Cluster Segments Identification Using H-K Means Algorithm

T. VENKATA SAI KRISHNA[1], Dr. YESU BABU ADIMULAM[2*] and Dr. R. KIRAN KUMAR[3]

[1]*Research Scholar, J N T UNIVERSITY-Kakinada, INDIA*

[*2]*Professor and HoD, Department of CSE, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, INDIA*

[3]*Assistant Professor, Department of Computer Science, KRISHNA UIVERSITY, Machillipatanam, Andhra Pradesh, INDIA*

Abstract - Clustering can be considered the most important un-supervised learning problem, so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. The determination of optimal number of clusters was carried out using elbow, silhouette and gap statistic does not reveal better clusters. NbClust, was used which has an exhaustive list of validity indices to estimate the number of clusters in a data set. Ward agglomeration and eucledian distance measures provided some meaningful insights on how many clusters are hidden in the data. The optimal number of clusters, k for the 5-HT receptor dataset was found to have 3 cluster solutions proposed by 7 indices. Therefore, initial value of k=3 was used to perform k-means, hierarchical followed by hybrid H-K means algorithm on the dataset. H-K means clustering method has modified the number of observations that appeared in cluster 3 segments.

Keywords - *NbClust, TDM, Clustering, H-K means algorithm.*

I. INTRODUCTION

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- Interpretability and usability.

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster, where the separation of points is achieved by a straight line on a bi-dimensional plane. On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering uses a completely probabilistic approach.

II. LITERATURE REVIEW

A. HIERARCHICAL CLUSTERING

Hierarchical cluster analysis of n objects is defined by a stepwise algorithm which merges two objects at each step, the two which have the least dissimilarity. Dissimilarities between clusters of objects can be defined in several ways; for example, the maximum dissimilarity (complete linkage), minimum dissimilarity (single linkage) or average dissimilarity (average linkage). Either rows or columns of a matrix can be clustered – in each case we choose the appropriate dissimilarity measure that we prefer. The results of a cluster analysis is a binary tree or dendrogram with n–1 nodes. The branches of this tree are cut at a level where there is a lot of 'space' to cut them that is where the jump in levels of two consecutive nodes is large. A permutation test is possible to validate the chosen number of clusters that is to see if there really is a non-random tendency for the objects to group together.

B. K-MEANS CLUSTERING

K-Means clustering is also an iterative clustering procedure, but it predefines the number of clusters that will be in the dataset. The algorithm begins by defining "centroids", which are points that will eventually migrate to the center of each cluster. The number of centroids chosen therefore determines the number of clusters in the dataset. The centroids are placed at random spots in the dataset. We then choose a distance metric to determine how far away each centroid is from each of the data objects. The distances of the objects that are closest to each of the centroids are then averaged, and the centroid is then moved to the center of the respective data objects. The process is then repeated by finding new distances from each of data objects to the centroids. The algorithm ends when the centroids no longer move within a certain threshold of distance. The closest data objects to each of the centroids are the resultant clusters. This process is similar to hierarchical in that it uses a distance metric to form clusters. It is different in that the number of clusters for k-means is predefined, where as hierarchical clustering creates levels of clusters.

C. HYBRID CLUSTERING HIERARCHICAL & K-MEANS

Hierarchical and k-means clustering are two major analytical tools for unsupervised datasets. However, both have their innate disadvantages. Hierarchical clustering cannot represent distinct clusters with similar expression patterns. Also, as clusters grow in size, the actual expression patterns become less relevant. K-means clustering requires a specified number of clusters in advance and chooses initial centroids randomly; in addition, it is sensitive to outliers.

Focus is on unsupervised clustering which is separated into two major categories: partition clustering and hierarchical clustering. There are many algorithms for partition clustering category, such as k-means clustering [1], k-medoid clustering, genetic k-means algorithm (GKA), Self-Organizing Map (SOM) and also graph-theoretical methods (CLICK, CAST). Among those methods, K-means clustering is the most popular one because of simple algorithm and fast execution speed. However, there are three major parts that require improvements: First, the number of k (clusters) must be decided before execution. Second, random choosing of the initial start points makes it impossible to obtain reliable results without much iteration of the entire clustering process. Third, it's sensitive to outliers. Although hierarchical clustering nests and represents the clusters as a dendrogram that provides an easy understanding of the data, the quality of clusters often degrades as more data are joined. It is becoming increasingly clear that none of the approaches alone are sufficient and that the application of various techniques will allow different aspects of the data to be explored

[2]. As a solution to this problem, a combined approach was proposed by Chen et al. (2005) [3], who first applied the k-means algorithm to determine the k clusters and then fed these clusters into the hierarchical clustering technique to shorten the merging cluster time and generate a treelike dendrogram. However, this solution still suffers from the limitation of determining the initial value for k [4] [5].

III. MATERIALS AND METHODS

In this paper, a new algorithm, hierarchical-k-means is proposed, which combines the advantages of both k-means and the hierarchical clustering algorithm to overcome their limitations. Initially the hierarchical clustering algorithm is applied and then the result used to decide the initial number of clusters and fed this information into k-means clustering to obtain the final clusters.

A. CLUSTER ANALYSIS

Cluster analysis [6] aims at classifying a set of observations into two or more mutually exclusive unknown groups based on combinations of variables. Thus, cluster analysis is usually presented in the context of unsupervised classification [7]. An important component of a clustering algorithm is the distance measure between data points. If all the components of the data instance vectors have the same physical units, it is then possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances.

B. DATASET

Serotonin (5-hydroxytryptamine (5-HT)) receptors has been associated in the aetiology of many disease conditions including depression, anxiety, social phobia, schizophrenia, obsessive-compulsive and panic disorders; migraine, hypertension, pulmonary hypertension, eating disorders, vomiting and irritable bowel syndrome [8]. The serotonergic system seems to be important in bulimia nervosa (BN). Modifications in brain serotonin function contributes to different aspects of eating disorders [9]. The data set used in this research contains physico-chemical properties [10] of 5-HT receptor drugs extracted from malacards database [11]. A total of 52 drugs which were tested clinically and being marketed worldwide are only selected. Properties such as Molecular Weight, logP, Heavy Atoms, H-bond Donors (HBD), H-bond Acceptors (HBA), polar surface area (PSA), number of freely rotatable bonds (RB) and half-life period of the drug created in the form of a table was subsequently used for analysis.

IV. RESULTS AND DISCUSSIONS

A. CLUSTER ANALYSIS

In this step, a hierarchical and k-means clustering followed by a hybrid hierarchical-k-means (HHK) algorithm was implemented. Before performing cluster analysis on a set of 5-HT receptor bound drugs were extracted from Malacards website and used further as dataset. Table 1 given below identifies the list of 52 drugs that are available in market towards reducing effect of 5-HT receptor activation.

In order to perform cluster analysis, few R packages were installed. They are facto extra, cluster and NbClust

From the figure 1, it was concluded that the optimal number of clusters, k for the 5-HT receptor dataset was found to have 3 cluster solutions. Therefore, initial value of k=3 was used to perform k-means, hierarchical followed by hybrid H-K means algorithm on the dataset.

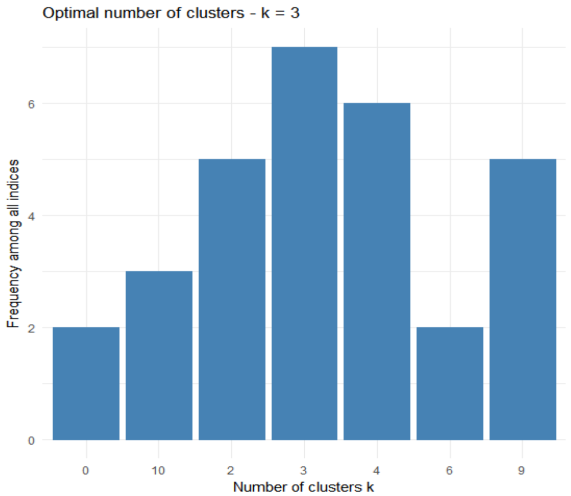

Fig.1: Consensus on optimal number of clusters obtained from NbClust package

Table 1: List of parameters selected for 5-HT receptor target.

row.names	Mwt	logP	Heavy_ Atoms	HBD	HBA	tPSA	RB	half_life
paroxetine	329.371	3.327	24	1	3	44	4	21
sertraline	306.236	5.18	20	1	0	16	2	24
citalopram	324.399	3.813	24	1	2	37	5	35
Clomipramine	314.86	4.528	22	1	1	7	4	32
Escitalopram	324.399	3.813	24	1	2	37	5	27
Fluoxetine	309.331	4.435	22	1	1	25	6	1
Fluvoxamine	318.339	3.202	22	1	3	58	9	15.6
Cocaine	303.358	1.868	22	1	4	57	3	0.5
Desipramine	266.388	3.533	20	1	1	19	4	7
duloxetine	297.423	4.631	21	1	2	25	6	12
imipramine	280.415	3.875	21	1	1	7	4	16
Methamphetamine	149.237	1.837	11	1	0	16	3	4
Methylphenidate	233.311	2.085	17	1	2	42	3	1
Milnacipran	246.354	1.771	18	1	1	47	5	6
Nortriptyline	263.384	3.826	20	1	0	16	3	16
Phentermine	149.237	1.966	11	1	0	27	2	7
Venlafaxine	277.408	3.036	20	2	2	33	5	5
Vilazodone	441.535	4.03	33	3	4	103	7	25.4
Amoxapine	313.788	3.429	22	1	3	41	0	8

Atomoxetine	255.361	3.725	19	1	1	25	6	5
Desvenlafaxine	263.381	2.733	19	3	2	44	4	10
Dexfenfluramine	231.261	3.246	16	1	0	16	4	32
Doxepin	279.383	3.962	21	1	1	13	3	6
Minaprine	298.39	2.196	22	1	5	50	5	2
Nefazodone	470.017	3.552	33	0	7	55	10	2
Protriptyline	263.384	4.302	20	1	0	16	4	6
Sibutramine	279.855	4.738	19	1	0	4	5	1.1
Tramadol	263.381	2.635	19	2	2	33	4	6.3
Trazodone	371.872	2.362	26	0	6	45	5	3
Trimipramine	294.442	4.121	22	1	1	7	4	11
amitriptyline	277.411	4.169	21	1	0	4	3	10
Mirtazapine	265.36	2.479	20	0	3	19	0	20
Mazindol	284.746	2.609	20	1	3	35	1	10
Pseudoephedrine	165.236	1.328	12	2	1	36	3	9
Vortioxetine	298.455	3.864	21	1	2	19	3	66
Dexmethylphenidate	233.311	2.085	17	1	2	42	3	2
Dextromethorphan	271.404	3.383	20	1	1	13	1	3
Mianserin	264.372	3.084	20	0	2	6	0	10
Amphetamine	135.21	1.576	10	1	0	27	2	10
Dopamine	153.181	0.599	11	3	2	68	2	0.02
Meperidine	247.338	2.213	18	1	2	30	3	3
verapamil	454.611	5.093	33	1	5	65	13	2.8
Loxapine	327.815	3.771	23	0	4	28	0	4
Olanzapine	312.442	1.746	22	1	5	30	0	21
Ondansetron	293.37	3.129	22	0	4	39	2	5.7
Quetiapine	383.517	2.856	27	1	6	48	5	6
Ribavirin	324.186	2.894	21	3	11	195	5	9.5
Phenelzine	136.198	0.692	10	2	2	38	3	1.2
Alitretinoin	300.442	5.603	22	0	2	40	5	2
Tegaserod	301.394	2.815	22	4	2	87	7	11
fenfluramine	231.261	3.246	16	1	0	16	4	20
Amineptine	337.463	4.499	25	1	2	56	8	0.48

B. HYBRID HIERARCHICAL-K MEANS CLUSTERING ALGORITHM

Clustering algorithms fragment a dataset into numerous groups or clusters which usually results in distributing the objects in some groups which represent a degree of similarity as possible and the objects in different groups show dissimilarity. From Figure 2, it is observed that the 3 cluster dendrogram was obtained from hierarchical clustering. It was found that two observations are with negative values and they probably placed in the wrong cluster. It is worth to mention that the average silhouette coefficient of complete dataset is similar (0.25) to that obtained from k means algorithm. The negative silhouettes were reported for 17 and 22 observations whereas in k means they are 7, 8, 24, 29, 50 and 52.

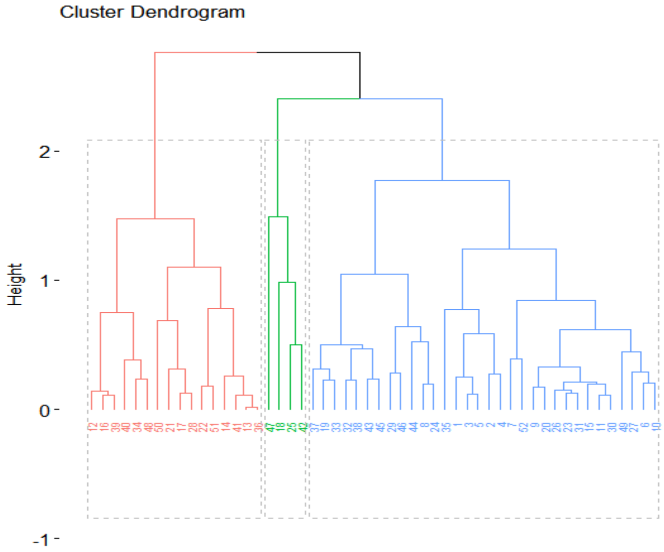

Fig.2: Hierarchical cluster dendrogram showing 3 clusters.

Table 2: 3 clusters formed by both kmeans and hierarchical clustering.

	Cluster 1	Cluster 2	Cluster 3	Kmeans cluster size
Cluster 1	27	3	0	**30**
Cluster 2	0	12	0	**12**
Cluster 3	5	1	4	**10**
Hierarchical cluster size	**32**	**16**	**4**	

Table 2 displayed rows being represented as 3 clusters of varying sizes by k means algorithm and column data represents cluster size by hierarchical clustering. It can be observed from cluster-1 that 3 of the observations belonging to cluster 1 by k means have been classified to cluster 2 in hierarchical clustering. The cluster 2 elements are 12 in both the cases, however, an additional 3 observations from cluster-1 and one observation from cluster-3 made the total observations to 16 in cluster-2 of hierarchical clustering. These differences are visualized in a dendrogram (Figure 3) and the final clustering data is also provided.

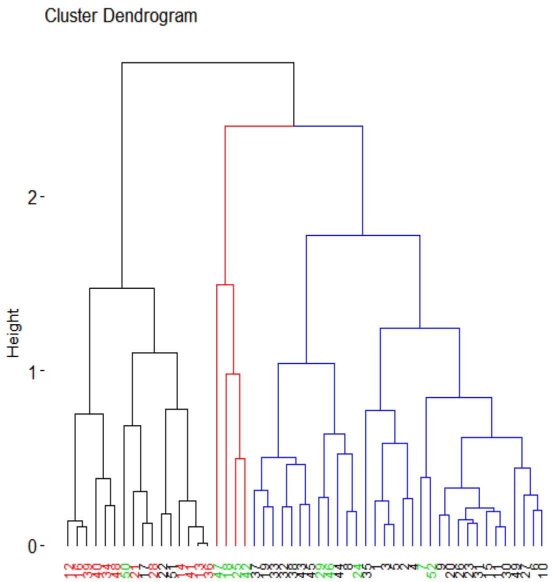

Fig.3: Modified hierarchical cluster dendrogram showing 3 clusters with marked variances.

Output of final H-K means clustering:

Table 3: 3 clusters formed by H-K means clustering.

	Cluster 1	Cluster 2	Cluster 3	HKmeans cluster
Cluster 1	26	4	0	**30**
Cluster 2	0	12	0	**12**
Cluster 3	2	2	6	**10**

Comparison of tables 2 and 3 revealed that the H-K means clustering method has modified the number of observations that appeared in cluster 3 segment.

V. CONCLUSION

The number of clusters carried out using elbow, silhouette and gap statistic does not reveal better clusters. Hence, NbClust, was used which has an exhaustive list of validity indices to estimate the number of clusters in a data set. Ward agglomeration and eucledian distance measures provided some meaningful insights on how many clusters are hidden in the data. The optimal number of clusters, k for the 5-HT receptor dataset was found to have 3 cluster solutions proposed by 7 indices. Therefore, initial value of k=3 was used to perform k-means, hierarchical followed by hybrid H-K means algorithm on the dataset.

VI. REFERENCES

[1]. MacQueen, J. Some methods for classification and analysis of multivariate observations. Proc: 5th Berkeley Symp. Math. Statist, Prob, 1:218-297, 1967.

[2]. http://www.lifesciencessociety.org/CSB2005/PDF2/043_chenb_hierarchical.pdf.

[3]. Chen, T.-S., Tsai, T.-H., Chen, Y.-T., Lin, C.-C., Chen, R.-C., Li, S.-Y., et al., 2005. A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray. In: Intelligent Signal Processing and Communication Systems, 2005. ISPACS 2005. Proceedings of 2005 International Symposium on, pp. 405–408.

[4]. Hasan, M.S., 2013. Investigating Gene Relationships in Microarray Expressions: Approaches Using Clustering Algorithms. The University of Akron.

[5]. Hasan, M.S., Duan, Z.-H., 2014. A hybrid clustering algorithms and functional study of gene expression in lung adenocarcinoma. In: Proceedings of the World Comp: International Conference on Bioinformatics and Computational Biology, pp. 23–29.

[6]. Jain AK, Murty MN, Flynn PJ: Data Clustering: A Review. ACM Computing Surveys (CSUR). 1999, 31 (3): 264-323.

[7]. Duda RO, Hart PE, Stork DG: Pattern Classification, ch.10: Unsupervised learning and clustering. Wiley, New York. 2001, 571.

[8]. https://pdfs.semanticscholar.org/bedf/1761ec0ab9d54634c353618447079aceb1f3.pdf.

[9]. Howard Steiger. Eating disorders and the serotonin connection: state, trait and developmental effects. J Psychiatry Neurosci. 2004 Jan; 29(1): 20–29.

[10]. https://pubchem.ncbi.nlm.nih.gov/.

[11]. http://www.malacards.org.