2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

An Entropy evaluation method of hierarchical clustering

Quan Tu, Tianyang Xu, Tingting Fang, WenWen Wang, Jie Jiang, Ping Zhu†
School of Science, Jiangnan University
Wuxi, China
e-mail: zhuping@jiangnan.edu.cn

***Abstract*—Based on the agglomerative hierarchical clustering algorithm, this paper proposes a new information entropy evaluation indicator—Average Discriminant Entropy(ADE), to measure the stability of cluster structure. After that, We designed the corresponding algorithm. In order to verify the validity of the indicator, six heterogeneous artificial data sets were used to simulate. By comparing ADE with other classic evaluation indicators, we found that ADE can obtain the best results under various data sets. Finally, a Monte Carlo experiment on the data with different noise levels proved the robust of ADE.**

***Keywords*-hierarchical cluster; information entropy; ADE; Monte-Carlo experiment**

I. INTRODUCTION

With the advent of the era of big data, data mining becomes more important [1]. The purpose of data mining is to extract useful information from the massive data. As one of the most important tool of data mining, the cluster method can divide a large amount data into several clusters, which contains abundant potential information. Therefore, the cluster method has been applied to statistics, machine learning, biology, marketing and so on [2]–[4].

The cluster method can be divided into partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods [5]. Particularly, the hierarchical clustering method can extract the information of different hierarchical structures, it is often used in the construction of biological evolution trees and plays an important role in bioinformatics. The hierarchical clustering method initially takes every sample as a cluster, and two clusters is combined to a new cluster until all samples are contained in a cluster or meet a certain termination condition. Nevertheless, how to determine the optimal cluster number is a main problem [7].

In recent years, a lot of indicators have been proposed to solve the problem, such as Davies-Bouldin(DB) index, Dunn index, etc [8]. However, most of them aren't fit for all data sets with different feature. In 1948, Shannon proposed the information entropy [9] to describe the uncertainty of system. In this paper, we introduce the information entropy for the evaluation of cluster. The Average Discriminant Entropy(ADE) indicator is developed to achieve it.

This paper is organized as follows. In the second section, we review the clustering method and hierarchical clustering method. Section 3 introduces a few of external validity indicators and internal effectiveness indicators for clustering. The fourth part introduces the information entropy and proposes the ADE indicator, and then gives the corresponding algorithm based on hierarchical clustering. Six heterogeneous artificial data sets are used to compare three classical indicators with ADE indicators, the result shows that ADE indicator perform well. Subsequently, a Monte Carlo experiment on the data with different noise levels is applied to prove the robust of ADE. Finally, a brief summary of this article is given.

II. THE INTRODUCTION OF CLUSTERING

A. Clustering

Considering the following data set:

$$\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{i,p}]^T \in R^p \quad (1)$$

$$X = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_N^T]^T \in R^{N\times p} \quad (2)$$

where $\boldsymbol{x}_i$ is the feature vector of the sample i, X is the set which consists of N samples, and the dimension for feature vectors is p.

The task of clustering is to classify these samples to get structure $\boldsymbol{C}$ without supervision, which requires [10]:

$$\begin{cases} \boldsymbol{C} = \{C_1, C_2, \cdots, C_{\mathrm{Q}}\}, \quad Q \le N \\ C_i \neq \emptyset, \quad i = 1, 2, \cdots Q \\ \bigcup_{i=1}^{Q} C_i = X \\ C_i \bigcap C_j = \emptyset, \quad i, j = 1, 2, \cdots Q, \quad i \neq j \end{cases} \quad (3)$$

where $\boldsymbol{C}$ represents the set of all clusters, and C_i represents each cluster.

B. Hierarchical Clustering

As an important branch of clustering method, hierarchical clustering is divided into agglomerated hierarchical clustering and divisive hierarchical clustering. The main difference of them depends on whether the process is top-down or bottom-up. Next we will focus on agglomerated hierarchical clustering, which is shown in (4), Where $\boldsymbol{H}$ represents the hierarchical structure:

$$\begin{cases} \boldsymbol{H} = \{H_1, H_2, \ldots, H_Q\}, Q \le N \\ H_i = \{\mathrm{C}_i^1, \mathrm{C}_i^2, \ldots, \mathrm{C}_i^{N-i+1}\} \\ H_i \cup H_{i+1} - H_i \cap H_{i+1} = \{\mathrm{C}_i^m, \mathrm{C}_i^n, \mathrm{C}_{i+1}^p\} \\ \mathrm{C}_i^m \cup \mathrm{C}_i^n = \mathrm{C}_{i+1}^p \end{cases} \quad (4)$$

In addition, H_i has to satisfy (3) as $\boldsymbol{C}$.

In general,the hierarchical clustering is shown in pedigree figure [11], as Fig.1:

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00066

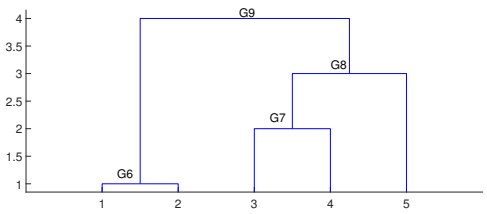

Fig. 1. A pedigree figure

III. Traditional evaluation indicators for clustering

To determine the optimal number of clusters, it is necessary to use the validity indicator to evaluate the clustering. Traditional evaluation indicators for clustering are divided into external and inner validity indicator. Because the external validity index needs the priori information, which is hard to meet, we just introduce several classic inner indicators.

A. Internal effectiveness indicator

1) Davies-Bouldin Index(DB)

In DB index [12], R_{km} represent the resolution among clusters:

$$R_{km} = (s_k + s_m)/d(\boldsymbol{o}_k, \boldsymbol{o}_m) \tag{5}$$

where s_k and s_m is the similarity in cluster C_k and C_m and $\boldsymbol{o}_k$ and $\boldsymbol{o}_m$ is the cluster center. s_k is calculated as follows:

$$s_k = \frac{\sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, \boldsymbol{o}_k)}{n_k} \tag{6}$$

where n_k is the sample number in cluster C_k. The DB is obtained as:

$$DB = \sum_{k=1}^{N_c} R_k/N_k, R_k = \max_{m=1\ldots C_k, m\neq k} R_{km} \tag{7}$$

In which N_c is the number of clusters, and for each cluster k, there is a maximum isolation R_k with other clusters. The smaller the DB value, the better clustering is.

2) Dunn index

As for Dunn index [13], the compactness in cluster is described by the cluster diameter and the distance among clusters stands for the isolation. The Dunn index is shown as follows:

$$D = \min_{k=1\ldots N_c} \{ \min_{m=1\ldots N_c} [\frac{D(k,m)}{\min\limits_{h=1\ldots N_c} diam(C_h)}] \} \tag{8}$$

The diameter is the largest distance among samples in a cluster:

$$diam(C_h) = \max_{x_i, x_j \in C_h} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{9}$$

The result is more reliable with a larger value of Dunn index.

3) COP coefficient

$$COP = \frac{1}{n} \sum_{C_k \in C} n_k \frac{(1/n_k) \sum_{i=1}^{n_k} d(\boldsymbol{x}_i, \boldsymbol{o}_k)}{\min\limits_{\boldsymbol{x}_j \notin C_l} \max\limits_{\boldsymbol{x}_i \in C_l} d_(\boldsymbol{x}_i, \boldsymbol{x}_j)} \tag{10}$$

where $\boldsymbol{o}_k$ is the cluster center, n_k is the number of samples in cluster k. The COP coefficient is better with a smaller value [14].

Considering most of internal effectiveness indicators aren't fit for data set with different type, we propose the Average Discriminant Entropy(ADE) under the inspiration of Shannon Entropy, which can evaluate the clustering in a new perspective.

IV. Entropy Evaluation Indicator

A. Entropy

Shannon proposed information entropy for the first time in 1948. Information entropy can be used to measure the uncertainty of the system, information entropy of discrete random variables is defined as follows:

$$H(X) = -\sum_{x \in X} p(x) \log_a p(x) \tag{11}$$

It is stipulated that all information entropy below uses bit as the unit, and the base is omitted.

Since the information entropy can describe the degree of chaos, we use it to extract the structure and approach the problem of determining the optimal number of clusters.

B. Average Discriminant Entropy (ADE)

Although the paper put forward clustering validity indicator based on the hierarchical clustering, it is also inspired by the K-means algorithm.

The K-means algorithm specifies the number of clusters and divides all sample points into the nearest cluster. The algorithm don't finish until all clusters does not change in a loop. An important step in this process is to calculate the distance from the sample point to each cluster center and select the nearest cluster center after comparison. If the distances between one sample point and two or more cluster centers are equal, the sample can be divided into different clusters. This situation can be understood as a greater uncertainty of this sample. We believe that an optimal cluster structure should have a low uncertainty, so information entropy is introduced to quantify the structural stability of the system.

In hierarchical clustering, the distance between the sample and it's center of the cluster is the smallest relative to the centers of the other clusters. Meanwhile, the minimum distance between the sample point and the center of the other clusters is selected. If the minimum distance to other clusters is close to the distance between the sample and it's cluster center, the uncertainty of this sample is relatively large, otherwise the uncertainty is small. In this way, we can obtain the Average Discriminant Entropy (ADE) of clustering evaluation indicator based on information entropy.

In a certain cluster structure of hierarchical clustering, suppose that there are N samples and c clusters. C_i denotes the cluster to which sample i belongs, C_k is another cluster. There exist $i = 1 \ldots N$ and $k, m \in \{1 \ldots c\}$.

Define a binary random variable $V(i)$ for each sample i:

$$\begin{cases} V(i,1) = \dfrac{d(x_i, C_i)}{d(i, C_i) + \min\limits_{i \notin Cm} d(x_i, C_m)} \\ V(i,2) = 1 - V(i,1) \end{cases} \tag{12}$$

The Discriminant Entropy(DE) for sample i is expressed as:

$$DE(i) = H(V(i)) \tag{13}$$

So the ADE for the structure is the average of DE of all samples, which is shown as:

$$ADE = \frac{\sum_{i=1}^{i=N} DE(i)}{N} \tag{14}$$

It is concise and applicable to get the information from the structure by using ADE. When the ADE is smaller, the less complex is the structure, that is to say, the extracted structure is stable. Next, the corresponding algorithm is designed on the base of ADE.

C. Algorithm design

A good designed algorithm is adaptable for different systems, it's vital to extract the structure of data when the characteristics of the data itself is complex. Fig.2 shows two cluster structures with different stability. As can been

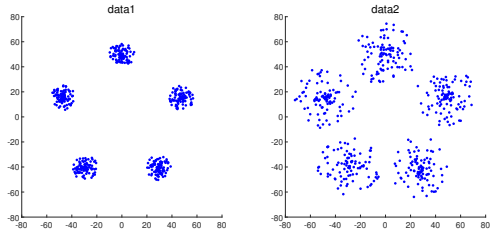

Fig. 2. Two structures with different stability

seen from Fig.2, the optimal cluster number is 5. However with the influence of structure complex, data 1 has is more obvious to be clustered into five clusters than data 2. In different systems, ADE can identify the structure exactly. Considering the fact that the 1 or N cluster has little research value, besides, the corresponding ADE is zero, so we don't take these situations into account.

The algorithm flows of entropy evaluation based on hierarchical is shown in **Algorithm 1**. Accordingly, The optimal cluster number can be worked out, from which we can get most information of a system.

V. Simulation

On the base of hierarchical cluster, we selected and compared the effect of DB Index, Dunn Index, COP coefficient with ADE using the artificial datas. The artificial data sets we design used to confirm the efficiency is shown as follows: In Fig.3, six types of data sets can clearly determine the optimal number of clusters. The first one is standard data set, and the other five are data sets

Algorithm 1 Framework of ADE to extract structure for complex system.

Input: The set of samples, $X = \left[\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_N^T\right]$;
Output: Ensemble of optimal clusters, C; The correspondingly ADE of the optimal structure;

1: Initial ensemble of clusters, $C = [C_1, C_2, \ldots, C_N]$;
2: Calculating the center of clusters with the help of C and X;
3: Obtaining the distance of every sample to all clusters;
4: Getting the ADE of current structure;
5: Updating the ensemble of clusters: combining the two closest clusters into one;
6: the number of clusters is $n = n - 1$, if $n \neq 1$, go to 2;
7: **return** C and ADE;

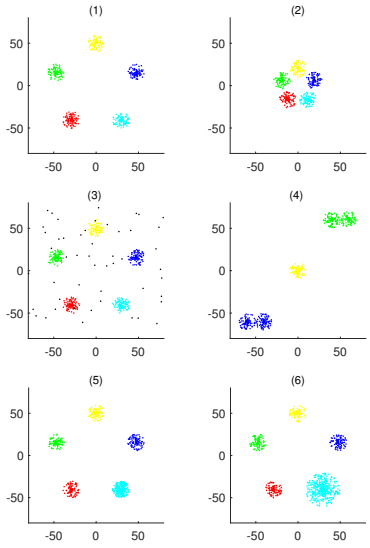

Fig. 3. The artificial data

adding different perturbations based on the standard data set. The disturbance factors are respectively inter-cluster separation, noise, sub-cluster, density and cluster radius. By testing six types of data sets, the result can fully reflect the comprehensive performance of the ADE indicator.

It is easy to compare and analyze with the evaluation results of various indicators by using artificial data sets, for the optimal clustering number of the sample can be gotten directly. The optimal clustering numbers of above six types of data sets are $5, 5, 5, 3, 5, 5$. Among 6 types of data set, the optimal clustering number of the fourth, which is influenced by sub-cluster, is controversial, that is to say, why it is not 5 but 3. At the scale of Fig.4, it can be seen that the two categories with sub-clusters can be clearly divided into two categories. On a larger scale, the sub-clusters will be in the same category and the

sub-clusters are not separated. However, the three major categories still have obvious separation characteristics, so in our artificial data sets, the optimal number of clusters for the fourth category of data is 3.

Table I to VI show the clustering evaluation results for the six types of artificial data sets when the number of clusters M is 2 to 6. The bold ones are the optimal values of the corresponding indicators.

It can be seen from Table I to VI, the evaluation results of Dunn index and COP coefficient for data set 3 are not satisfying, and the result of the COP indicator on the data set 4 also deviates from the real value. DB index and ADE perform well on six data sets. It indicates that the ADE is more adjustable to multiple systems compared to other selected indicators. Furthermore, the value of ADE indicator can reflects the stability of system, which has a practical significance. Therefore, concluded from the above analysis, the ADE indicator can be widely applied in the evaluation of clustering.

Table I
CLUSTER EVALUATION RESULTS OF DATA SET 1

	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
DB	1.0386	0.6681	0.4482	**0.1701**	0.4560
$DUNN$	0.3567	0.5187	0.5187	**2.0437**	0.0891
COP	100.8738	51.3525	29.7428	**8.5112**	11.1527
ADE	0.8772	0.7731	0.5890	**0.3792**	0.4752

Table II
CLUSTER EVALUATION RESULTS OF DATA SET 2

	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
DB	1.0676	0.7231	0.5503	**0.4277**	0.6877
$DUNN$	0.0832	0.1110	0.1112	**0.2396**	0.0497
COP	102.5458	55.7836	37.9334	**20.6919**	0.8403
ADE	0.8724	0.8038	0.7276	**0.6359**	0.6868

Table III
CLUSTER EVALUATION RESULTS OF DATA SET 3

	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
DB	1.0516	0.6859	0.4806	**0.2602**	0.4985
$DUNN$	0.1008	0.1113	0.1562	0.1637	**0.2055**
COP	103.7948	51.3515	32.4342	12.0899	**11.6058**
ADE	0.8737	0.7808	0.6183	**0.4343**	0.4646

Table IV
CLUSTER EVALUATION RESULTS OF DATA SET 4

	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
DB	0.3590	**0.1991**	0.3491	0.4458	0.7159
$DUNN$	0.5204	**1.4012**	0.0909	0.0734	0.0359
COP	51.6602	22.1823	21.5122	**23.0624**	23.0876
ADE	0.5864	**0.4792**	0.5278	0.5912	0.6492

Next, in order to analyze the robustness of ADE indicators, Monte-Carlo repeat experiments were introduced. For the standard data in data set 1, there exist 5 clusters in total, and the center of each cluster is taken from five points evenly distributed on a circle with $(0, 0)$ as the center and 50 as the radius. The cluster radius is 10 and there randomly distribute 100 scattered points in each cluster. For noise points (x_k, y_k), let x_k and y_k are taken from a uniformly distributed random sequence obeying $[-80, 80]$. Repeat the test 100 times with noise ratios of 5%, 10%, 15% and 20%. To prevent large deviations, the number of candidate clusters is set 2 from to 11. It is verified whether the indicators can get the optimal clusters under the above conditions. Among them, one set of data is taken under each of the four noise levels, and the results are shown in Fig. 4. It can be seen that the data sets under each level of noise still have obvious characteristics of 5 categories:

Table V
CLUSTER EVALUATION RESULTS OF DATA SET 5

	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
DB	0.8546	0.6145	0.4458	**0.1763**	0.4055
$DUNN$	0.3481	0.3481	0.5130	**1.9947**	0.0375
COP	127.2204	67.2145	31.7492	**16.5361**	36.1592
ADE	0.6946	0.5997	0.5323	**0.3910**	0.5773

Table VI
CLUSTER EVALUATION RESULTS OF DATA SET 6

	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$
DB	0.9203	1.5595	0.4614	**0.2199**	0.4252
$DUNN$	0.2646	0.2660	0.3917	**0.7621**	0.0292
COP	139.4750	85.5576	48.2121	**33.2816**	54.1409
ADE	0.7303	0.8093	0.6006	**0.4700**	0.6053

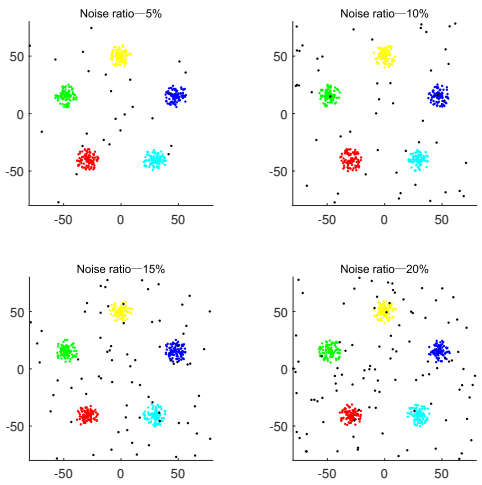

Fig. 4. Data sets at each noise level

Perform 100 random experiments each, describe the optimal number of clusters during each experiment with data points, and mark inaccurate results with red circles, as shown in Fig. 5:

Among them, the accuracy of evaluation at each level is 93%, 90%, 89% and 81%, that is, the clustering indicator has a accuracy of close to 90% when the noise ratio is lower than 15%. Therefore, the ADE indicator is considered to have good robustness and relatively stable and accurate in the evaluation of clustering.

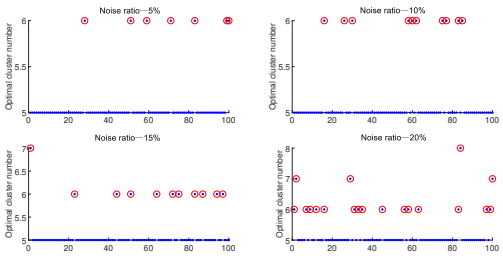

Fig. 5. Results of 100 random experiments at each noise level

VI. Conclusion

In this paper, we put forward an ADE indicator to quantify the uncertainty of a system in clustering method, and design a corresponding algorithm by introducing hierarchical clustering. Comparing DB indicator, Dunn indicator, COP coefficient and ADE with six artificial data sets, it's obvious that the ADE indicator has a comprehensive and outstanding performance. Furthermore, to analyze the robustness of ADE indicators, A Monte-Carlo repeat experiment is carried out. The method proposed in this paper can be extended to other fields, such as commerce, biology, Internet etc., which has a bright future.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 11271163) and the Funds of College Students' Innovative Entrepreneurial Training Plan Program(No.201910295047)

The authors would like to thank the Associate Editor and the anonymous reviewers for their constructive and helpful comments and suggestions to improve the quality of this paper.

References

[1] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *INFORMATION SCIENCES*, vol. 275, pp. 314–347, AUG 10 2014.

[2] M. Templ, P. Filzmoser, and C. Reimann, "Cluster analysis applied to regional geochemical data: Problems and possibilities," *Applied Geochemistry*, vol. 23, no. 8, pp. 2198 – 2213, 2008.

[3] S. Sedita, A. Caloffi, and L. Lazzeretti, "The invisible college of cluster research: a bibliometric core cperiphery analysis of the literature," *Industry and Innovation*, vol. 27, no. 5, pp. 562–584, 2020.

[4] B. J. Winterhoff, M. Maile, A. K. Mitra, A. Sebe, M. Bazzaro, M. A. Geller, J. E. Abrahante, M. Klein, R. Hellweg, S. A. Mullany, K. Beckman, J. Daniel, and T. K. Starr, "Single cell sequencing reveals heterogeneity within ovarian cancer epithelium and cancer associated stromal cells," *Gynecologic Oncology*, vol. 144, no. 3, pp. 598 – 606, 2017.

[5] P. Haldar, I. D. Pavord, D. E. Shaw, M. A. Berry, M. Thomas, C. E. Brightling, A. I. Wardlaw, and R. H. Green, "Cluster analysis and clinical asthma phenotypes," *AMERICAN JOURNAL OF RESPIRATORY AND CRITICAL CARE MEDICINE*, vol. 178, no. 3, pp. 218–224, AUG 1 2008.

[6] K. Biswas, J. He, I. D. Blum, C.-I. Wu, T. P. Hogan, D. N. Seidman, V. P. Dravid, and M. G. Kanatzidis, "High-performance bulk thermoelectrics with all-scale hierarchical architectures," *NATURE*, vol. 489, no. 7416, pp. 414–418, SEP 20 2012.

[7] N. M. Kopelman, J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose, "Clumpak: a program for identifying clustering modes and packaging population structure inferences across K," *MOLECULAR ECOLOGY RESOURCES*, vol. 15, no. 5, pp. 1179–1191, SEP 2015.

[8] M. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *PATTERN RECOGNITION*, vol. 37, no. 3, pp. 487–501, MAR 2004.

[9] J. LIN, "DIVERGENCE MEASURES BASED ON THE SHANNON ENTROPY," *IEEE TRANSACTIONS ON INFORMATION THEORY*, vol. 37, no. 1, pp. 145–151, JAN 1991.

[10] H. Teichgraeber and A. R. Brandt, "Systematic comparison of aggregation methods for input data time series aggregation of energy systems optimization problems," in *13th International Symposium on Process Systems Engineering (PSE 2018)*, ser. Computer Aided Chemical Engineering, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds. Elsevier, 2018, vol. 44, pp. 955 – 960.

[11] H. Hexmoor, "Chapter 6 - diffusion and contagion," in *Computational Network Science*, ser. Emerging Trends in Computer Science and Applied Computing, H. Hexmoor, Ed. Boston: Morgan Kaufmann, 2015, pp. 45 – 64.

[12] R. M. Alguliyev, R. M. Aliguliyev, and L. Sukhostat, V, "Weighted consensus clustering and its application to Big data," *EXPERT SYSTEMS WITH APPLICATIONS*, vol. 150, JUL 15 2020.

[13] N. Ilc, "Modified Dunn's cluster validity index based on graph theory," *PRZEGLAD ELEKTROTECHNICZNY*, vol. 88, no. 2, pp. 126–131, 2012.

[14] L. Ozgener, "Coefficient of performance (COP) analysis of geothermal district heating systems (GDHSs): Salihli GDHS case study," *RENEWABLE & SUSTAINABLE ENERGY REVIEWS*, vol. 16, no. 2, pp. 1330–1334, FEB 2012.