7792	0.1038
			Product.	0.2200	**0.6200**	0.5800	0.5333	0.4766	0.4600	0.4600
			Gender.	**0.4490**	0.3437	0.2611	0.2367	0.1566	0.1398	0.1145
			Shill	0.2200	**0.6200**	0.5800	0.5333	0.4766	0.4600	0.4600
			Electric.	**0.4160**	0.2988	0.2380	0.1736	0.1380	0.1336	0.1240
			HTRU2	**0.9074**	0.1530	0.0846	0.0726	0.0516	0.0422	0.0297

		Magic.	**0.8468**	0.3223	0.1575	0.1390	0.1213	0.0927	0.1099
Pr	max	HCV	0.2441	**0.3147**	0.1681	0.1492	0.4167	0.2180	0.2579
		Product.	0.1358	**0.5448**	0.3685	0.2760	0.2801	0.1973	0.1727
		Gender.	0.3116	**0.3153**	0.2448	0.2068	0.1556	0.1557	0.1276
		Shill	0.1358	**0.5448**	0.3685	0.2760	0.2801	0.1973	0.1727
		Electric.	**0.4075**	0.3020	0.2373	0.1784	0.1226	0.1159	0.1085
		HTRU2	**0.7077**	0.2624	0.1979	0.2048	0.1316	0.0789	0.0589
		Magic.	**0.8485**	0.3626	0.1836	0.2077	0.1706	0.1299	0.1250
Rc	max	HCV	0.2485	0.3411	0.1088	0.1177	**0.4420**	0.2694	0.1636
		Product.	0.3431	**0.5902**	0.3994	0.2809	0.2179	0.1755	0.1536
		Gender.	**0.3376**	0.3122	0.1841	0.1303	0.0666	0.0626	0.0344
		Shill	0.3431	**0.5902**	0.3994	0.2809	0.2179	0.1755	0.1536
		Electric.	**0.4015**	0.2049	0.1247	0.0696	0.0395	0.0330	0.0269
		HTRU2	**0.7351**	0.0876	0.1332	0.0920	0.0465	0.0076	0.0049
		Magic.	**0.8817**	0.2331	0.0788	0.0529	0.0399	0.0252	0.0265
F-m	max	HCV	0.2445	0.2948	0.0852	0.0738	**0.3934**	0.2268	0.1583
		Product.	0.1944	**0.5343**	0.3616	0.2630	0.2271	0.1747	0.1529
		Gender.	**0.3082**	0.2861	0.1877	0.1420	0.0848	0.0740	0.0495
		Shill	0.1944	**0.5343**	0.3616	0.2630	0.2271	0.1747	0.1529
		Electric.	**0.4008**	0.2409	0.1602	0.0993	0.0598	0.0513	0.0430
		HTRU2	**0.7202**	0.0945	0.0999	0.1012	0.0442	0.0127	0.0080
		Magic.	**0.8437**	0.2657	0.1066	0.0838	0.0639	0.0418	0.0434
ARI	max	HCV	0.1422	0.6222	**0.6960**	0.6016	0.5617	0.4877	0.2134
		Product.	0.3244	0.3546	**0.3726**	0.3521	0.2722	0.3138	0.3096
		Gender.	**0.0101**	0.0056	0.0030	0.0009	-5.441	3.3841	0.0003
		Shill	0.3244	0.3546	**0.3726**	0.3521	0.2722	0.3138	0.3096
		Electric.	0.0276	0.0303	**0.0626**	0.0526	0.0291	0.0322	0.0288
		HTRU2	**0.3918**	0.2817	0.0926	0.0748	0.0600	0.0342	0.0340
		Magic.	**0.4812**	0.0020	0.0250	0.0240	0.0312	0.0239	0.0298
Hom.	max	HCV	0.0526	0.3204	0.4453	0.4668	0.5289	0.4656	**0.5719**
		Product.	0.3043	0.3744	0.4371	0.4770	0.4378	**0.5658**	0.5655
		Gender.	0.0104	0.0170	0.0130	0.0179	0.0206	0.0241	**0.0253**
		Shill	0.3043	0.3744	0.4371	0.4770	0.4378	0.5658	**0.5655**
		Electric.	0.0279	0.0376	0.1049	0.1077	0.0912	**0.1098**	0.1066
		HTRU2	0.2042	0.3086	0.3135	0.3372	0.3550	0.3454	**0.3584**
		Magic.	**0.5222**	0.0062	0.0347	0.0418	0.0638	0.0527	0.0906
Cm.	max	HCV	0.2965	0.4042	**0.4853**	0.4366	0.3991	0.3257	0.2505
		Product.	**0.4350**	0.3555	0.3598	0.3600	0.2791	0.3471	0.3373
		Gender.	0.0129	**0.0135**	0.0080	0.0098	0.0098	0.0106	0.0104
		Shill	**0.4350**	0.3555	0.3598	0.3600	0.2791	0.3471	0.3373
		Electric.	0.0266	0.0226	**0.0500**	0.0443	0.0338	0.0374	0.0341
		HTRU2	**0.1858**	0.1216	0.0737	0.0680	0.0631	0.0534	0.0522
		Magic.	**0.4890**	0.0037	0.0166	0.0180	0.0246	0.0182	0.0293
MI	max	HCV	0.0706	0.3307	**0.4378**	0.4033	0.4222	0.3433	0.3069
		Product.	0.3555	0.3604	0.3890	0.4030	0.3318	**0.4208**	0.4113
		Gender.	0.0104	**0.0133**	0.0076	0.0098	0.0100	0.0111	0.0107
		Shill	0.3555	0.3604	0.3890	0.4030	0.3318	**0.4208**	0.4113
		Electric.	0.0269	0.0278	**0.0672**	0.0621	0.0485	0.0550	0.0507
		HTRU2	**0.1943**	0.1741	0.1190	0.1127	0.1067	0.0919	0.0950
		Magic.	**0.5050**	0.0044	0.0221	0.0248	0.0351	0.0266	0.0438
Purity	max	HCV	0.8766	0.8831	0.8961	0.9090	0.9220	0.9090	**0.9285**
		Product.	0.6400	0.7133	0.7300	0.7566	0.7333	**0.7866**	0.7833
		Gender.	0.6015	0.6015	0.6015	0.6015	0.6015	0.6015	**0.6015**
		Shill	0.6400	0.7133	0.7300	0.7566	0.7333	**0.7866**	0.7833
		Electric.	0.6276	0.6276	0.6456	0.6620	0.6632	0.6708	**0.6752**
		HTRU2	0.9153	0.9153	0.9153	0.9153	0.9184	0.9153	**0.9153**

		Magic.	**0.8468**	0.6403	0.6403	0.6807	0.6851	0.6609	0.6874
	Silho.	max	HCV	**0.6106**	0.2484	0.2354	0.2676	0.4913	0.2546	0.1911
			Product.	0.6475	**0.6593**	0.6508	0.3962	0.3086	0.4011	0.3634
			Gender.	**0.5103**	0.4543	0.4174	0.4556	0.4312	0.4036	0.3897
			Shill	0.6475	**0.6593**	0.6508	0.3962	0.3086	0.4011	0.3634
			Electric.	0.3126	**0.3415**	0.3245	0.3088	0.3316	0.3247	0.3089
			HTRU2	**0.8335**	0.6848	0.5650	0.5092	0.4957	0.4718	0.4637
			Magic.	**0.6261**	0.0044	0.0221	0.0248	0.0351	0.0266	0.0438
	CH	max	HCV	216.11	129.29	98.30	119.30	**248.45**	85.62	77.39
			Product.	**2115.41**	1906.7	1653.0	755.46	686.13	589.52	509.26
			Gender.	**7351.18**	6026.2	5870.5	7014.1	6911.0	6182.6	6617.67
			Shill	**2115.41**	1906.7	1653.0	755.46	686.13	589.52	509.26
			Electric.	4904.66	6455.2	6714.5	6366.1	**6824.8**	6771.4	6530.39
			HTRU2	49148.44	47099.43	47238.35	46581.91	**55227.00**	53219.09	51733.36
			Magic.	**57005.1**	14278.3	12621.1	15536.7	9886.6	8811.7	12766.1
	DB	min	HCV	0.9964	1.0465	1.0668	**0.9867**	0.9613	1.1640	1.1628
			Product.	0.5101	**0.4634**	0.5329	0.8037	0.8721	1.0470	1.0684
			Gender.	**0.7355**	0.8980	0.8044	0.7402	0.7659	0.9255	0.8360
			Shill	0.5101	**0.4634**	0.5329	0.8037	0.8721	1.0470	1.0684
			Electric.	1.2699	0.9187	0.9034	0.9006	**0.8126**	0.8557	0.8793
			HTRU2	**0.4779**	0.5851	0.5954	0.6033	0.6181	0.6222	0.6254
			Magic.	**0.5007**	0.8974	0.8949	0.8859	0.9441	0.9186	0.9021

Table 6. Results of OPTICS algorithm

Method	İnd.	Best	Dataset	Number of clusters (*k*)						
				2	**3**	**4**	**5**	**6**	**7**	**8**
OPTICS	Ac	max	HCV	0.0324	0.0324	0.0324	0.0324	**0.0974**	0.0974	0.0974
			Product.	0.0033	0.0033	0.0033	0.0033	**0.0066**	0.0066	0.0066
			Gender.	0.0025	0.0025	0.0050	0.0050	0.0101	0.0101	**0.0109**
			Shill	0.0033	0.0033	0.0033	0.0033	**0.0066**	0.0066	0.0066
			Electric.	0.0016	0.0016	0.0048	0.0048	0.0048	0.0048	**0.0060**
			HTRU2	0.0008	0.0008	0.0008	**0.0044**	0.0044	0.0044	0.0044
			Magic.	0.0010	0.0008	0.0008	0.0010	0.0014	0.0014	**0.0016**
	Pr	max	HCV	**0.1666**	0.1666	0.1666	0.1666	0.1666	0.1666	0.1666
			Product.	0.0045	0.0045	0.0047	0.0055	0.0142	**0.0166**	0.0166
			Gender.	0.0069	0.0071	0.0095	0.0099	0.0140	**0.0159**	0.0152
			Shill	0.0045	0.0045	0.0047	0.0055	0.0142	**0.0166**	0.0166
			Electric.	0.0085	0.0092	0.0206	0.0363	0.0555	0.1054	**0.1892**
			HTRU2	0.0047	0.0049	0.0055	0.0069	0.0089	0.0109	**0.0138**
			Magic.	0.0051	0.0074	0.0094	0.0133	0.0192	0.0250	**0.0312**
	Rc	max	HCV	0.0061	0.0061	0.0061	0.0061	**0.0185**	0.0185	0.0185
			Product.	0.0002	0.0002	0.0002	0.0003	0.0022	**0.0026**	0.0026
			Gender.	5.0546e-05	5.203e-05	9.905e-05	0.0001	**0.0002**	0.0002	0.0002
			Shill	0.0002	0.0002	0.0002	0.0003	0.0022	**0.0026**	0.0026
			Electric.	4.5244e-05	4.895e-05	0.0002	0.0004	0.0006	0.0012	**0.0019**
			HTRU2	4.60e-06	4.85e-06	5.42e-06	3.53e-05	4.56e-05	5.61e-05	**7.07e-05**
			Magic.	**1.5457**	9.8032e-06	1.2392e-05	2.1893e-05	4.4208e-05	5.7471e-05	8.2101e-05
	F-m	max	HCV	0.0119	0.0119	0.0119	0.0119	**0.0333**	0.0333	0.0333
			Product.	0.0005	0.0005	0.0005	0.0006	0.0038	**0.0045**	0.0045
			Gender.	0.0001	0.0001	0.0001	0.0002	0.0004	0.0004	**0.0005**
			Shill	0.0005	0.0005	0.0005	0.0006	0.0038	**0.0045**	0.0045
			Electric.	8.9909	9.7278	0.0004	0.0008	0.0012	0.0023	**0.0039**
			HTRU2	9.20e-06	9.70e-06	1.08e-05	7.04e-05	9.08e-05	**0.0001**	0.0001
			Magic.	3.0823e-05	1.9580e-05	2.4752e-05	4.3715e-05	8.8214e-05	**0.0001**	0.0001

ARI	max	HCV	**-0.0487**	-0.048	-0.048	-0.048	-0.100	-0.100	-0.1004
		Product.	0.0760	0.0760	0.0847	0.0776	**0.1066**	0.1044	0.1044
		Gender.	0.0010	0.0009	**0.0025**	-0.003	-0.005	-0.003	-0.0071
		Shill	0.0760	0.0760	0.0847	0.0776	**0.1066**	0.1044	0.1044
		Electric.	-0.0016	-0.001	-0.003	-0.010	-0.003	**0.0016**	-0.0001
		HTRU2	-0.0107	-0.013	-0.016	-0.011	**0.0007**	-0.002	-0.0163
		Magic.	-0.0006	-0.038	-0.037	-0.032	-0.028	-0.024	**0.0306**
Hom.	max	HCV	0.0080	0.0080	0.0080	0.0080	**0.0250**	0.0250	0.0250
		Product.	**0.4481**	0.4481	0.4436	0.4164	0.3736	0.3322	0.3322
		Gender.	**0.1077**	0.1070	0.1016	0.1001	0.0910	0.0770	0.0775
		Shill	**0.4481**	0.4481	0.4436	0.4164	0.3736	0.3322	0.3322
		Electric.	**0.0709**	0.0685	0.0606	0.0368	0.0258	0.0141	0.0101
		HTRU2	**0.2430**	0.2343	0.2203	0.1989	0.1908	0.1610	0.1235
		Magic.	**0.2567**	0.0823	0.0716	0.0592	0.0468	0.0412	0.0362
Cm.	max	HCV	0.0303	0.0303	0.0303	0.0303	**0.0424**	0.0424	0.0424
		Product.	0.1930	0.1930	0.1952	0.1938	**0.2022**	0.1988	0.1988
		Gender.	0.0248	**0.0249**	0.0234	0.0227	0.0204	0.0183	0.0184
		Shill	0.1930	0.1930	0.1952	0.1938	**0.2022**	0.1988	0.1988
		Electric.	0.0531	**0.0532**	0.0527	0.0486	0.0477	0.0450	0.0446
		HTRU2	**0.0336**	0.0331	0.0325	0.0334	0.0365	0.0352	0.0318
		Magic.	**0.0881**	0.0448	0.0437	0.0448	0.0453	0.0481	0.0493
MI	max	HCV	**-0.0125**	-0.012	-0.012	-0.012	7.8348	7.8348	7.8348
		Product.	0.2313	0.2313	0.2340	0.2321	**0.2356**	0.2245	0.2245
		Gender.	-0.0009e-05	-8.20e-05	1.8113e-05	**0.0009**	7.4919e-05	-0.001	8.8179e-05
		Shill	0.2313	0.2313	0.2340	0.2321	**0.2356**	0.2245	0.2245
		Electric.	0.0412	**0.0417**	0.0404	0.0312	0.0258	0.0170	0.0133
		HTRU2	0.0399	0.0391	0.0387	0.0413	**0.0473**	0.0449	0.0388
		Magic.	**0.1147**	0.0409	0.0400	0.0399	0.0375	0.0373	0.0358
Purity	max	HCV	**0.8766**	0.8766	0.8766	0.8766	0.8766	0.8766	0.8766
		Product.	**0.7633**	0.7633	0.7633	0.7500	0.7366	0.7166	0.7166
		Gender.	**0.6208**	0.6208	0.6192	0.6183	0.6149	0.6149	0.6133
		Shill	**0.7633**	0.7633	0.7633	0.7500	0.7366	0.7166	0.7166
		Electric.	**0.6560**	0.6552	0.6516	0.6396	0.6376	0.6344	0.6320
		HTRU2	**0.9340**	0.9331	0.9318	0.9305	0.9305	0.9278	0.9244
		Magic.	**0.7375**	0.6561	0.6527	0.6511	0.6481	0.6481	0.6479
Silho.	max	HCV	**-0.2072**	-0.208	-0.231	-0.333	-0.387	-0.400	-0.4078
		Product.	-0.0238	0.2313	0.2340	0.2321	**0.2356**	0.2245	0.2245
		Gender.	**-0.0357**	-0.041	-0.0640	-0.0936	-0.1417	-0.2142	-0.2442
		Shill	-0.0238	-0.023	-0.041	-0.092	**-0.018**	-0.146	-0.1773
		Electric.	-0.1255	-0.130	-0.143	-0.172	-0.215	**0.250**	-0.277
		HTRU2	**-0.1462**	-0.146	-0.159	-0.197	-0.243	-0.289	-0.3251
		Magic.	-0.4595	**-0.144**	-0.156	-0.194	-0.233	-0.267	-0.2944
CH	max	HCV	5.4203	5.5357	**6.1745**	1.5527	1.9178	2.1096	2.4034
		Product.	16.11	16.21	16.73	**17.61**	10.96	11.19	13.56
		Gender.	15.85	16.34	16.88	17.480	18.52	19.44	**20.31**
		Shill	16.11	16.21	16.73	**17.61**	10.96	11.19	13.56
		Electric.	12.06	12.27	12.77	13.49	14.36	15.42	**15.99**
		HTRU2	13.73	14.62	14.11	15.78	15.23	16.54	**17.57**
		Magic.	12.21	13.76	14.25	15.15	16.21	18.34	**19.70**
DB	min	HCV	2.2407	2.2647	2.3823	3.3334	**2.2067**	2.4761	2.4740
		Product.	1.9412	2.0677	2.0753	**1.6510**	1.6811	1.8260	2.5566
		Gender.	1.3336	1.3165	**1.2953**	1.3194	1.3362	1.3376	1.2989
		Shill	1.9412	2.0677	2.0753	**1.6510**	1.6811	1.8260	2.5566
		Electric.	**1.5902**	1.6038	1.6166	1.6507	1.6600	1.6443	1.6903
		HTRU2	2.2208	2.1834	**2.1425**	2.1504	2.1845	2.2079	2.1610
		Magic.	**3.2155**	13.76	14.25	15.15	16.21	18.34	19.70

5. Analysis

Initially, the results obtained using a hierarchical algorithm are analyzed according to Table 2. The results obtained by the indices Ac, Pr, Rc, F-m and Cm are considered effective. Although incorrect results are obtained with ARI, Silho, CH, and DB indices for small and medium data, accurate results are obtained for big data. With Hom, MI and Purity, the results are completely wrong.

According to Table 3, the results obtained using the k-means algorithm are analyzed. While the Ac, Pr, Rc, F-m, ARI, and Silho indices are considered to be effective for small and medium data, the results are quite accurate for big data. An accurate result is obtained with the Cm index. With CH, DB, Hom, MI, and Purity indices, completely wrong results are obtained.

According to Table 4, the results obtained using the BIRCH algorithm are analyzed. The Ac, Pr, Rc, F-m, Silho, and DB indices provide accurate results for big data. With ARI, MI, CH, Cm, Hom, and Purity indices, the number of classes is estimated incorrectly.

According to Table 5, the results obtained using the k-medoids algorithm are analyzed. The Ac, Pr, Rc, F-m, ARI, Cm, MI, and Silho indices provide accurate results for big data. The results obtained with the Hom, Purity, CH, and DB indices are completely incorrect.

Using the OPTICS method, only the number of classes with the Hom and Purity indices is accurately estimated. With other indices, completely incorrect results are obtained.

Based on the results of the experiment, a comparative analysis of cluster quality assessment indices is analyzed. There are 35 cases based on 7 sets of data and 5 algorithms.

The overlap of the results of the cluster quality assessment indices is investigated.

According to the previous tables, the following table is obtained.

Table 7. Number of common results of indices

	Ac	**Pr**	**Rc**	**F-m**	**ARI**	**Hom**	**Cm**	**MI**	**Purity**	**Silho**	**CH**	**DB**
Ac	-											
Pr	21	-										
Rc	**26**	24	-									
F-m	**26**	**25**	**30**	-								
ARI	9	11	12	9	-							
Hom	2	1	3	2	3	-						
Cm	14	8	11	10	12	6	-					
MI	5	8	7	7	9	9	18	-				
Purity	5	4	4	3	3	25	6	13	-			
Silho	19	12	14	14	8	6	10	6	10	-		
CH	8	6	10	8	4	5	8	3	6	6	-	
DB	14	11	15	14	6	6	9	4	5	14	13	-

As can be seen from Table 7:

- Ac, Pr, Rc, and F-m show similar results in most cases (Rc and F-m indices are 86% - 30 out of 35, Ac, Pr, and F-m indices are 73% - 26 out of 35).
- Hom shows different results with the Ac, Pr, F-m in most cases (96% - 2 out of 35).
- Silho shows a similar result to the DB (40% - 14 out of 35).
- Silho contradict the CH (80% - 6 out of 35).
- Silho and Ac (58% - 19 out of 35) show similar results.
- CH and MI (86% - 3 out of 35) show different results.

12 indices are comparatively analyzed. Similar and opposite results are determined from them. It can be concluded from here that the results of Ac, Pr, Rc, and F-m indices overlap.

6. Conclusion

In this paper, cluster analysis, its applications, external and internal indices of cluster quality assessment were studied and compared. Thus, different size data sets were

taken, k-means, k-medoids, agglomerative hierarchical, BIRCH, and OPTICS algorithms were applied to them. Some external and internal evaluation indices were used to assess the quality of clustering.

12 indices were examined and compared in the article. According to the results of the experiment, Ac, Pr, Rc and F-m indices show similar results in in group determining in a given clustering structure. It can be a good reference for people who intend to contribute to the study of research and machine learning based on the evaluation of the clustering quality in data sets in the future. The result is practically important in various areas of human activity (medicine, business, science, analysis of social networks, etc.) assessing the quality of obtained clusters by applying the clustering algorithm to data.

References

Aliguliyev, R. M. (2009). "Performance evaluation of density-based clustering methods". Information Sciences, 179(20): 3583-3602. https://doi.org/10.1016/j.ins.2009.06.012

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern recognition, 46(1), 243-256. https://doi.org/10.1016/j.patcog.2012.07.021

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. Machine learning, 50(1), 5-43. https://doi.org/10.1023/A:1020281327116ss

Ankerst, M., Breunig, M., Kriegel, H., Sander, J. (1999) OPTICS:ordering points to identify the clustering structure. ACM SIGMOD Record , pp 49–60. https://doi.org/10.1145/304181.304187

Alguliyev, R. M., Aliguliyev, R. M., & Abdullayeva, F. J. (2019). Privacy-preserving deep learning algorithm for big personal data analysis. Journal of Industrial Information Integration, 15, 1-14. https://doi.org/10.1016/j.jii.2019.07.002

Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-28349-8_2

Chana I. A., & Arora S. (2014). Survey of clustering techniques for big data analysis. 5th International Conference - Confluence the Next Generation Information Technology Summit, 59-65. https://ieeexplore.ieee.org/abstract/document/6949256/

Chandra, E., & Anuradha, V. P. (2011). A survey on clustering algorithms for data in spatial database management systems. International Journal of Computer Applications, *24*(9), 19-26. 10.5120/2969-3975

Aggarwal, C. C., & Reddy, C. K. (2013). Data Clustering: Algorithms and Applications, ser. https://dokumen.tips/data-analytics/data-clustering-algorithms-and-applications.html?page=1

Mammadova, L. R. (2021). Finding the optimal number of clusters in hierarchical clustering. https://2021.nscf.ru/TesisAll/05_AI_MachineLearning/260_MammadovaLeRa.pdf

Ahmadov, E. (2021). Comparative Analysis of K-Means, K-Means++ and Mini Batch K-Means Algorithms in Python Environment. Problems of information technology, 5(2), 3-16 (in Azeri). https://www.sciencegate.app/document/10.25045/jpit.v12.i2.11

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *14*(1). https://www.researchgate.net/publication/311925975_Clustering_Algorithms_A_Comparative_Approach

Mammadova, L. (2021). Some external evaluation indices for clustering. The 2nd international scientific conferences of students and young researchers dedicated to the 98th anniversary of the National Leader of Azerbaijan Heydar Aliyev, pp. 445-446 (in Azeri). http://www.bhos.edu.az/noduploadeditor/files/Tezisler_2021_17x24sm_Final%20(1).pdf

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier., 457-460. https://www.google.com/books?hl=ru&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=Data+mining:+concepts+and+techniques.&ots=tAIx1-mz_Y&sig=jZzODUwidvm-Xau17UWmMTMKJog

Alguliev, R. M., & Aliguliyev, R. M. (2005). Fast genetic algorithm for clustering of text documents, Artificial Intelligence 3, 698–707 (in Russian). https://doi.org/10.1155/2011/416308

Aliguliyev, R. M. (2007). Automatic document summarization by sentence extraction, Journal of Computational Technologies 12, 5–15. https://cyberleninka.ru/article/n/automatic-document-summarization-by-sentence-extraction

Aliguliyev, R. M. (2006). A novel partitioning-based clustering method and generic document summarization. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops (pp. 626-629). IEEE. https://ieeexplore.ieee.org/abstract/document/4053329

Sajana, T., Rani, C. S., & Narayana, K. V. (2016). A survey on clustering techniques for big data mining. Indian journal of Science and Technology, 9(3), 1-12. 10.17485/ijst/2016/v9i3/75971

Zhang, T., Ramakrishnan, R. & Livny, M. (1997).BIRCH: A New Data Clustering Algorithm and Its Applications. Data Mining and Knowledge Discovery **1,** 141–182 (1997). https://doi.org/10.1023/A:1009783824328

Parimala, M., Lopez, D., & Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. International Journal of advanced science and technology, 31(1), 59-66. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.643.6121&rep=rep1&type=pdf

Zhang, Q., & Couloigner, I. (2005). A new and efficient k-medoid algorithm for spatial clustering. In International conference on computational science and its applications

(pp. 181-189). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/11424857_20

UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets.php

Rafsanjani, M. K., Varzaneh, Z. A., & Chukanlo, N. E. (2012). A survey of hierarchical clustering algorithms. The Journal of Mathematics and Computer Science, pp. 229-240. https://www.researchgate.net/publication/281377211_A_Survey_Of_Hierarchical_Clustering_Algorithms

Hamerly, G. J. (2003). Learning structure and concepts in data through data clustering. University of California, San Diego. https://www.proquest.com/openview/60c11b639be3a8be3f419b66322251a5/1?pq-origsite=gscholar&cbl=18750&diss=y

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), 846-850. https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356

Kvalseth, T. O. (1987). Entropy and correlation: Some comments. IEEE Transactions on Systems, Man, and Cybernetics, 17(3), 517-519. https://ieeexplore.ieee.org /abstract/document/4309069/

Rosenberg, A., & Hirschberg, J. (2007, June). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) (pp. 410-420). https://aclanthology.org/D07-1043.pdf

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. The computer journal, 26(4), 354-359. https://academic.oup.com /comjnl/article-abstract/26/4/354/377434

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. https://www.sciencedirect.com/science/article/pii/0377042787901257

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27. https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227. https://ieeexplore. ieee.org/abstract/document/4766909/