A Survey of Advances in Hierarchical Clustering Algorithms and Applications

Amr Munshi†
aaamunshi@uqu.edu.sa
† *Computer Engineering Department, Umm Al-Qura University, Saudi Arabia*

Summary
Hierarchical clustering methods have been proposed for more than sixty years and yet are used in various disciplines for relation observation and clustering purposes. In 1965, divisive hierarchical methods were proposed in biological sciences and have been used in various disciplines such as, and anthropology, ecology. Furthermore, recently hierarchical methods are being deployed in economy and energy studies. Unlike most clustering algorithms that require the number of clusters to be specified by the user, hierarchical clustering is well suited for situations where the number of clusters is unknown. This paper presents an overview of the hierarchical clustering algorithm. The dissimilarity measurements that can be utilized in hierarchical clustering algorithms are discussed. Further, the paper highlights the various and recent disciplines where the hierarchical clustering algorithms are employed.

Keywords:
Fuzzy partitioning, electric grid, photovoltaic, solar panel

1. Introduction

Clustering algorithms can be divided generally into two main categories: partitional and hierarchical. Partitional algorithms divide the data into non-overlapping clusters. The most commonly known partitional algorithm is k-means. Hierarchical algorithms can either be agglomerative (bottom-up) or divisive (top-down) [1]. An agglomerative algorithm starts with considering each data point as an individual cluster, and then similar clusters are merged at successive steps. Conversely, divisive hierarchical algorithms start with all data points being in one cluster and then they are split successively until each data point is an individual cluster.

Hierarchical clustering methods have been proposed for more than sixty years and yet are used in various disciplines for relation observation and clustering purposes. In 1965, divisive hierarchical methods were proposed in biological sciences [2] and have been used in various disciplines such as, anthropology [3], and ecology [4]. Also, several agglomerative hierarchical methods have been discussed by [5] in 1973 [6]. In addition, hierarchical methods are much used for data reduction purposes. Unlike most clustering algorithms that require the number of clusters to be specified by the user, hierarchical clustering is well suited for situations where the number of clusters is unknown. It only requires a dissimilarity measurement criterion to decide which clusters should be grouped together [7]. However, various dissimilarity measurements may yield different clustering results even when the same dataset is used [8].

This paper presents an overview of the hierarchical clustering algorithm, and is organized as the following. Section 2 illustrates the dissimilarity measurements that can be utilized in hierarchical clustering algorithms. The types of agglomerative hierarchical clustering algorithms and how they function is presented in Section 3. Section 4 presents the divisive hierarchical clustering algorithm. Section 5 highlights various disciplines where the hierarchical clustering algorithm is employed. The concluding remarks are mentioned in Section 6.

2. Dissimilarity Measurements

There are various distance metrics that can express the (dis)similarity between pairs of data points. The most commonly used distance metric is the Euclidean distance. The Euclidean distance is the squared root of the sum of the squared difference between the variables' values and is given by [8]:

$$d(i,K) = \sqrt{\sum_{j=1}^{n} (xij - xkj)^2} \quad (1)$$

Another alternative distance measure is the Manhattan distance or sometimes called city-block distance. It calculates the distance between data points by using the sum of the variables' absolute values and is given by:

$$d(i,K) = \sum_{j=1}^{n} |xij - xkj| \quad (2)$$

The Chebyshev distance is a similarity matric that is suitable when working with ordinal data sets and is given by the following formula:

$$d(i,K) = \max\left(\sum_{j=1}^{n} |xij - xkj|, \sum_{j=1}^{n} |yij - ykj|\right) \quad (3)$$

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.4

The Chebyshev distance calculates the maximum of the absolute difference in the clustering variables' values [8]. Fig. 1 illustrates the interrelation between the Euclidean, Manhattan and Chebyshev distances in a two-dimensional space between two-points.

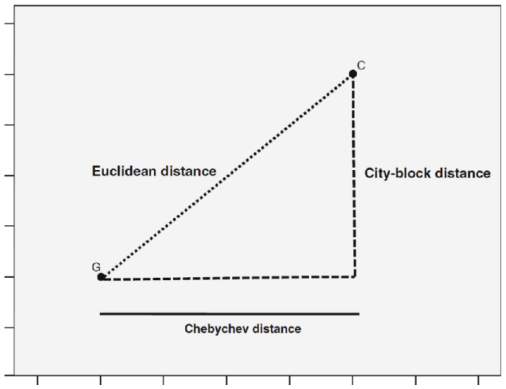

Figure 1: The interrelation between the Euclidean, Manhattan (City-block) and Chebyshev distances in a two-dimensional space between point G and C. Retrieved from [8]

In data sets that contain uncorrelated features, the Mahalanobis distance which is equivalent to the Euclidean distance can be used to measure the similarity between data points. In many situations this distance metric causes some computational burden [9]. The Mahalanobis distance formula is given by:

$$d(i,K) = \quad (xi - xj)^T S^{-1}(xi - xj) \qquad (4)$$

The choice of the distance metric to use is not critical in improving the underlying structure of clusters, whereas, the choice of the clustering algorithm (next section) is much more important [8].

3. Agglomerative Hierarchical Clustering

After determining the dissimilarity measure based on the variables of the dataset, a clustering algorithm that groups similar data points together is to be applied. A widely used hierarchical clustering approach is agglomerative clustering. It starts with considering each data point as an individual cluster, and merges a selected pair of clusters at successive steps. The choice of merging a pair of clusters is based on the smallest intergroup dissimilarity [7]. Eventually, all clusters are combined into a single cluster. This recursive combining of clusters can be represented in a convenient tree-like structure called a dendrogram (Fig. 2). Each level of the hierarchy represents a particular grouping of data objects into disjoint clusters. It is a user task to decide which level represents the desired clustering formation and how many clusters are desired in a sense that observations (data points) within each cluster are sufficiently more similar to each other than to those in other groups [7].

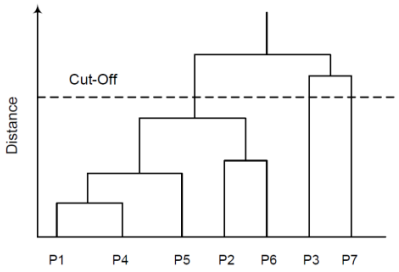

Figure 2: A dendrogram representing hierarchical clustering. Retrieved from [10]

The steps to perform a general agglomerative hierarchical algorithm are:

1. Assign each data point to a separate cluster.
2. Evaluate all pair-wise distances between clusters.
3. Construct a distance matrix using distance metrics.
4. Look for the pair of clusters with the shortest distance.
5. Merge the pair of clusters and remove them from the distance matrix.
6. Evaluate all distances from this new cluster to all other clusters, and update the distance matrix.
7. Repeat from step 2 until the all the clusters are grouped into one cluster.

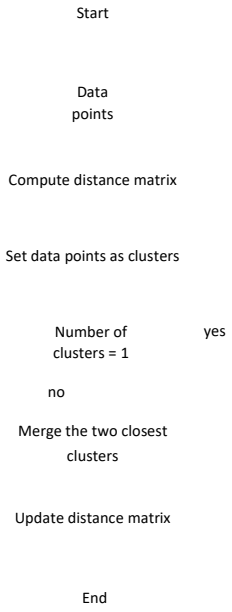

Figure 3: Flow chart of the agglomerative hierarchical clustering algorithm

3. 1. Single Linkage

The single linkage (nearest-neighbour) is based on the minimum distance between two data points in two different clusters. In other words, it merges clusters based on the most similar data points from each cluster. Single linkage tends to form clusters that may lead to heterogeneous data points clustered together [11]. This procedure is sensitive to outliers, as a new data point can extremely alter the hierarchical clustering structure [12].

3. 2. Complete Linkage

Complete linkage (farthest-neighbour) is based on the maximum distance between two data points in two different clusters. The cluster similarity is based on the most dissimilar data points from each cluster. It tends to form compact sphere-like clusters [11]. This procedure finds compact clusters with small diameters; however, some data points in a certain cluster may be much closer to other clusters than the other data points in its cluster [12].

3. 3. Average Linkage

The average linkage also called UPGMA (un-weighted pair-group method using arithmetic averages) is a compromise of single and complete linkages. It is based on the average distance between all the pairs of data points of two clusters. In other words, it calculates the minimum and maximum of all the pairwise distances between data points of two clusters to average them. Consequently, the resulted clusters tend to almost have equal within cluster variability [11].

3. 4. Centroid Linkage

In this procedure the centroid, which is an existing representative data point, of each cluster is determined first. Then the merging is based on the distance between the two centroids.

Each linkage procedure has its properties and can present entirely different results, even when applied on the same dataset. In general, the single linkage procedure is considered to be the most versatile. Contrariwise, the complete linkage procedure is significantly affected by outliers as it is based on the maximum distances. Average linked and centroid linked procedures produce similar size clusters with low within cluster variance. Also, both procedures are affected by outliers but not as much as complete linkage [8].

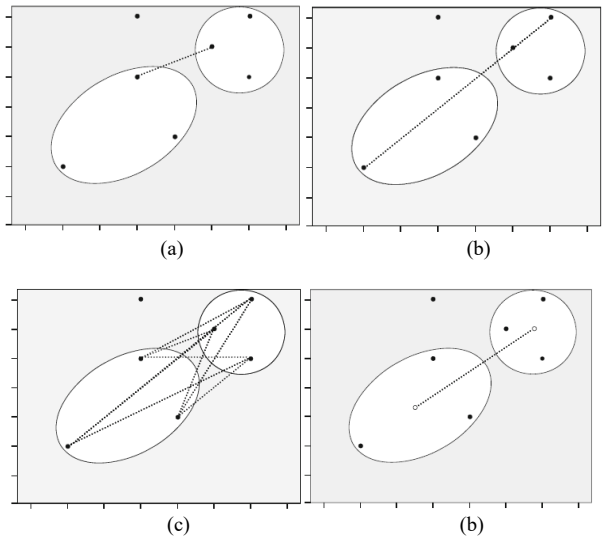

Figure 4: Agglomerative hierarchical clustering procedures (a) single linkage (b) complete linkage (c) average linkage (d) centroid linkage. Retrieved from [8]

Table 1: Formulas and computational complexity of agglomerative hierarchical clustering procedures

Algorithm	Formula	Complexity [16]	Capability of handling high dimensional data
Single linkage	$D(X,Y) = \min_{x\in X, y\in Y} d(x,y)$	O(n^2)	No
Complete linkage	$D(X,Y) = \max_{x\in X, y\in Y} d(x,y)$	O(n^2)	No
Average linkage	$D(X,Y) = \frac{1}{\lvert X\rvert . \lvert Y\rvert}\sum_{x\in X}\sum_{y\in Y} d(x,y)$	O(n^2)	No
Centroid linkage	$D(X,Y) = \lvert\lvert \bar{X} - \bar{Y}\rvert\rvert^2$	O(n^2)	No
Ward's	$D(X,Y) = \frac{\lvert\lvert \bar{X} - \bar{Y}\rvert\rvert^2}{\frac{1}{Nx} + \frac{1}{Ny}}$	O(n^2)	No

3. 5. Ward's Method

Another commonly used procedure in hierarchical clustering is Ward's method. This procedure differs from the other mentioned procedures, as it utilizes the variance to evaluate the distances between clusters. It merges clusters if such merging increases the overall within cluster variance to the smallest possible degree. In general, Ward's method is considered to be efficient [8]. In addition, it is appropriate to use when equally sized clusters are expected and the data set is free from outliers [8].

4. Divisive Hierarchical Clustering

Divisive hierarchical clustering begins with the entire data set in a single cluster, and then the most dissimilar clusters are split-off recursively into two clusters. It continues splitting until each cluster represents its own [11]. This algorithm has not been used as widely as agglomerative algorithms in the clustering literature. A possible utilization of this algorithm is when the interest is to partition the data set into a relatively small number of clusters.

The steps to perform a general divisive hierarchical algorithm are [13]:

1. Start with all data points in one cluster.
2. Calculate the diameter (maximum distance between data points) of each cluster and choose the cluster with the maximal diameter.
3. Calculate the distances between data points in that cluster.
4. Split where the most dissimilar data point occurs.
5. Repeat from step 2 until each data point is represented into an individual cluster.

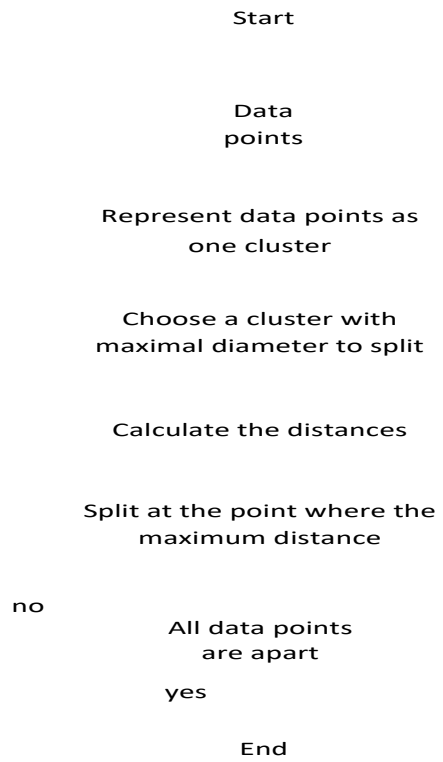

Figure 5: Flow chart of the divisive hierarchical clustering algorithm

The divisive algorithm is conceptually more complex than the agglomerative algorithm. Its computational complexity is considered to be high with by O(2^n). Divisive algorithms can be applied by using another algorithm in the process as a subroutine. This is by recursively employing combinational methods such as k-means or k-medoids, with

the number of clusters set to 2 *(k=2)* at each iteration. However, in this case the splitting depends heavily on the starting configuration at each step [11]. Accordingly, the splitting results may differ in each run, even when using the same data set.

5. Applications of Hierarchical Clustering

The significant growth of available data in various disciplines motivates the adaption of data mining techniques to promote innovative research. The application of various clustering techniques on a dataset in order to investigate the appropriate method for establishing classification is a common practise in data mining [14]. The clustering results of different clustering algorithms can be evaluated using validity indices to select the most efficient clustering algorithm for a particular dataset. Hierarchical clustering is considered to be one of the efficient and widely used methods in various research disciplines. The following sub-sections are some fields of studies that hierarchical clustering had a significant role in research conductance.

5. 1. Gene Clustering [15]

Microarrays are tools to obtain genomic information in a comparative and parallel way. Observing biological activities and cellular changes under various conditions at molecular level by conducting microarray experiments poses various challenging statistical and computational issues. A major obstacle in the process is gene clustering. The main purpose of gene clustering is to search for the group of genes that have similar patterns of biological functions or interactions.

Hierarchical clustering has been powerful in clustering genes and samples, and is considered to be a standard tool for such purposes. However, it suffers from providing solutions based on the iterative mergences between pair-wise distances instead of a global criterion. In addition, hierarchical clustering suffers from the lack of robustness and cut procedures to find the number of clusters. However, hierarchical clustering can be beneficial to visualize global patterns that express the data, but not suitable to present cluster information for further biological exploration.

5. 2. Microarray Data Analysis on Ovarian Cancer [16]

Ovarian cancer is considered to be the second most common cause of death for gynecological cancers, excluding breast cancer, for women in western countries [17]. The cause of this particular type of cancer is unknown. The survival rate is low due to the significant spread of this cancer beyond the ovaries at diagnosis. Moreover, ovarian cancer may recur to women who have had a complete response to treatment. Many attempts have been conducted to detect ovarian cancer at early stages, such as measuring serum CA125 antibody concentrations. CA125 is considered to be robust in following the progression of the disease; however, it cannot be used as a diagnostic marker.

The analysis of the ovarian cancer microarray data can be involved in order to detect molecular markers of such cancer at early stages. In this research, the performance of three unsupervised clustering algorithms namely; SOM, Fuzzy c-means and hierarchical clustering were employed to analyze the ovarian cancer microarray data. The dataset used included 15 samples with 9600 genes. These samples included 5 benign ovarian tumors (OVT), 1 borderline ovarian malignancy (OVTT), 4 ovarian cancers at stage I (OVCAI), and 5 ovarian cancers at stage III (OVCAIII). A regression analysis was used to reduce the dimensions of the dataset and obtain 9600 residuals of genes. The 100 largest and 100 smallest residuals of genes were chosen for analyzing using the analysis of variance (ANOVA). The ANOVA analysis presented 12 gene markers that can be used to distinguish between the tumors and cancers: OVT, OVTT, OVCAI and OVCAIII. The 12 gene markers were performed clustering by the three clustering algorithms and the results were compared. The average hierarchical clustering algorithm with Euclidean distance presented the best performance in distinguishing between the OVT, OVTT, OVCAI and OVCAIII.

5. 3. Classifying Electricity Customers [18]

Identifying the consumption patterns of customers and grouping them together based on their load diagram in order to formulate tariff offers is of interest to electricity service providers. This research investigated the effectiveness of various clustering algorithms namely; modified follow-the-leader, k-means, fuzzy k-means, hierarchical (average distance and Ward's criterion) and SOM in clustering a set of 234 non-residential load diagrams. The clustering results present a representative load diagram for each customer class. Further, load profiles for tariff purposes can be established. The results of this research show that the modified follow-the-leader and the hierarchical clustering based on the average distance linkage criterion algorithms presented the best clustering results. Both algorithms provided remarkable detailed separation between clusters. However, the modified follow-the-leader was the most efficient algorithm as it comprised clustering adequacy and computational speed.

5. 4. Protein Sequence Clustering [19]

In biomedical research the functions and structures of protein sequences have remarkable significance. It is time and resource consuming to understand the function and structure of unknown protein molecules. Proteins with similar sequences often belong to the same protein family and have similar functions and structures. Thus, predicting the sequence of an unknown protein can aid in classifying which family it belongs to and accordingly use the common functions and structures of that family as its estimates. This process can be automated in order to robustly predict the family group of any unknown protein sequences. For that, similar protein sequences are clustered together into groups based on their sequence homology and then a representative model for each group can be built.

This research proposed an unsupervised approach for protein sequences clustering. It utilizes the hierarchical clustering algorithm based on the single-linkage criterion to pre-cluster the protein sequences in the first phase. In the second phase it adopts a partitional clustering algorithm to refine the clustering results. The experimental results of this research demonstrated the robustness and effectiveness of the proposed model in clustering protein sequences and accordingly, understanding its function and structure.

5. 5. Malware Categorization [20]

Malware attacks such as viruses, backdoors, spyware, trojans and worms, present security threats to computer users. The common defense tool against malwares is anti-virus products. Anti-virus products detect, remove and characterize those malwares. The characterization of those malware depends on a method to categorize the properties of malware into groups. For this purpose, clustering malware into various groups is of interest for computer security research.

This research proposed a parameter-free hybrid clustering algorithm (PFHC) based on the hierarchical clustering and k-means clustering algorithms for malware clustering. It utilizes the agglomerative hierarchical clustering algorithm as the frame, starting with singleton clusters that include one sample. Then it reuses the centroids of upper level in every level and merges the two nearest clusters. Finally, it adopts the k-means clustering algorithm for iteration to obtain an approximate global optimal division. The experimental results suggest that this algorithm was stable and reliable to achieve initial seeds also; it had an adequate approach to explore the number of clusters. The PFHC algorithm effectively categorized a set of malware profiles into their family groups. Moreover, PFHC out-performed other clustering algorithms such as hierarchical clustering and k-means clustering algorithms.

5. 6. Image Retrieval [21]

The traditional content-based image retrieval (CBIR) is a computer vision technique that searches for digital images in large databases based on analysing the contents of the image. However, this retrieval technique has been unable to meet efficiency for large and high-dimension image databases. Many researches have been conducted to extract data and potential information from general collections of images [22] [23]. This research introduced a digital image retrieval approach that utilizes the hierarchical clustering algorithm for hierarchical indexing to an image database. The proposed approach takes advantage of the agglomerative hierarchical, k-means and ART2 clustering algorithms. As the agglomerative hierarchical clustering algorithm consumes more that 90% of its total time at the initial iteration [24], a preprocessing step that reduces this consumed time is essential. For this purpose, ART2 clustering algorithm is used firstly to obtain the initial clustering results. After that, the hierarchical indexing is established by applying the hierarchical agglomerative clustering algorithm. Finally, the k-means clustering algorithm is used to calculate the pattern center to restrain the centroids drift which is a disadvantage of ART2.

The simulation results of the proposed image retrieval approach showed better results in computational time, efficiency and clustering results compared to the traditional CBIR approach.

5. 7. Land Cover Mapping Using Satellite Images [25]

The adaption of satellite images can lead to accurate planning and usage of lands. Satellite images offer to extract temporal data that can be beneficial to gain knowledge related to land use. The adaption of data analysing techniques has established a vast research area in presenting solutions for the land cover mapping problem for city planning and land-usage.

This research utilized various hierarchical clustering algorithms for the land cover mapping problem. The three hierarchical algorithms used in this research differ in their splitting methods. The splitting methods are used to search

Table 2: Summary of the aforementioned studies that hierarchical clustering had a significant role

Study	Role of Hierarchical Clustering	Comments
Gene Clustering	Group genes that have similar patterns of biological functions or interactions	- Hierarchical clustering solutions based on the iterative mergences between pair-wise distances instead of a global criterion - Lack of robustness and cut procedures to find the number of clusters in hierarchical clustering - Beneficial to visualize global patterns that express genes data - Not suitable to present cluster information for further biological exploration
Microarray Data Analysis on Ovarian Cancer	Analyze ovarian cancer microarray data and distinguish between the OVT, OVTT, OVCAI and OVCAIII	The average hierarchical clustering algorithm with Euclidean distance presented the best performance in distinguishing between the OVT, OVTT, OVCAI and OVCAIII
Classifying Electricity Customers	Identifying consumption patterns of customers and grouping them together based on their load diagram in order to formulate tariff offers	- The modified follow-the-leader and hierarchical clustering based on the average distance linkage criterion algorithms presented the best clustering results - Both algorithms provided remarkable detailed separation between clusters - The modified follow-the-leader was the most efficient algorithm as it comprised clustering adequacy and computational speed.
Protein Sequence Clustering	Hierarchical clustering based on the single-linkage criterion to pre-cluster the protein sequences in the first phase of the process	Robustness and effectiveness of the proposed model in clustering protein sequences and accordingly, understanding its function and structure
Malware Categorization	Clustering malware into different categories	- The hybrid (hierarchical & k-means) PFHC algorithm effectively categorized a set of malware profiles into their family groups - PFHC out-performed other clustering algorithms such as hierarchical clustering and k-means clustering algorithms
Image Retrieval	Hierarchical indexing of digital images is established by applying agglomerative hierarchical clustering	The proposed image retrieval approach showed better results in computational time, efficiency and clustering results compared to the traditional CBIR approach
Land Cover Mapping Using Satellite Images	Three hierarchical algorithms with different splitting methods; Mean Shift Clustering (MSC), Niche Particle Swarm Optimization (NPSO) and Glowworm Swarm Optimization (GSO) were used to present solutions for the land cover mapping problem for city planning and land-usage	It was observed that the hierarchical clustering based on GSO splitting method was the most accurate and robust algorithm for land cover mapping purposes

for the best possible number of clusters and its centroids, these splitting methods are Mean Shift Clustering (MSC), Niche Particle Swarm Optimization (NPSO) and Glowworm Swarm Optimization (GSO). The performance comparison of the proposed hierarchical clustering algorithms is presented using two typical multi-spectral satellite images Landsat and QuickBird. Based on the results, it was observed that the hierarchical clustering based on GSO splitting method was the most accurate and robust algorithm.

6. Conclusions

This paper presented an overview of the hierarchical clustering method and its utilizations in various research disciplines. Hierarchical algorithms can either be agglomerative or divisive. Agglomerative algorithms start with considering each data point as an individual cluster, and merge pairs of clusters at successive steps. There are various agglomerative approaches that have different distance definitions between clusters. The most common agglomerative procedures are single linkage, complete linkage and average linkage. Divisive hierarchical clustering begins with the entire data set in a single cluster, and then the most dissimilar clusters are split-off recursively into two clusters. It continues splitting until each cluster represents its own. Divisive hierarchical clustering has not been used as widely as agglomerative hierarchical clustering in the clustering literature.

There are various distance metrics that can express the (dis)similarity between pairs of data points such as, the Euclidean distance, Manhattan distance, Chebyshev distance and Mahalanobis distance. However, the most commonly used distance metric is the Euclidean distance. The choice of the distance metric is not critical in improving

the underlying structure of clusters, whereas, the choice of the clustering algorithm is much more significant.
Hierarchical clustering is considered to be one of the efficient and widely used methods in various research disciplines. It has been powerful in clustering genes and samples and is considered to be a standard tool for such purposes. Also, hierarchical clustering can be employed to analyze the ovarian cancer microarray data and detect cancers and tumors at early stages. In biomedical research, the hierarchical clustering algorithms are involved to understand the functions and structures of unknown protein sequences; in addition, it represented robustness and effectiveness into the procedure. Another utilization of hierarchical clustering is to identify the consumption patterns of customers and group them together based on their load diagrams in order to formulate tariff offers from electricity service providers to their customers. In addition, hierarchical clustering algorithms are used for computer security purposes. This involves hierarchical algorithms in procedures to categorize malwares based on their properties into groups. In content-based image retrieval, hierarchical clustering showed better results in computational time, efficiency and clustering results compared to other traditional CBIR approaches. Moreover, the adaption of various hierarchical clustering algorithms has established a vast research area in presenting solutions for the land cover mapping problem for city planning and land-usage.

References

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, ser. Prentice-Hall Advanced References series. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[2] A. Vr. F. Edwards and, L.L. CavaIli-Sforza, *A method for Cluster Analysis*, *Biometrics,* 1965, pp. 362-375.

[3] C. Peebles, *Monothetic-Divisive Analysis of Moundville Burials*. Newsletter of Computer Archaeology, 1972.

[4] H. T. Clifford and W. Stephenson, *An introduction to numerical classification.* New York: Academic Press, 1975.

[5] P. H. A. Sneath and R. R. Sokal, *Numerical taxonomy,* San Francisco: W. H. Freeman, 1973.

[6] J. W. Beckstead, "Using Hierarchical Cluster Analysis in Nursing Research," *Western Journal of Nursing Research*, vol. 24, no. 3, pp. 307-319, 2002.

[7] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The Elements of Statistical Learning: Data Mining: Inference and Prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83-85, 2005.

[8] E. A. Mooi, and M. Sarstedt, *A Concise Guide to Market Research,* Springer-Verlag Heudelberg, 2011.

[9] X. Rui and D. Wunsch, "Survey of clustering algorithms," *Neural Networks, IEEE Transactions*, vol. 16, no. 3, pp. 645-678, May 2005.

[10] A. A. Munshi, and Y. A.-R. I. Mohamed, "Photovoltaic power pattern clustering based on conventional and swarm clustering methods," *Solar Energy*, vol. 124, pp. 39-56, 2016.

[11] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Hoboken, NJ, USA, 2005.

[12] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques,* Morgan Kaufmann, San Francisco, 2005.

[13] N. Rajalingam and K. Ranjini, "Hierarchical Clustering Algorithm - A Comparative Study," *International Journal of Computer Applications*, vol. 19, no. 3, April 2011.

[14] T. G. Nikolaou, D. S. Kolokotsa, G. S. Stavrakakis and I. D. Skias, "On the Application of Clustering Techniques for Office Buildings' Energy and Thermal Comfort Classification," *Smart Grid, IEEE Transactions on* , vol. 3, no. 4, pp. 2196-2210, Dec. 2012.

[15] G. C. Tseng, "A comparative review of gene clustering in expression profile," *Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th*, vol. 2, pp. 1320-1324 Dec. 2004

[16] M-H Tsai, C-H Lai, S-Jr Lu and S-F Su, "Performance Comparisons between Unsupervised Clustering Techniques for Microarray Data Analysis on Ovarian Cancer," *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference*, vol. 5, pp. 3685-3690, Oct. 2006.

[17] G. Gatta and M. B. Lasota, et al., "Survival of European women with gynaecological tumours, during the period 1978-1989," Eur J Cancr, vol. 34, no. 14, pp. 2218-2225, 1998.

[18] G. Chicco, R. Napoli and F. Piglione, "Application of clustering algorithms and Self Organising Maps to Classify Electricity Customers", *Proc. IEEE Bologna PowerTech*, June 2003.

[19] W-B. Chen, C. Zhang and H. Zhong, "An unsupervised protein sequences clustering algorithm using functional domain information," *Information Reuse and Integration, 2008. IRI 2008*, pp. 76-81, July 2008

[20] Z-X. Han, S. Feng, Y. Ye and Q. Jiang, "A Parameter-Free Hybrid Clustering Algorithm Used for Malware Categorization," *Anti-counterfeiting, Security, and Identification in Communication, 2009. ASID 2009. 3rd International Conference*, pp. 480-483, Aug. 2009.

[21] C-Y. Zhao, B-X. Shi, M-X. Zhang and Z-W. Shang, "Image retrieval based on improved hierarchical clustering algorithm," *Wavelet Analysis and Pattern Recognition (ICWAPR), 2010 International Conference*, pp.154-157, July 2010.

[22] D. Singh and A. Singh, "A New Framework for Texture Based Image Content with Comparative Analysis of Clustering Techniques," *Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference*, pp. 232-236, Nov. 2012.

[23] L. Meng and A-H. Tan, "Semi-supervised hierarchical clustering for personalized web image organization," *Neural Networks (IJCNN), The 2012 International Joint Conference*, pp. 1-8, June 2012.

[24] LI Zhao-peng and LI Ken-l, "Parallel data preprocessing based on hierarchical clustering," *Microelectronics & Computer*, vol. 24, no. 10, 2007

[25] J. Senthilnath, S. N. Omkar, V. Mani, N. Tejovanth, P. G. Diwakar and B. A. Shenoy, "Hierarchical Clustering Algorithm for Land Cover Mapping Using Satellite Images," *Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 3, pp. 762-768, June 2012.