Utilizing Data Mining in Biology: Insights with MATLAB

Alina Ivanova [1] & Ivan Petrov [2]
[1 & 2] Department of Computer Science, Sofia University, Sofia, Bulgaria

ABSTRACT

FCM allows one piece of data to belong to two or more clusters. Results based on different clusters in both algorithms. K-means is the centroid based technique. We are also compared k-means and FCM results in this research. Comparison results show that the k-means is better than FCM. With the help of this research we can remove complexity from data sets in future. So the result shows that proteins are close to each other and k-means algorithm remove data set complexity with high accuracy and less consuming time and found large sum of distance in among the statistics peak's association to FCM algorithm. Data mining techniques is very important in the analysis of real environmental data.

Keywords: *K-Mean Algorithm, MATLAB, MEROPS, Unsupervised Learning*.

I. INTRODUCTION

Data mining is an information finding from data immoral. Data finding is the computer process of excavating finished examining of data and afterward that removing the sense of the data. Data quarrying implement accommodating for response business query that conventionally were too fewer period overwhelming.

We can usage statistics withdrawal instrument for additional clothes. They scrub data sets discovery prognostic and other data that specialists may failure, since it lays outdoor their prospects. Data mining trappings and approaches permitting commercial to kind active, forecast performances, economics, knowledge-driven conclusions, impending tendencies and bio-informatics.

Data mining is applied and main for businesses in a comprehensive variety of businesses counting trade, industrial transport; fitness upkeep, medicinal skill; economics and atmosphere are previously by data mining outfits and methods to income benefit of ancient data. With the assist of unspoiled thanks skills and arithmetical and scientific procedures to examine finished warehoused information, data mining analysts recognize significant facts, patterns, and relationship exceptions.

For business purpose data mining is used to relationships in the data in order to help make better business decisions and discover patterns. Data mining can forecast customer devotion, grow keener advertising movements and it is assistance to commercial sales trends.

Data mining investigator housing pronouncement plants (1960s). Data mining investigator also developed delivery vector machines (1990s). Data mining is obliging to smearing these approaches. These approaches are contact concealed designs in large data sets. Data mining bonds the hole from practical statistics, bio-informatics and reproduction intellect. Data mining delivers the accurate contextual to catalog organization by abusing. In Data mining data is stowed and indexed indatabases to tool the sure knowledge and detection procedures more capably. Data miningletting such approach to be useful to a big of data circles.

II. FCM ALGORITHM BY FUZZY LOGIC TOOLBOX AND K-MEANS PROCEDURE BYMATLAB ON PROTEIN DATA SET

In early periods numerous kinds of statistics failure job and numerous kindsof data in setting inclined to grip somewhat minor data sets. Same enormous amount ofdata has been calm from hereditary information corrective, forte upkeep and inenvirons. The database is enlarged even earlier and twitches additional thickness andproblems in data. Data Mining is consciousness discovery from Data.

Data mining has changed and endures to change from the divide of investigation arenas such as machine teaching, natural knowledge, healthcare, manufacturing, pattern appreciation databases, and digits and with the assistance of data mining remove difficulty in data. In this investigation we contend MEROPS connected instrument for protein

dataset. Protein orders those are earlier to each other. So with the contribution of datamining technique we at hand how can take absent this problematic and last the standby in dataset with the assistance of clustering. We at pointer this investigation MEROPS operational instrument for protein data set, FCM algorithm, Fuzzy logic toolbox, K-means algorithm, MATLAB, bioinformatics toolbox in MATLAB.

Gathering is the procedure of group a usual of data substances. In bunching data objects in to numerous collection or bunches so that substances with in a tall resemblance, nonetheless these objects are actual unlike to substances in additional group. Dissimilarities and similarities are measure stand on the power ideals unfolding the objects and often involve distance measures. We used two algorithms first is K-means and second is fuzzy C Means. K-Means is a Centroid-Based Method usages the centroid of clusters. Fuzzy c-means (FCM) is a technique of clustering and in this way stands one helping of statistics to go to binary or additional clusters. We usage bio informatics tool box in MATLAB for conspiracy the protein dataset. MATLAB (matrix laboratory)R2007b is an arithmetical calculating setting.

It is built-up by mathematical calculation. MATLAB permits project of customer borders, matrix operations, data process of algorithms and interfacing with programs in print in other languages. MATLAB is connecting C,C^{++}, Java, and FORTRAN. It is fourth-invention programming language. FuzzyLogic Toolbox is a collection of purposes constructed on the MATLAB®. It supplies mechanisms for you to make and oversee fuzzy belief governments inside the support of MATLAB. If you favor, you can joint composed your fuzzy schemes into imitations with Simulink. This toolbox trusts importantly on graphical user interface (GUI) gears to assistance can exertion completely from the custody streak. Chief opinion in this investigation is protein inspection with the assistance of data mining instrument. We use bioinformatics toolbox too.

III. RESULTS AND DISCUSSION

K-means algorithm

Table 1, Table2, Table 3 current the consequences for consecutively k-Means for secure value of n and how presentation limits are varying in each run. Here the value of n =2,3,4,5 and a protein data set of X= 436 opinions is used to discovery these limits. It also demonstrations that how different accuracies that are got throughout each run of algorithm.

Table 1: Determine the Performance of k-means algorithm in MATLAB (Comparison between No of clusters n=2, 3, 4, 5)

No of clusters	No of iterations	No of misclassified points	Accuracy
2	5	0	100%
3	5	5	97.85%
4	5	15	95.10%
5	5	10	94.47%

Table 2: Determine the Performance of k-means algorithm in MATLAB (Comparison between No of clusters n=3,4,5)

No of clusters	No of iterations	No of misclassified points	Accuracy
3	8	1	99%
4	8	6	97.62%
5	8	9	96.93%

Table 3: Regulate the Presentation of k-means algorithm in MATLAB (Comparison between No of clusters n=4,5)

No of clusters	No of iterations	No of misclassified points	Accuracy
4	12	1	98%
5	12	5	97.85%

Table 4 shows that silhouette value for k = 2 are better thus desired numbers of clusters are 2.

Table 4: Regulate the Presentation of k-means algorithm in MATLAB

(Silhouette value index, Better clusters according to silhouette value)

No of clusters	Silhouette value index	Better clusters according to silhouette value
2	0.7	2
3	0.65	2
4	0.52	1
5	0.41	2 and 4

Figure 1, Figure 2, Figure 3 and Figure 4 demonstration outline worth for algorithm. We can assess that silhouette plots demonstration that the right amount ofclusters. With the assistance of these figures we can control better cluster for data. We can see that in these plots silhouette value for cluster 2 is better for data set.

The results of close cram were in the overflowing agreement of preceding investigators (Singh and Mahajan, 2014), who also optional the gathering algorithm such as k-means algorithm and Fuzzy c means algorithm. K-means algorithm gives healthier consequence in this investigation.

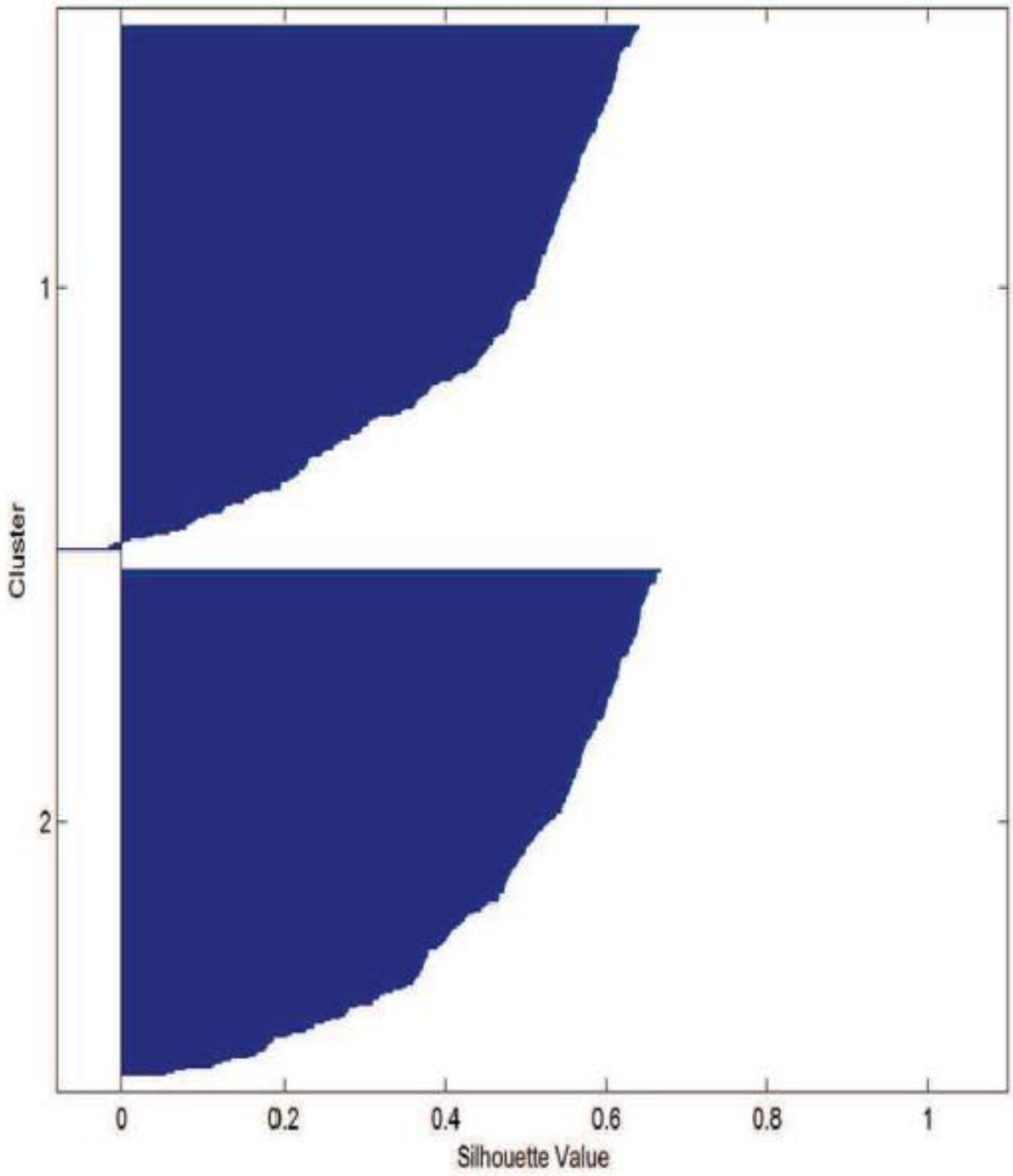

Figure 1: Silhouette values for clusters 2

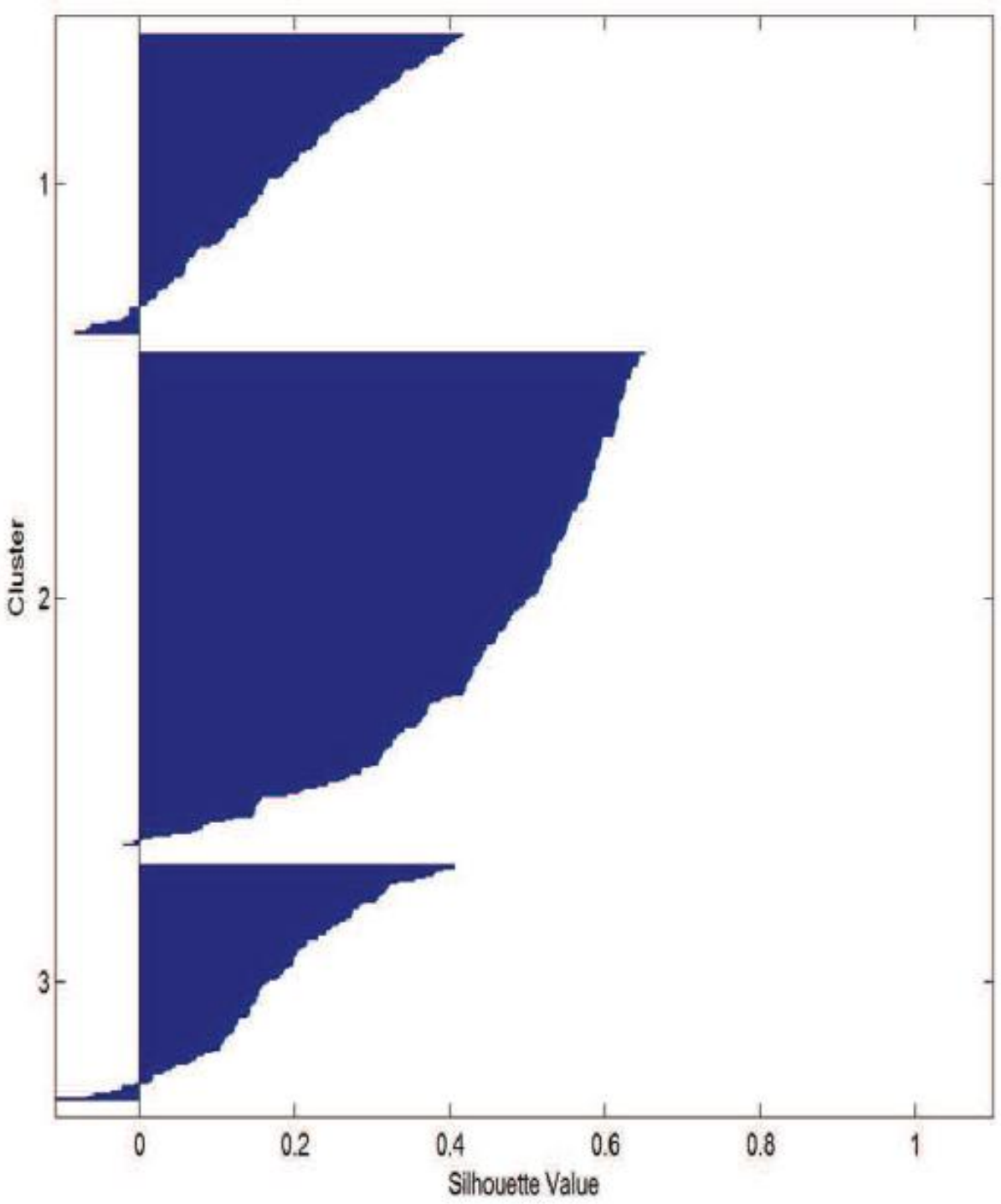

Figure 2: Silhouette values for clusters 3

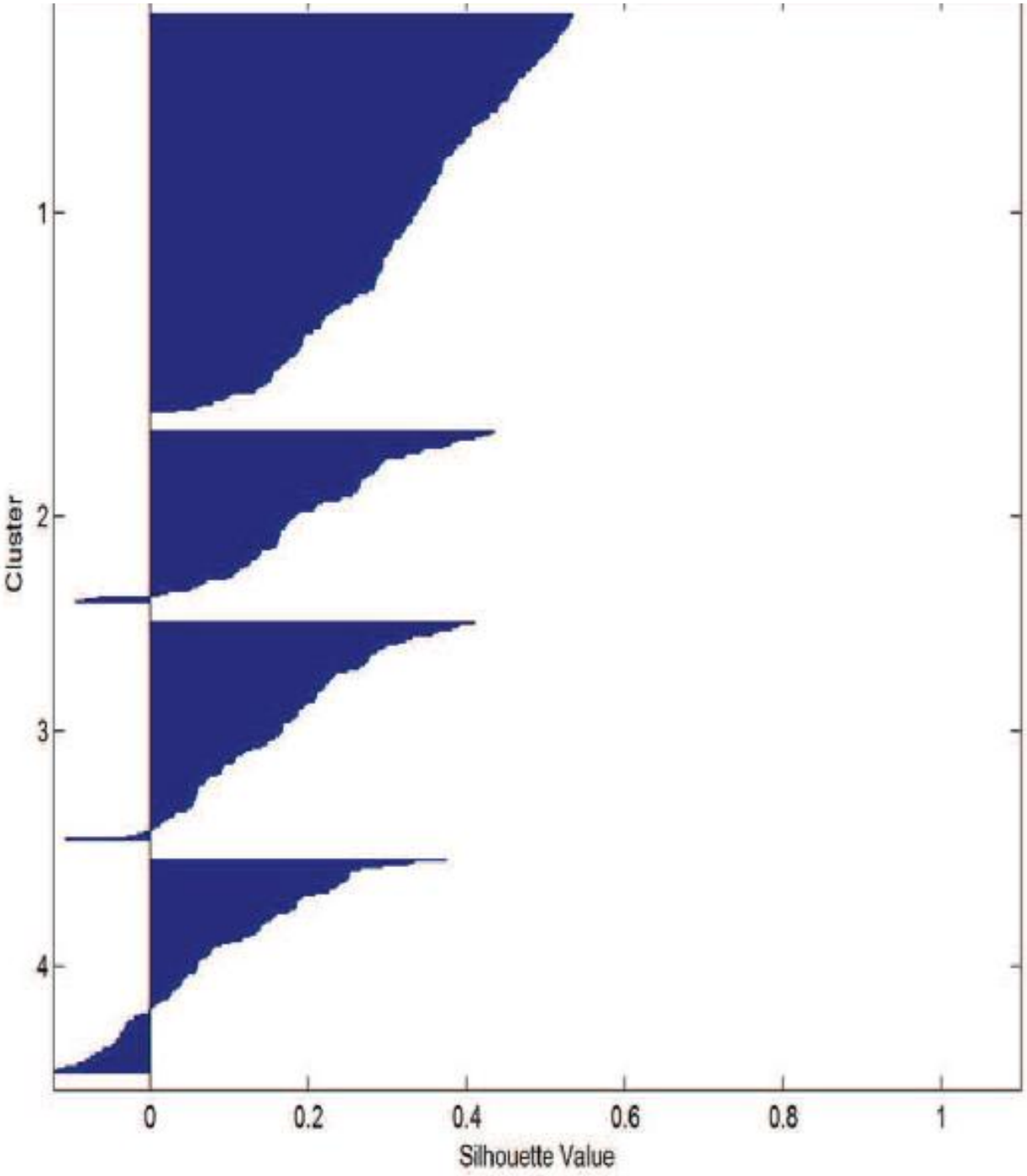

Figure 3: Silhouette values for clusters 4

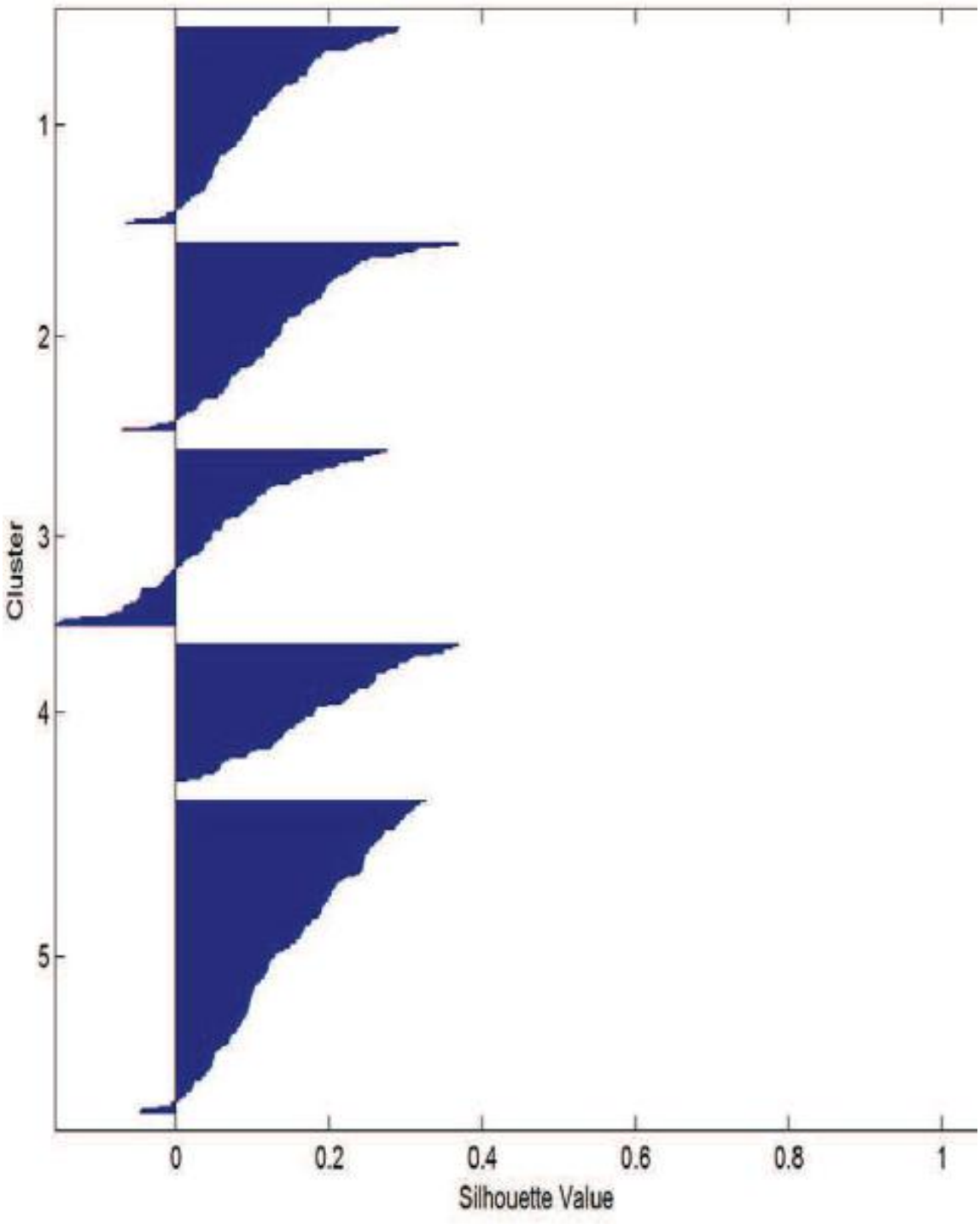

Figure 4: Silhouette values for clusters 5

The results discoursed in Table 4.5 presents k-means evade local minima. In this result kmeans up surges the sum of coldness. It's a robust point of k-means.

Table 5: Increases sum of distance and avoid local minima

No of clusters	Increases sum of distance

2	342.2992
3	282.5109
4	462..0820
5	450.5523

Table 6 shows result for k-means for accuracy according to the total sum of distance and also shows that how performance parameter varying in each cluster. We determine accuracy test between the clusters.

Table 6: Performances measure the different no clusters of k-means.

No of clusters	No of iterations	Total sum of distance	Accuracy
2	5	629.514	82.802571%
3	8	916.561	94.844868%
4	12	1210.48	94.781267%
5	22	1495.08	92.421245%

Table 7: Show that statistic value of data

Statistic	Value clusters 2	Value (clusters 3)	Value (clusters 4)	Value (clusters 5)
Mean	0.0345	-0.0476	-0.0013	0.0070
Mode	-3.6442	-4.0291	-4.0737	-4.073
Median	-0.0944 -	0.0452	0.0088	-0.0030
Std	1.3832 1	.3925	1.3813	1.3806

Figure 5 shows separated random data for two clusters. Proteins are close together to each other. With the help of k-means algorithm we remove this problem.

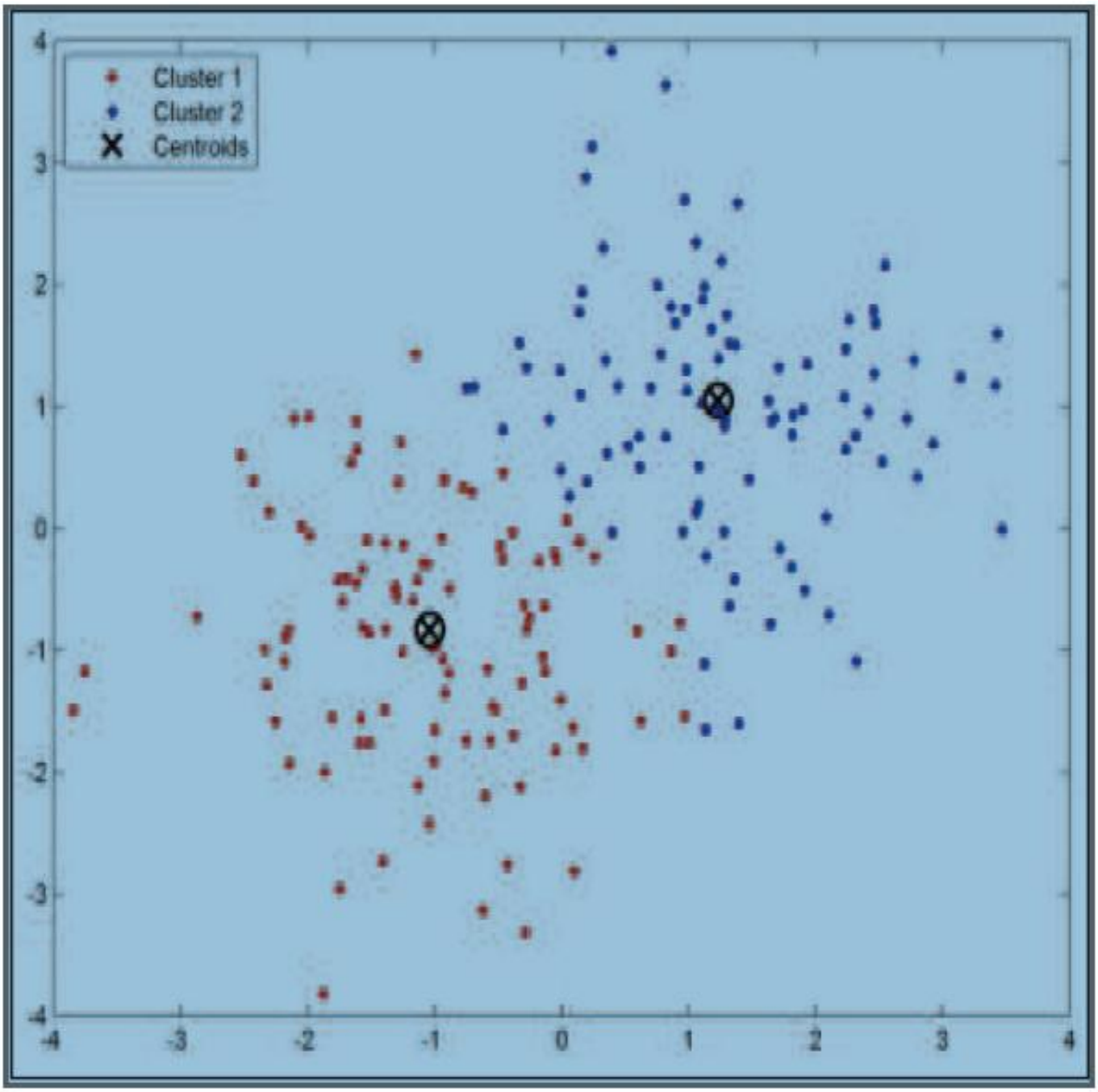

Figure 5. Plot a graph of separated random data for two clusters

REFERENCES

1. *Arthur David and Vassilvitskii Sergei (2007), "k-means++: The Advantages of Careful Seeding", SODA '07: Proceedings of the eighteenth annual ACMSIAM, 1027-1035.*
2. *Banerji Geetali and Saxena Kanak (2012), "An Efficient Classification Algorithm for Real Estate domain", International Journal of Modern Engineering Research (IJMER), 2(4): 2424-2430.*
3. *Basheer M., Al-Maqaleh, Hamid Shahbazkia, (2012), "A Genetic Algorithm for Discovering Classification Rules in Data Mining", International Journal ofComputer Applications (0975 – 8887), 41(18): 40-44.*
4. *Bauer E. and Kohavi R. (1999), "An empirical comparison of voting classification algorithms bagging, boosting and variants", Machine Learning, 36(1/2):105–142.*

5. *Bifet A., Holmes G., Pfahringer B., Kirkby R., Gavald`a R.(2009), "New ensemble methods for evolving data streams", In KDD, ACM, 139-148.*
6. *Blake CL. and Merz CJ. (1998), "UCI Repository of machine learning databases. http:// www.ics.uci.edu/_mlearn/MLRepository.html, University of California, Irvine, Department of Information and Computer Science".*
7. *Bora, Jyoti, Dibya and Dr. Gupta ,Anil ,Kumar (2014), "A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm", International Journal of Computer Trends and Technology (IJCTT), 10 (2): 108 -113.*
8. *Breiman L. (1996), "Bagging predictors", Machine Learning, 24(2): 123–140. Chamundeswari, G., Prof. Varma, G., Pardasaradhi& Prof. Satyanarayana, Ch. (2012), "An Experimental Analysis of K-means Using MATLAB", International Journal of Engineering Research & Technology (IJERT), 1 (5): 1-5.*
9. *Cieslak D. and N. Chawla N.(2009), "A framework for monitoring classifiers performance: when and why failure occurs?)", KAIS. 18(1): 83–108.*
10. *DerSimonian, R. (1986), "Maximum likelihood estimation of a mixing distribution", J. Roy. Statist. Soc. C., 35:302–309.*
11. *Dietterich T. (2000), "An experimental comparison of three methods for constructing ensembles of decision trees Bagging, boosting, and randomization", Machine Learning, 40(2), 139–158.*
12. *Dietterich T. (2002), "Ensemble learning, in The Handbook of Brain Theory and Neural Networks", 2nd ed., M. Arbib, Ed., Cambridge MA: MIT Press.*
13. *Dr. Bhatia, M.P.S. and Khurana, Deepika (2013), "Experimental study of Data clustering using k-Means and modified algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP),3(3): 17-30.*
14. *Dr. Hemalatha M. and Saranya Naga N. (2011), "A Recent Survey on Knowledge Discovery in Spatial Data Mining" IJCSI International Journal of Computer Science, 8 (3):473-479.*
15. *Dzeroski S. and Zenko B. (2004), "Is combining classifiers with stacking better than selecting the best one? Machine Learning", 255–273.*
16. *Efron B. and Tibshirani R., (1993) "An Introduction to the Bootstrap", Chapman & Hall, New York. Elena Makhalova, (2013), "Fuzzy C means Clustering in MATLAB", The 7th International Days of Statistics and Economics, Prague, 19(21): 905-914.*
17. *Ester Martin, Kriegel Peter Hans, Sander Jörg (2001), "Algorithms and Applications for Spatial Data Mining Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS", Taylor and Francis, 1-32.*