Clustering Techniques and Their Applications: A Review

Arjun Dutta
Member of CSIR(2018), India.
Email: arjuncode47@gmail.com

Abstract*— *This paper deals with concise study on clustering: existing methods and developments made at various times. Clustering is defined as an unsupervised learning where the *targets* are sorted out on the foundation of some similarity inherent among them. In the recent times, we dispense with large masses of data including images, video, social text,DNA, gene information, etc. Data clustering analysis has come out as an efficient technique to accurately achieve the task of categorizing information into sensible groups. Clustering has a deep association with researches in several scientific fields. *k-means algorithm was suggested in 1957.*K-mean is the most popular partitional clustering method till date. In many commercial and non-commercial fields, clustering techniques are used.The applications of clustering in some areas like image segmentation, object and role recognition and data mining are highlighted. In this paper, we have presented a brief description of the surviving types of clustering approaches followed by a survey of the areas.

Index Terms*— *Data Clustering, DNA, K-means, Gene data.

I. INTRODUCTION

We ARE living in a world full of data. Everyday, people encounter a large measure of information and store or represent it as data, for further analysis and management. Ace of the vital means of sharing with these data is to classify or group them into a set of classes or clusters. In reality, as one of the most primitive activities of human organisms.[1]
Since last two years 90% of the information in the world has been piled up. Data output in 2017 alone has been approximately 2.5 quintillion bytes per day. Cluster analysis is a key that is employed to discover the characteristics of the cluster and to concentrate on a particular cluster for further analysis. Clustering is unsupervised learning and do not rely on predefined categories. In clustering, we measure the dissimilarity between objects by measuring the space between each couple of targets. These measures includes the Euclidean and Manhattan distance.
Basically, classification systems are either supervised or unsupervised, depending on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively([3], [4], [5])

The goal of clustering is to divide a finite unlabeled data set into a finite and distinct set of "natural," hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability distribution ([6],[7]). This can get to the task of clustering fall outside of the framework of unsupervised predictive learning problems, such as vector quantization [7],,and entropy maximization [8].

Learning problems can be segmented into two distinct categories:
• Supervised: where all training data is labelled
Supervised Learning, Supervised learning was based on training a data sample from a data source with the correct classification already assigned. Such techniques are utilized in feed-forward or Multi Layer Perceptron (MLP) models. These MLP has three distinctive features:
1. One or more layers of hidden neurons that are not part of the input or output layers of the network that enable the web to learn and work out any complex problems
2. The nonlinearity reflected in the neuronal activity is differentiable and,
3. The interconnection model of the network presents a high level of connectivity

• Unsupervised: where input data is unlabelled Clustering, a prominent part of unsupervised learning is a more difficult and challenging problem than classification[9].

Self-organizing neural networks learn using an unsupervised learning algorithm to identify hidden patterns in unlabeled input data. This un supervised refers to the ability to discover and organize information without providing an error signal to measure the possible resolution. The lack of counselling for the learning algorithm in unsupervised learning can sometimes be advantageous, since it allows the algorithm to see back for patterns that have not been previously considered [10].
.
The main principal characteristics of Self-Organizing Maps (SOM) are[11]:

1. It transforms an incoming signal pattern of arbitrary dimension into one or two dimensional map and execute this transformation adaptively.
2. The network represents a feed-forward structure with a single computational layer consisting of nerve cells arranged in rows and columns
3. At each point of representation, each input signal is held in its proper context and,
4. Neurons dealing with closely related pieces of information are close together and they communicate through synaptic connections.

The primary aim of a clustering algorithm is to acquire a technique that will distinguish the natural groupings in the unlabeled data. Data clustering has been employed for the following principal purposes: • To gain useful knowledge from data,i.e. generates hypotheses, detect anomalies, and identify salient features within given data • To identify the level of similarity among forms, organisms or points that comprise the data • As a method for organizing the data and summarizing it through cluster prototypes.

In this paper, we constitute a survey of the diverse areas where data clustering has been successfully employed to harness new, relevant and useful information that has been utilized to achieve organized results within the said areas. Part 2 focuses on categorization of algorithms, highlight their key details and provide a few examples. Part 3 is comprised of various domains like banking, health urban development, privacy protection etc., where data clustering analysis has been carried out with success. Finally, section 4 details the points presented in the paper in the form of conclusion. The goal of this theme is to survey the core concepts and techniques in the large subset of cluster analysis with its roots in statistics and decision theory. Where appropriate, citations will be made to key concepts and techniques arising from clustering methodology in the machine-learning and other communities

II.TYPES OF CLUSTERING ALGORITHMS :-

The typical pattern clustering activity involves the following steps [Jain and Dubs 1988]:
(1) Pattern representation,
(2) definition of the pattern proximity measure appropriate to the data field,
(3) Clustering or grouping,
(4) Data abstraction, and
(5) assessment of output .

All these can be grouped under 8 distinct subsets :partitioning algorithm, hierarchical algorithm, density based algorithm, grid based algorithm, model based algorithm soft computing and bi-clustering. Here we are presenting 8 important clustering algorithms. Fig. 2 caves in a clean delineation of the terminology of clustering algorithms.

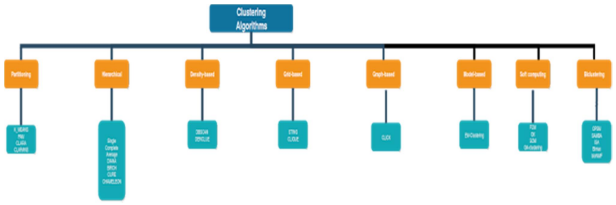


fig 2:Different Clustering Algorithms

Hierarchical Clustering Algorithms:

A representative algorithm of this kind is hierarchical clustering, which is implemented in the popular numerical software MATLAB [13]. This algorithm is an agglomerative algorithm that takes in variations depending on the metric used to measure the spaces among the clusters. The Euclidean distance is usually used for individual points . There are no known criteria, of which clustering distance should be used, and it seems to depend strongly on the dataset.

Among the most used variations of the hierarchical clustering based on different distance measures are [14]:

1. Average linkage clustering The dissimilarity between clusters is estimated using median values.
The mean length is computed from the distance between each peak in a cluster and all other points in another cluster. The two cluster with the lowest average distance are connected together to make the new cluster.

2. Centroid linkage clustering This variation uses the group centroid as the norm. The centroid is defined as the core of a cloud of points.

3. Complete linkage clustering (Maximum or Furthest-Neighbor Method)The dissimilarity between two groups is equal to the greatest dissimilarity between a member of cluster i and a member of cluster j. This method tends to get very tight clusters of similar cases.

4. Single linkage clustering (Minimum or Nearest-Neighbor Method): The dissimilarity between two clusters is the minimum dissimilarity between members of the two clusters. This method produces long chains which form loose, straggly clusters.

5. Ward's Method: Cluster membership is assigned by calculating the total sum of squared deviations from the mean of a cluster. The criterion for fusion is that it should produce the smallest possible increase in the error sum of squares.

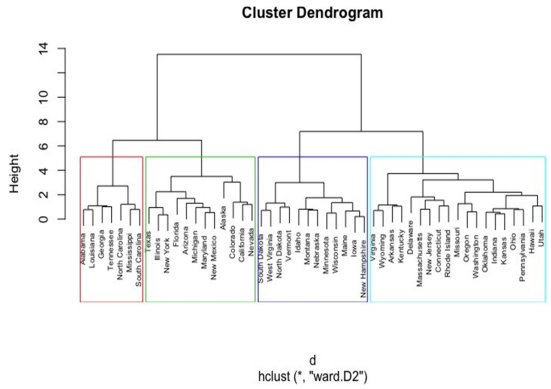


Fig 3([15])

K-means algorithm : The K-means algorithm, probably the first ace of the clustering algorithms proposed, is founded on a very simple idea: Given a set of initial clusters, assign each point to one of them, and so each cluster center is replaced by the mean point on the respective cluster [16]. These two simple steps are iterated until convergence. A period is assigned to the cluster, which is close in Euclidean distance to the stage. Although K-means has the outstanding advantage of being easy to implement, it delivers two big drawbacks [17].

Firstly, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success[18]

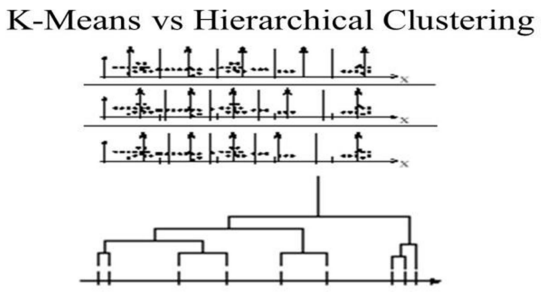


Fig 4[19]

Partitional Clustering: This is an iterative approach which finds similarity among intra-cluster points with regards to their distances from the cluster-centroid. It brings into consideration two assumptions [20].

A partitional clustering algorithm obtains a single division of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical technique. *Partitional* methods have advantages in applications, involves large data sets for which the construction of a dendrogram is computationally prohibitive[21].

Examples: Examples: K-Modes, K-Means, K-Medoids, etc.

Density based Clustering:

Density-based clustering is highlighted by number of applications. Substancial work has been executed in this field of Density based clustering. One approach has been built-up the incremental clustering for mining large database. This approach presents the first incremental clustering algorithm based on DBSCAN which is applicable on any database containing data in a measured distance. Referable to the density-based nature of DBSCAN, the introduction or omission of an object affects the current clustering only in the adjacent of this object [22].

Density based clustering includes three techniques: - DBSCAN [23] which grows clusters with respect to a density-based connectivity analysis. - OPTICS [24] extends DBSCAN to produce a cluster ordering obtained from a spacious scope of parameter settings. - DENCLUE [25] clusters objects based on a set of density distribution functions.

[26]DBSCAN forms more clusters,as compared to OPTICS also, , it was remarked that the time taken by Euclidian is always more as compared to Manhattan distance,but the number of clusters formed by using Euclidian as a length criterion is more as compared to using Manhattan

III. SOFT-COMPUTING TECHNIQUES USED IN CLUSTERING:

Soft computing is a promising problem solving technology which is suitable for unpredictable and nonlinear problems [27]. Hither,the basic techniques for solving clustering are, Genetic Algorithm, FLS(Fuzzy logic system), Neural Network etc. The fuzzy logic system (FLS) is an inference scheme which makes up the human thoughts and its basic form consists of a fuzzifier, some fuzzy IF-THEN rules, a fuzzy inference engine and a de-fuzzifier. NN-Neural networks (NNs) imitate the human mind to achieve intelligent tasks[28]. GA-Genetic algorithms (GAs) are numerical optimization algorithms inspired from genetics and have been given to a broad scope of troubles. GA typically maintains a population of persons that symbolize a set of candidate key for the considered problem.

Rough clustering extends the theory of rough or approximation sets. Rough k-means is first introduced by Lingras [29]. Referable to the ability of handling impreciseness, uncertainty, and vagueness for real-world problems, soft clustering is more realistic than hard clustering. Soft clustering as a partitioning algorithm is good for big data due to the heterogeneous structure of real large data [30].

Grid-based Clustering : The grid-based algorithms do not employ the database at one time. It establishes a uniform grid by collecting the data from the database by using statistical methods. The performance of the algorithm depends on the size of the grid and not on the

size of the actual data space. After running the grid, it computes the density of each cell in the gridiron. If the density of the cell is below the threshold value, the cell is thrown away. Ultimately, the clusters are formed from adjacent groups of dense cells. In the Grid-Based method, it quantizes the object space into a finite number of cells that form a grid structure. The primary advantage of this advance is its quick processing time, which is independent of the number of data objects and dependent on the number of cells in each proportion in the quantized space

Examples: STING, CLIQUE, Wavecluster, OptiGrid, etc.

Graph based clustering:

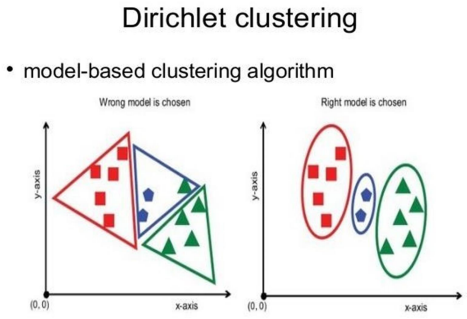
A graph-based clustering method was proposed to cluster protein sequences into families using a weighted linkage graph based on sequence similarity. It improves the single linkage clusters automatically using a graph partitioning algorithm, which applies a heuristic of the FM algorithm. The effectiveness of our method was shown by comparison with InterPro families in all mouse proteins in SWISS-PROT[32].

The number of protein sequences in public databases grows rapidly with the advancement of experimental technologies in molecular biology and large genome projects of late. The great size of sequence data makes it hard to recognize the relationships among a big lot of sequences. It is widely applied to cluster large data into meaningful or manageable groups [33, 34].

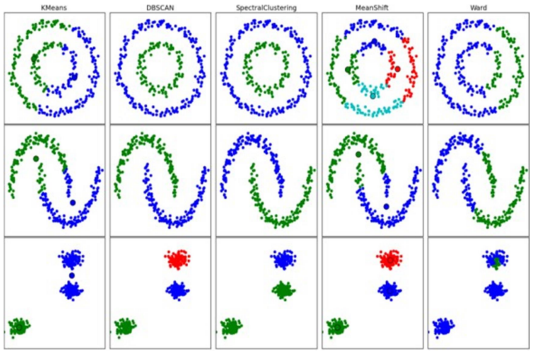
Model-based Clustering:

The picture indicates the two cases where k-means fails. Since the centers of the two clusters almost coincide, the k-means algorithm fails to distinguish the two clusters. This is imputable to the fact that k-means algorithm uses only a single parameter, i.e. the mean and hence can find only circular clusters. The model--based algorithms make use of many predefined statistical or mathematical models to make clumps. The turn of clusters may be pre-defined, though it is not necessary. This algorithm operates on the assortment of underlying probability and creates clusters on the base of it[35].

Examples:COBWEB, Neural Network .,etc



:



Biclustering classification of cluster algorithm:

The term bi-clustering has been used by Mirkin (1996) to describe "simultaneous clustering of both row and column sets in a data matrix"[37].

From visual inspection of the plots one can insure that this bi-clustering approach works equallyeasily as conventional clustering methods, when there were clear patterns over all attributes.Biclustering allows rows and columns to be included in multiple biclusters, and so allows one gene or one circumstance to be keyed out by more than one function categories[38]

APPLICATIONS:

Cluster analysis has been successfully implemented in various fields to extract useful patterns in data. In this section we explore the various fields where clustering technique has been used successfully.

Banking: Nowadays banking sectors collects a great quantity of information. Clustering analysis imparts a convenient answer to the threats. The biggest threat which is happening today is money laundering. To contain this menace, the DBSCAN clustering algorithm has been implemented in AMLRAS (Anti Money Laundering Regulations Application System). The function of DBSCAN algorithm is to observe and report suspicious financial transactions. AMLRAS has been successfully tried out on large financial data from where it could identify potential fraudulent transactions thus preventing laundering

Ogwueleka (2011) presented a Credit Card Fraud (CCF) detection model using Neural Network technique. The selforganizing map neural network (SOMNN) technique was applied to solve the problem of carrying out optimal classification of each transaction into its associated group since the output is not predetermined.

Maisarh, Yaseen; Omar in 2008 conducted a study to predict the failure of the Yemeni banks using logistic regression, discriminant analysis and Multiple Linear Regression.[41]The technique of fuzzy logic is one of the most important machine learning techniques used to detect financial failure in commercial banks[42].The combined effect of DBSCAN and Rule base data mining prediction algorithms on detection of card fraudulent transactions is presented.

The combined algorithms were demonstrated to be more effective in detecting or predicting card frauds than the single use of DBSCAN algorithm alone[43].

Healthcare : Clustering of medical X-ray images has been accomplished by various clustering techniques [44].

Medical image analysis also uses spectral clustering for a clear study ([45][46]).

Considering a small subset of medical datasets, algorithms have been formed and accurate results have been achieved by some of them [47]. Applied to a much larger generalized dataset, for every medical field, obtaining accurate results has yet been very difficult.

Liad[48] is an expert system program that uses Bayesian classification to estimate the posterior probabilities of various diagnoses under consideration, given the symptoms present in a case.

Urban development :

Quantifiable design criteria for urban design tasks include purely geometrical or topological measures, such as the length of roads or space accessibility[49], as well as social aspects, especially the perception of space, e.g. streetscape security[50].

Except the aforementioned applications of Clustering algorithms are spread through many other fields, like, Marketing (finding groups of customers with similar interests and behaviour given a large database of customer data containing their properties and past buying records), Medicine(IMRT segmentation, Analysis of antimicrobial activity), Medical imaging etc[51].

IV. CONCLUSION:

Wireless sensor network is a revolutionary invention in the domain of computer networking and automation. In various domains several efforts have been made by many researchers of the field. Clustering is one ,among the several methods of improving the efficiency of WSN. This paper provides a comprehensive review of the different clustering algorithms and their applications. Due to the large quantity of data maintained in several domains, clustering analysis has become indispensable in extracting patterns from bulk data thus aiding the process of useful knowledge discovery. Upon examining various clustering algorithms and their uses in different fields, we gather that out of several algorithms published till date, some are more popular than the others. Top clustering algorithms commonly used along with their fields of application have been highlighted in this paper. While some algorithms like the K-means and DBSCAN have been well explored and explained. A significant thing is that no single clustering algorithm has been found to dominate all areas of implementation. clustering technique has shifted from probabilistic model to soft computing or fuzzy logic model. Thus, we can say that ,there still exists much scope for research and development. With couple of few years we will see more extensive applications of the data clustering approach in diverse application scopes.

There are many elements that are usually known, and can be helpful in choosing an algorithm. One of the most important elements is the nature of the data and the nature of the desired cluster. Another issue to keep in mind is the kind of input and tools that the algorithm requires.

REFERENCES:

[1]M. Anderberg, Cluster Analysis for Applications. New York: Academic, 1973.

[2] Hale, T. (2018, June 18). How Much Data Does The World Generate Every Minute? Retrieved July 5, 2018, from http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

[3] C. Bishop, Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995.

[4] V. Cherkassky and F. Mulier, Learning From Data: Concepts, Theory, and Methods. New York: Wiley, 1998.

[5] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed. New York: Wiley, 2001.

[6] A. Baraldi and E. Alpaydin, "Constructive feedforward ART clustering networks—Part I and II," IEEE Trans. Neural Netw., vol. 13, no. 3, pp. 645–677, May 2002.

[7] V. Cherkassky and F. Mulier, Learning From Data: Concepts, Theory, and Methods. New York: Wiley, 1998.

[8] B. Fritzke. (1997) Some competitive learning methods. [Online]. Available: http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/JavaPaper

[9] Jain, A.K. (2008). Data Clustering: 50 Years Beyond K-means. ECML/PKDD

[10] T. Kohonen, O. Simula, "Engineering Applications of the SelfOrganizing Map", Proceeding of the IEEE, Vol. 84, No. 10, 1996, pp.1354 – 1384

[11] Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification R. Sathya Professor, Dept. of MCA, Jyoti Nivas College (Autonomous),Professor and Head, Dept. of Mathematics, Bangalore, India. Annamma Abraham Professor and Head, Dept. of Mathematics B.M.S.Institute of Technology, Bangalore, India.

[12] An Analysis on Clustering Algorithms in Data Mining Mythili S1 , Madhiya E2
IJCSMC, Vol. 3, Issue. 1, January 2014, pg.334 – 340

[13] G. Salton, "Automatic text processing: the transformation,"Analysis and Retrieval of Information by Computer, 1989.

[14] D. Isa, V. P. Kallimani, and L. H. Lee, "Using the self organizing map for clustering of text documents," Expert Systems With Applications, vol. 36,pp.9584-9591,2009

[15]Cluster Dendogram,https://www.google.co.in/search?q=cluster+dendrogram&source=lnms&tbm=isch&sa=X&ved=0ahUKEwje3IGO_IXdAhUlR48KHea5DkwQ_AUICigB&biw=1058& bih=734#imgrc=ZCW00wK8V2SELM:

[16] Shi Na,LiuXumin, "Research on k-means Clustering Algorithm",IEEE Third International Conference on Intelligent Information Technology and Security Informatics,2010.

[17] K. J. Cios, W. Pedrycz, and R. M. Swiniarsk, "Data miningmethods for knowledge discovery," IEEE Transactions on Neural Networks, vol. 9, pp. 1533-1534, 1998

[18]An Analysis on Clustering Algorithms in Data Mining Mythili S1 , Madhiya E2 Assistant Professor, PSGR Krishnammal College for Women, Coimbatore1 Assistant Professor, PSGR Krishnammal College for Women, Coimbatore2,IJCSMC, Vol. 3, Issue. 1, January 2014, pg.334 – 340

[19]https://www.google.co.in/search?q=hierarchical+clustering+and+k-means+clustering&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjUoK3W_4XdAhWFq48KHbOuCWYQ_AUICygC&biw=1058&bih=684#imgrc=5VRRLOnKULMMAM:

[20]Fahad, A., Alshatri, N.,Tari,Z., Alamri, A.,Khalil, I.,Zomaya, A.Y., Foufou,S.,& Bouras, A.(2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Transactions on Emerging Topics in Computing, 2, 267-279.

[21] Guha, Meyerson, A. Mishra, N. Motwani, and O. C. ."Clustering data streams: Theory and practice ." IEEETransactions on Knowledge and Data Engineering, vol. 15,pp. 515-528, 2003.

[22] Martin Ester,Hans-Peter Kriegel,Jorg Sander,Michael Wimmer,Xiaowei Xu,"Incremental clustering for mining in a data ware housing", University of Munich Oettingenstr. 67, D-80538 München, Germany.

[23] M. Ankerst, M. M. Breunig, H. P. Kriegel and J. Sander, OPTICS: Ordering Points To Identify Clustering Structure, at International Conference on Management of Data, Philadelphia, ACM 1999

[24] Yong-Feng Zhou, Qing-Bao Liu, S. Deng, Q. Yang, An Incremental Outlier Factor Based Clustering Algorithm, Proceedings of First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 Nov 2002

[25] Alexander Hinneburg, Hans-Henning Gabriel, DENCLUE 2.0: Fast Clustering based on kernel Density Estimation", Martin-Luther-University, Germany

[26]International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue- 1, March 2012,An Empirical Evaluation of Density- Based Clustering Techniques Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra

[27] M. Vinoth Kumar, Dr. T. Lalitha, Soft Computing: Fuzzy Logic Approach in Wireless Sensors Networks, Scientific Research Publishing, Circuits and Systems, 2016, 7, 1242-1249

[28] A REVIEW PAPER ON SOFT COMPUTING BASED CLUSTERING ALGORITHM Srutipragyan Swain1 , Manoj Kumar Das Mohapatra2 1Computer Science Department, Imit,Cuttack (India) 2 Independent Research (India)

[29] Lingras, Pawan, and Georg Peters. "Applying rough set concepts to clustering." Rough Sets: SelectedMethods and Applications in Management and Engineering. Springer London, 2012. 23-37.

[30] IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.1, January 2017 Soft Clustering for Very Large Data Sets Min Chen State University of New York, New Paltz, NY, USA

[31] Data Clustering: Theory, Algorithms, and Applications. (n.d.). Retrieved July 12, 2018, from https://epubs.siam.org/doi/abs/10.1137/1.9780898718348.ch12

[32] Genome Informatics 12: –102 (2001) 93 A Graph-Based Clustering Method for a Large Set of Sequences Using a Graph Partitioning Algorithm Hideya Kawaji1,2 Yosuke Yamaguchi,

[33] Kaufman, L. and Rousseeuw, P.J., Finding Groups in Data: an Introduction to Cluster Analysis, Wiley, 1990.

[34] Hartigan, J.A., Clustering Algorithms, Wiley, New York, 1975.

[35] Chamroukhi, F. (2013). Robust EM algorithm for model-based curve clustering. The 2013 International Joint Conference on Neural Networks (IJCNN), 1-8