A Novel Approach to Clustering of Proteins

[1]R.Bhramaramba
Associate Professor,
Department of Computer Science and Engineering,
VNR Vignana Jyothi Institute of
Engineeering and Technology,
Hyderabad-500090, India

[2]Allam Appa Rao,
Vice-Chancellor, JNTU, Kakinada, India,

[3]Vakula Vijay Kumar
Dean, Godavari Institute of Engineering and Technology,
Rajahmundry, Hyderabad- India

***Abstract*— Data Analysis plays an indispensable role for understanding various phenomena. Clustering algorithms are a class of important tools for data analysis. K-means cluster analysis is considered to cluster protein variates across 3 species using SPSS 16.0. In this Paper we describe an approach to k-means cluster analysis which grouped the sample data of the three species under study into four apriori groups showed that the three groups are different from one another as evident from marginal inter group differences in the mean values of the clusters as well as the number of cases in each cluster. The analysis also shows that there are sharp intra group differences as reflected in the disproportionate number of cases in each cluster.**

Keywords- Clustering; Data analysis; K-means cluster analysis; protein variates

I. Introduction

Data analysis plays an indispensable role for understanding various phenomena [1]. Clustering algorithms are a class of important tools for data analysis[2]. Clustering is the process of grouping data into classes or cluster so that objects within the cluster are similar to the other. Cluster analysis is used to categorize genes with similar functionality [3]. It sorts out the relationship among proteins according to some objective criterion [4].The goal of clustering is to minimize the intracluster distances and to maximize the intercluster distances. It classifies classes eg. Respondents into clusters (groups) based on their variable values or characteristics. Cases within clusters are similar to each another and clusters are dissimilar from one another. A variable used to classify the set of cases is the cluster variate. This is specified by the researcher and not estimated empirically. Cluster analysis is a descriptive, non-inferential exploratory technique with roots in many areas, including data mining, statistics, biology and machine learning. This is highly dependent on the variables used. The objective of cluster analysis are to partition cases into groups based on similarities of characteristics to form a taxonomy (empirically based classification), to compare with a topology (theoretically based classification) to simplify data structure (clusters can be profiled by their general characteristics) and to reveal hidden relationships among cases. This analysis has been widely used in numerous applications viz. pattern recognition, data analysis and image processing. By clustering one can identify the correlations between data attributes [5]. With reference to the computational complexity of K-means cluster analysis, the time complexity is O(Nkd) and space complexity is O(N+k), the complexity becomes near linear to the number of samples in the data sets. Since k and d are usually much less than N, K-means can be used to cluster large data sets [1].

II. Related works

For purpose of clustering the protein attributes of the sample species under study, k-means cluster analysis method is used due to its known advantages over other clustering methods viz. k-means cluster analysis technique is considered efficient primarily because it does not compute the distances between all pairs of cases as do many clustering algorithms including that used by hierarchical clustering technique. Distances are computed using simple Euclidean distance. Its main advantage is it is much faster than hierarchical clustering technique.

According to J.B. Mac Queen [6], k-means is one of the simplest unsupervised learning algorithm that solves the well known clustering problem. The name comes from representing each of k clusters c by the mean or weighted average of c points, the so called centroid.

The k-means algorithm is a down clustering method. It is outlined as follows. Initial cluster centres are selected either randomly or through some other means. Observations are assigned to the closest centre to perform a partition of the data. Euclidean distance is the most common pairwise distance measure. The observations of each cluster are averaged to produce the new values for the center vector of that cluster [7].

The k-means cluster analysis procedure is simple. It classifies the data set through a certain number of clusters (assume k clusters) fixed apriori. This is followed by defining k centroids, one for each cluster. These centroids are placed at different locations as they produce different results. To the extent possible, these centroids are placed far away from one another. The next step is to associate each point in the data set to the nearest centroid. After completion of all parts in the data set, the first step comes to an end and early groupage is done. At this stage the user has to recalculate the k new centroids as barycenters of the clusters resulting form the previous step. After finding the new k centroids a new binding or loop will be done between the same data set point and the nearest new centroid. Through this loop the k centroids change their location step by step until no more changes are done i.e. centroids do not move any more. Finally this algorithm aims at minimizing an objective function, in this case a squared error function.

Although it can be proved that the procedure will always terminate, the k-means algorithm doesn’t necessarily find the most optimal configuration corresponding to the global objective function minimum. A popular heuristic for k-means clustering is Lloyd’s algorithm [8].

There is no general theoretical solution to find the optimal number of clusters for any given data set. Normally, the results of the multiple runs are compared with different k classes to choose the best one according to the criterion for instance Schwarz criterion. To conclude, k-means clustering is partitioning and relocation method primarily based on analysis of variance.

The data set and protein attributes considered for clustering are the same as outlined [9]. The study has taken 556 samples of Diabetes related proteins across each of three species and eight variables pertaining to their physicochemical properties. The Sample Size is taken after preprocessing 5000 sample out of 22,000 available due to lack of space. The data sample has been clustered by considering a data structure which follows object by variable structure. This represents n objects such as proteins with p variables such length, % basic, % acidic, % hydrophobic, %aromatic and % polar. The structure is in the form of a n by p matrix. The data had been subjected to logarithmic transformation for better accuracy. The number of clusters has been chosen to be four at random.

III. EXPERIMENTAL STUDY & ANALYSIS

The results are examined with reference to initial cluster centers (Table 1), Changes in cluster centres(Table 2), final cluster centres(Table 3), Distances between final cluster(Table 4), ANOVA(Table 5) and number of cases in each cluster(table 6).

Panel A: Homo Sapiens

	Cluster			
	1	2	3	4
lnlength	6.63	4.58	8.42	6.21
lnperbasic	3.44	2.85	2.23	1.89
lnperacidic	3.22	2.73	2.92	2.47
lnperhydroph	2.26	3.35	2.52	3.46
lnperaromatic	1.48	.71	2.28	2.65
lnperpolar	3.35	3.42	3.80	3.51

Panel B: Mus Musculus

	Cluster			
	1	2	3	4
lnlength	5.08	6.62	8.46	3.61
lnperbasic	1.93	3.45	2.62	3.07
lnperacidic	2.24	3.21	3.10	2.79
lnperhydroph	3.57	2.20	3.29	3.19
lnperaromatic	2.71	1.45	2.20	.99
lnperpolar	3.42	3.38	3.29	3.48

Panel C: Rattus Norvegicus

	Cluster			
	1	2	3	4
lnlength	4.58	5.89	8.46	5.97
lnperbasic	2.85	2.95	2.61	2.87
lnperacidic	2.66	2.71	3.11	3.16
lnperhydrophobic	3.35	2.63	3.28	2.03
lnperaromatic	.02	1.71	2.21	3.44
lnperpolar	3.52	1.71	3.29	3.74

Table 1. Initial Cluster Centers

Panel A: Homo Sapiens

	Change in Cluster Centers			
Iteration	1	2	3	4
1	1.117	1.241	.995	.946
2	.229	.215	.194	.088
3	.105	.063	.012	.050
4	.041	.019	.062	.011
5	.013	.022	.023	.017
6	.007	.010	.000	.009
7	.005	.008	.012	.006
8	.000	.006	.000	.004
9	.000	.007	.000	.005
10	.000	.003	.000	.002

a Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is .003. The current iteration is 10. The minimum distance between initial centers is 2.390.

Panel B: Mus Musculus

	Change in Cluster Centers			
Iteration	1	2	3	4
1	1.217	1.344	1.112	1.031
2	.046	.092	.084	.330
3	.025	.053	.050	.129
4	.014	.022	.022	.067
5	.011	.006	.012	.035
6	.011	.002	.000	.022
7	.008	.002	.000	.014
8	.007	.004	.000	.010
9	.013	.004	.000	.022
10	.011	.002	.000	.021

a Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is .016. The current iteration is 10. The minimum distance between initial centers is 2.417.

Panel C: Rattus Norvegicus

	Change in Cluster Centers			
Iteration	1	2	3	4
1	1.694	1.809	1.424	1.670
2	.283	.126	.085	.093
3	.169	.109	.014	.039
4	.079	.059	.007	.021
5	.047	.070	.028	.019
6	.021	.025	.000	.004
7	.013	.017	.005	.004
8	.012	.012	.000	.003
9	.013	.000	.004	.010
10	.017	.000	.000	.011

a Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is .013. The current iteration is 10. The minimum distance between initial centers is 2.772.

Table 2 - Iteration History(a)

Tables 1 and 2 i.e. initial cluster centres and changes in cluster centers show that the distance between the clusters across species and within the species vary widely in respect of all the five protein variates. This difference is more manifested between humans and the other two species which showed greater resemblance among them despite variations. Almost similar inferences emerge in case of final cluster centres. The distance between final cluster centres showed a clear descending order while in the other two species the distance between clusters increased uniformly in the case of mouse and rat.

Panel A: Homo Sapiens

	Cluster			
	1	2	3	4
lnlength	6.67	5.23	7.62	6.02
lnperbasic	2.58	2.60	2.56	2.54
lnperacidic	3.01	2.95	3.00	2.90
lnperhydroph	3.11	3.15	3.07	3.18
lnperaromatic	2.03	2.03	2.05	2.22
lnperpolar	3.48	3.46	3.52	3.46

Panel B: Mus musculus

	Cluster			
	1	2	3	4
lnlength	5.70	6.37	7.27	4.90
lnperbasic	2.54	2.58	2.57	2.70
lnperacidic	2.90	2.97	3.02	2.95
lnperhydroph	3.18	3.13	3.13	3.18
lnperaromatic	2.22	2.09	2.04	1.88
Lnperpolar	3.46	3.48	3.48	3.42

Panel C: Rattus Norvegicus

	Cluster			
	1	2	3	4
lnlength	5.17	6.05	7.03	6.09
lnperbasic	2.61	2.62	2.57	2.53
lnperacidic	2.95	2.99	3.00	2.92
lnperhydrophibic	3.19	3.05	3.12	3.19
lnperaromatic	2.08	1.76	2.05	2.33
lnperpolar	3.42	3.54	3.49	3.42

Table 3 - Final Cluster Centers

ANOVA table 5 considered as the central statistical output of k-means cluster analysis shows the size of mean square error and F values. It shows that F values magnitude of length, aromatic and acidic variables discriminate between different clusters in the human sample as well as mouse though the relative significance of the variables differ in each case. However, in the case of rat, length, aromatic and hydrophobic variables differentiate between various clusters in the sample.

Panel A: Homo Sapiens

Cluster	1	2	3	4
1		1.439	.949	.695
2	1.439		2.386	.810
3	.949	2.386		1.617
4	.695	.810	1.617	

Panel B: Mus Musculus

Cluster	1	2	3	4
1		.686	1.584	.889
2	.686		.903	1.499
3	1.584	.903		2.387
4	.889	1.499	2.387	

Panel C: Rattus Norvegicus

Cluster	1	2	3	4
1		.957	1.867	.957
2	.957		1.026	.608
3	1.867	1.026		.993
4	.957	.608	.993	

Table 4 - Distances between Final Cluster Centers

Panel A: Homo Sapiens

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Lnlength	89.357	3	.072	552	1243.800	.000
Lnperbasic	.095	3	.037	552	2.570	.054
Lnperacidic	.398	3	.032	552	12.470	.000
Lnperhydroph	.222	3	.031	552	7.114	.000
Lnperaromatic	1.580	3	.083	552	18.923	.000
Lnperpolar	.067	3	.024	552	2.836	.038

Panel B: Mus Musculus

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Lnlength	90.454	3	.099	552	911.966	.000
Lnperbasic	.464	3	.037	552	12.579	.000
Lnperacidic	.291	3	.037	551	7.788	.000
Lnperhydroph	.093	3	.031	552	3.015	.030
Lnperaromatic	2.109	3	.095	552	22.291	.000
Lnperpolar	.075	3	.021	552	3.628	.013

Panel C: Rattus Norvegicus

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Lnlength	77.557	3	.102	552	760.790	.000
Lnperbasic	.255	3	.035	552	7.335	.000
Lnperacidic	.228	3	.031	552	7.302	.000
lnperhydrophibic	.487	3	.031	552	15.624	.000
Lnperaromatic	7.041	3	.070	552	100.413	.000
Lnperpolar	.360	3	.028	552	12.880	.000

Table 5 – ANOVA

Panel A: Homo Sapiens

Cluster	1	157.000
	2	144.000
	3	47.000
	4	208.000
Valid		556.000
Missing		.000

Panel B: Mus Musculus

Cluster	1	181.000
	2	214.000
	3	84.000
	4	77.000
Valid		556.000
Missing		.000

Panel C: Rattus Norvegicus

Cluster	1	127.000
	2	89.000
	3	141.000
	4	199.000
Valid		556.000
Missing		.000

Table 6 - Number of Cases in each Cluster

Finally, a comparison of the number of cases across clusters in the three species (Table 6) also clearly showed that the three groups differ from one another. This is clearly manifested in the proportion of sample distribution across four clusters among the three species. The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

IV. CONCLUSION

To sum up k-means cluster analysis which grouped the sample data of the three species under study into four apriori groups showed that the three groups are different from one another as evident from marginal inter group differences in the mean values of the clusters as well as the number of cases in each cluster. The analysis also shows that there are sharp intra group differences as reflected in the disproportionate number of cases in each cluster.

References

[1] Rui Xu and Donald Wunsch II, Survey of Clustering Algorithms, IEEE Transactions on Neural Networks, vol. 16, No.3, May 2005, 645-678.

[2] Yi Hong and Sam Kwong, Learning Assignment Order of Instances for the constrained K-means clustering algorithm, IEEE Transactions ib systems, Man and cybernetics – Part B: Cybernetics, Vol. 39, No.2, April 2009, 568-574.

[3] Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers. 2006. 24-25, 123, 336.

[4] S.C.Rastogi, Namita Mendiratta and Parag Rastogi. Bioinformatics – Concepts, Skills and Applications. CBS Publishers & Distributors. 2006. 9, 19, 24, 30, 245, 285, 290, 306.

[5] Kristina Machova, Valentin Matak and Peter Bednar. The role of the Clustering in the Field of Information Retrieval. 2004

[6] J. B. MacQueen. "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press. 1967. 1:281-297.

[7] Kjersti Aas. Micro Array Data Mining – A Survey. Norwegian Computing Center. Oslo, Norway. 2001

[8] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu. An Efficient K-means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 24, No. 7. July 2002. 881.

[9] Bhramaramba R, Allam Appa Rao, Vakula Vijay Kumar and GR Sridhar. Comparative Study in Rodents and Human Beings: Principal Component Analysis (pca) of Proteins Related to Type 2 Diabetes Mellitus. Journal of Applied and Theoretical Information Technology. Feb 2009. Vol. 5. No.2. 143-179.

AUTHORS PROFILE

Ravi Bhramaramba received M.S. Software Systems degree from BITS Pilani in 1999. She received her Ph.D. degree in Computer Science & Engg from Jawaharlal Nehru Technological University, Hyderabad (JNTUH) in 2011. She has total 11 years of teaching experience, currently she is an Associate Professor in the Department of Computer Science & Engineering. She is a life member of ISTE and CSI. Her research interests include Databases, Data mining, and Bioinformatics.

Dr Allam Appa Rao, Vice-Chancellor of the JNTU Kakinada, is an iconic and towering personality in the field of education and research. His contributions to the field of Computer Engineering have been exemplary and spilled over into numerous other areas of science and technology, making him a pioneer of scientific advancements meant for the benefit of society. His areas of specialization include Bioinformatics and Computational Biology, Knowledge Management, Software Engineering and Network Security. He has shared his prowess with students, fellow-engineers and scientists across the globe by authoring more than 150 research papers published in international journals and conference proceedings.

Vakulabharanam Vijaya Kumar received integrated M.S. Engg, degree from Tashkent Polytechnic Institute (USSR) in 1989. He received his Ph.D. degree in Computer Science from Jawaharlal Nehru Technological University (JNTU) in 1998. He has served the JNT University for 13 years as Assistant Professor and Associate Professor and taught courses for M.Tech students. He has been Dean for Dept of CSE and IT at Godavari Institute of Engineering and Technology since April, 2007.His research interests include Image Processing, Pattern Recognition, Network Security, Steganography, Digital Watermarking, and Image retrieval. He is a life member for CSI, ISTE, IE, IRS, ACS and CS. He has published more than 120 research publications in various National, Inter National conferences, proceedings and Journals. He has established Srinivasa Ramanujan Research Forum at GIET, Rajahmundry, India for promoting research activities.