Data Mining for Biological Problems Using K-Means Algorithm: A Clustering Approach for Biomedical Signal and Genomic Data Analysis

Dr. Arjun Mehta[1]*, Dr. Priya Sharma[1]
[1]Indian Institute of Technology Bombay, Department of Computer Science and Biomedical Data Analytics, Mumbai, India

ABSTRACT

Data Mining is a knowledge discovery from data and it treats as mining ofknowledge from large amount of data in every field. The algorithms are implemented using MATLAB and fuzzy logic tool box and results are evaluated based on performance parameter in both algorithms. After doing this research experiment results show that how k-means and fuzzy C means implemented on protein data set. In this research work we present the problem that show proteins are highly affiliated to each other.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. Forest fire is important to the forest ecosystem.

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

I. INTRODUCTION

Data mining knowledge and corporal perfect from data has exposed for periods. Initial methods of categorizing projects in data include regression examination (1800s) and Bayes' theorem (1700s). The propagation, ubiquity and rising control of computer knowledge has better approach talent, supposed, data collection and larder. Data mining ideas build up more facility for study. As data sets have adult in size has more and more been augmented with unintended, mechanical data dispensation, assisted by other detections in information technology and computer science such as neural networks, cluster analysis, genetic algorithms (1950s).

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 in databases to tool the sure knowledge and detection procedures more capably. Data mining letting such approaches to be useful to a big of data circles.

Data mining is a flawlessly inter-displinary business. Data mining can be clarifies in numerous dissimilar conducts. Data mining should have been additional different called first is information mining from database and another is information detection from data usual.

Numerous persons give data mining and thoughts as a substitute. It's aimed at additional commonly charity period and the others opinion data mining as just a vital stage in the sequence of data mining.

We are living in a sphere wherever lots amounts of data are tranquil every day. Investigating such numbers is an actual important essential. Data mining can meet this need by providing tools; methods and techniques to discover knowledge from real data .We examine how data mining can be viewed as a result of the progression of information technology, science and engineering. We are living in the information age. This age is a usual saying; though, we are truthfully revenue in the data stage which is collection alike to a best. Data mining damaged to the World Wide Web (www) and varied data larder plans each day from commercial, medicinal discipline, medication, bio-informatics, discipline, engineering unevenly each additional part of everyday lifetime. In these fields increase of offered data

volume is a result of the mechanization of our society. In businesses worldwide create gigantic data sets, including stock, transactions, sales trading records, promotions, company profile and routine and customer pointer.

In current ages, data mining have been damaged in numerous parts of information and engineering, such as bio-informatics, medication, fitness attention, physics and environment. In the studies of humanoid heredities, inheritances, arrangement withdrawal assistances challenge the actual significant goalmouth of tolerant. This considerate the charting the bury VIP differences in existence DNA/RNA classification and protein cable the randomness in illness honesty. In simple terms, it aims to discover how the changes in an folks DNA/RNA sequence distresses the dangers of emerging shared illnesses such as cancer, The data mining arrangement is secondhand to do this job. This is recognized as multi-factor dimensionality discount.

The MEROPS database is an in a noise standby for peptidases and the proteins that decrease them. The MEROPS database usages and group founded classification of the peptidases. In this, composed peptidases are selected to a relative on the basis of resemblances in amino acid sequence and relations that are view to be homologous are gathered together in a Clan. We chat about gastricsin PEPTIDASE. Its MEROPS ID is A01.003. It's additional name pepsin C, seminal progastricsin. Action location is person and mouse. Polymorphism relates with gastric ulceration. We are taking 436 sequence dimensions from H-Sequence in protein data set.

Protein name- gastricsin other name- Pepsin C, seminal progastricsin
Cleavages-50
Classification-CLAN AA>> SUBClan(none)>>Family A1>> Subfamily A>> A01.003
Family A1-Pepsin A(Homo sapiens)

H sequence

1	MRGLVVFLAVFALSEGQCHHQVSVNLLRVPLHKGKSLRRALRERRLLEDFLRNHHYAVSR	60
61	KHSSSGVVASESLTNYLDCEWLCLQCQYFGKIYIGTPPQKFTLVFDTGSPDIWVPSVYCN	120
121	SDACREXPVFQKNHQRFDPSKSSTHQNMGKSLSIQYGTGSMRGLLGYDTVTVSGVSCRVS	180
181	NIVDPHQTVGLSTQEPGDIFTYSEFDGILGLAYPSLASEXSVPVFDNTMQRHLVAQDLFS	240
241	VYMEQVRAVGFRNDQGSMLTLRAIDLSYYTGSLHWIPMTQEYWQFTVDRWVSPLESVIID	300
301	GVVVACDGGCQAILDTGTSLLVGPGGHILNIQQAIGATAGQYNEVRSGFFSFDIDCGRLS	360
361	SIPTAVFEIHGKKYPLPPSAYTSQVXVSFPQDQGFCTSGFQGDYSSQQWILGNVFIWEYY	420
421	SVFDRTNNRVGLAKAV	436

II. OVERVIEW OF K-MEANS ALGORITHM

The k-Means clustering algorithm (Ghosh and Dubey, 2013) is unique of the mainly in dividing algorithms. It is the greenest variety of cluster inspection is dividing. In this methods list data into k clusters where k is the contribution restraint separate in continue finished iterative transfer process which touches to local least. Initially K-means is to found k centroids at chance one for every Cluster. Then is to intended standby amongst truths opinions for datasets and the cluster centers and transmission the data point. Data points are combined in certain clusters then first association is complete. New middles are thenintended by charming the regular of opinions in the clusters. This is complete since of inclusion of novel points may companion to differ in cluster middles. The worth of cluster can be exact by the inside cluster difference which is the sum of shaped mistake amid all substances is clear as:

$$E = \sum_{j=1}^{n} \sum_{p \in c_i} dist(p, m_i)^2$$

Where E is the adding of the plaza responsibility for each bit of material in the data set p is the data point and $m_i$is the centroid of cluster c_i.

K-means allocates each viewing to the cluster. Which pays within-cluster sum ofsquares (WCSS)? Sum of squares is label the Eulidean coldness.

$$E_i^n = \{d_m : ||d_m - v_i^n|| \leq ||d_m - v_j^n|| \; \forall, 1 \leq j \leq k\}, \qquad (2)$$

Where each d_m is assigned to unerringly one E_i^n, even if it could be assigned to twoor more of them.

1. Enter the rate of k.
2. Pick a primary n centers C = {c_1, c_2, ……,c_k}
3. Replicate
4. Re (assign) every one things k to the cluster to which the entity is the mostparallel. It is based on the mean rate of the object in the cluster k.
5. Calculate the normal of the inspection in each cluster to obtain k newcentroid site.
6. Undecided no adapt.

III. RESULTS AND DISCUSSION

Fuzzy C Means algorithm-
For clusters 2,3,4,5
Table 1, Table 2, Table 3 and Table 4show that FCM value of n where n is the no of clusters which change for each run. FCM shows that how performance parameter varies for each run. We determine Sum of distance value of objective function during iteration for data. For n=2,3,4,5 n is the no of clusters.
Data set X=436

Table 1: Sum of distance according to iteration n=2

Iterations	Sum of distances (obj_fun)
1	42.283798
58	30.895203

Table 2: Sum of distance according to iteration n=3

Iterations	Sum of distances (obj_fun)
1	46.460376
98	34.335733

Table 3: Sum of distance according to iteration=4

Iterations	Sum of distances (obj_fun)
1	48.407830
100	36.660185

Table 4: Sum of distance according to iteration=5

Iterations	Sum of distances (obj_fun)
1	48.298797
19	36.266996

Figure 1, Figure 2, Figure 3 and Figure 4show that the objective function value during iteration.

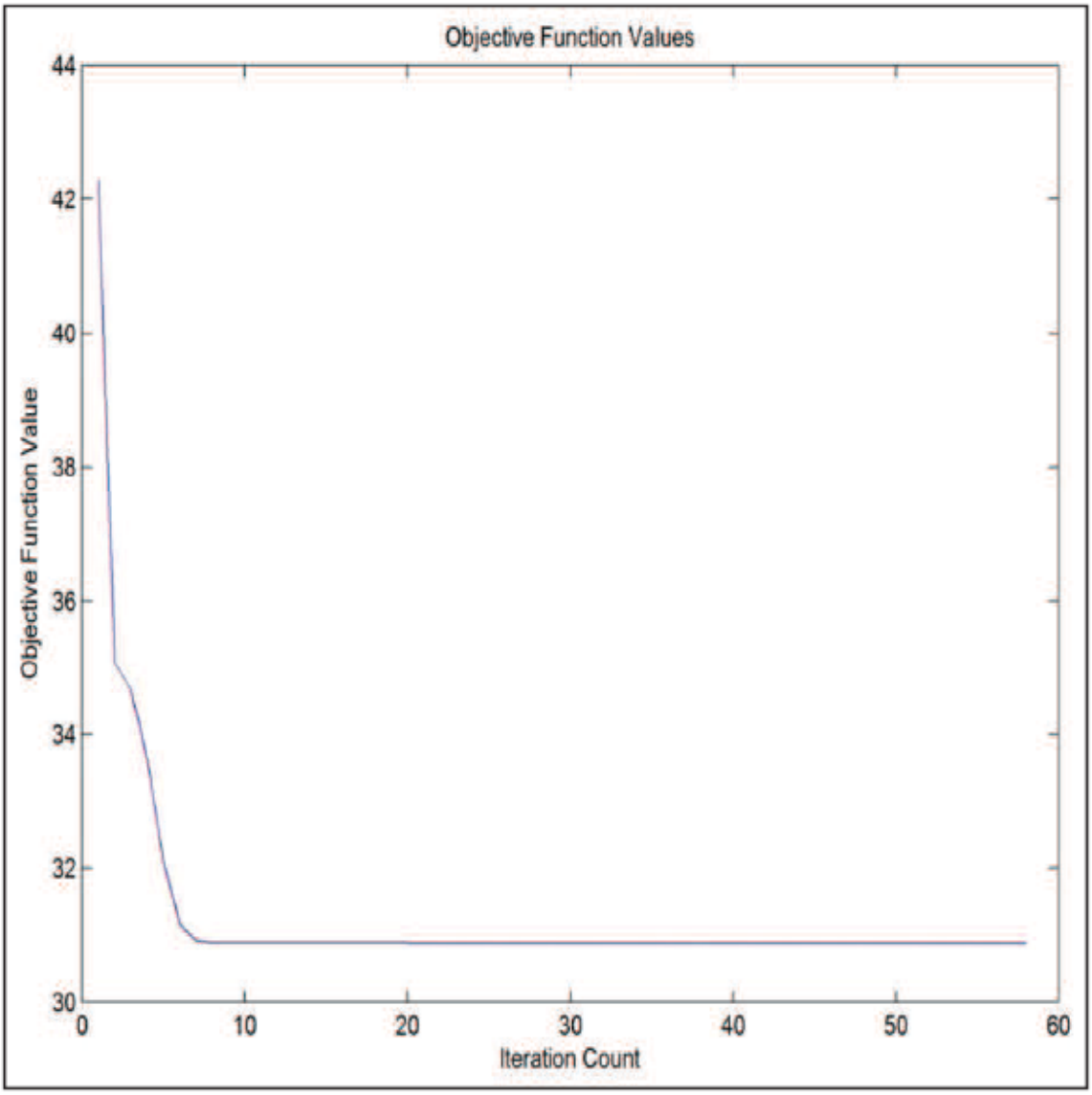

Figure 1: Graph of objective function value for n=2

Graph of clusters 2

Figure 2 shows that data point in the data set a degree of membership toeach cluster in fuzzy logic toolbox.

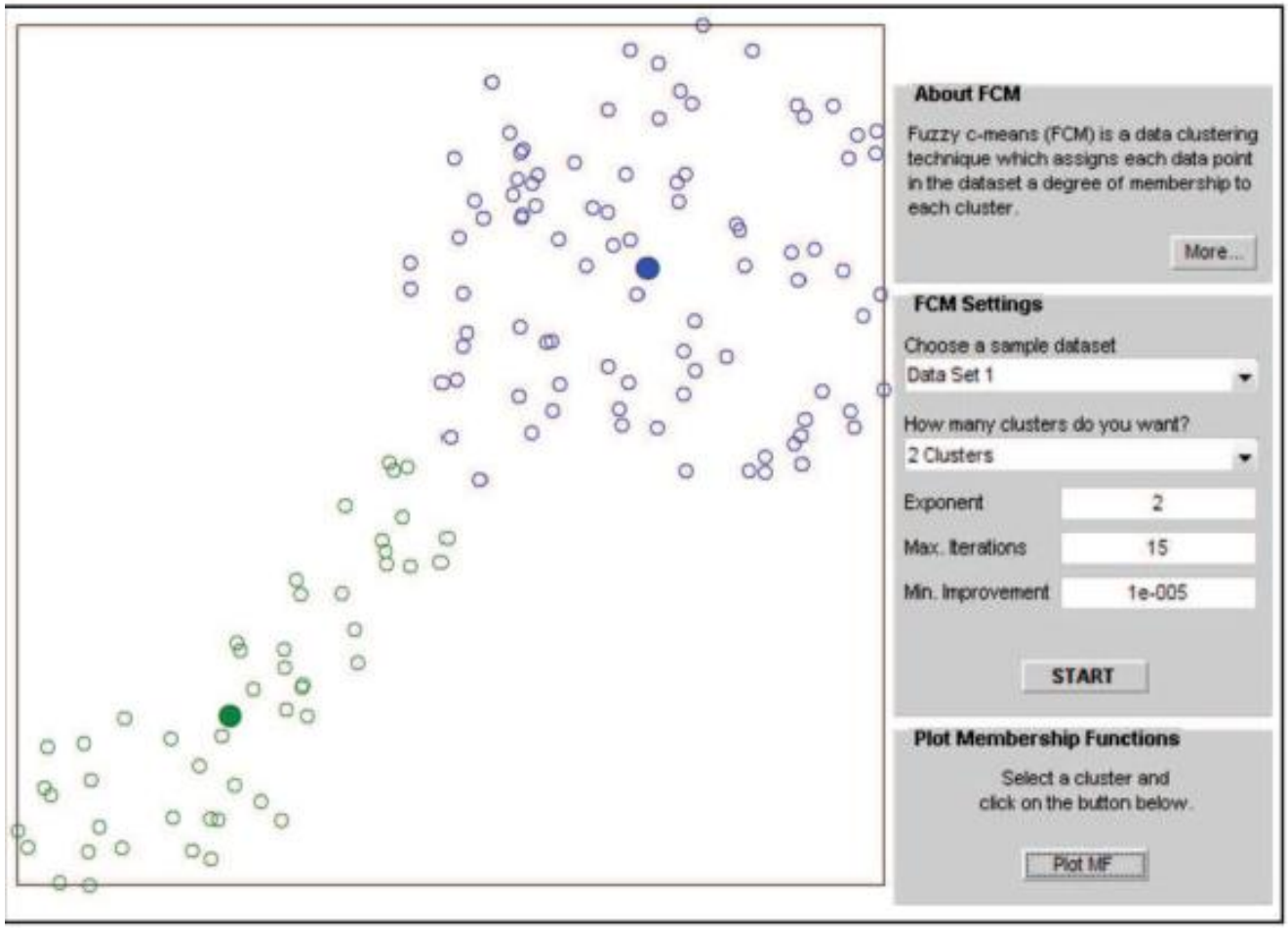

Figure 2: 2D Image data point in the data set a degree of membership to each cluster

.

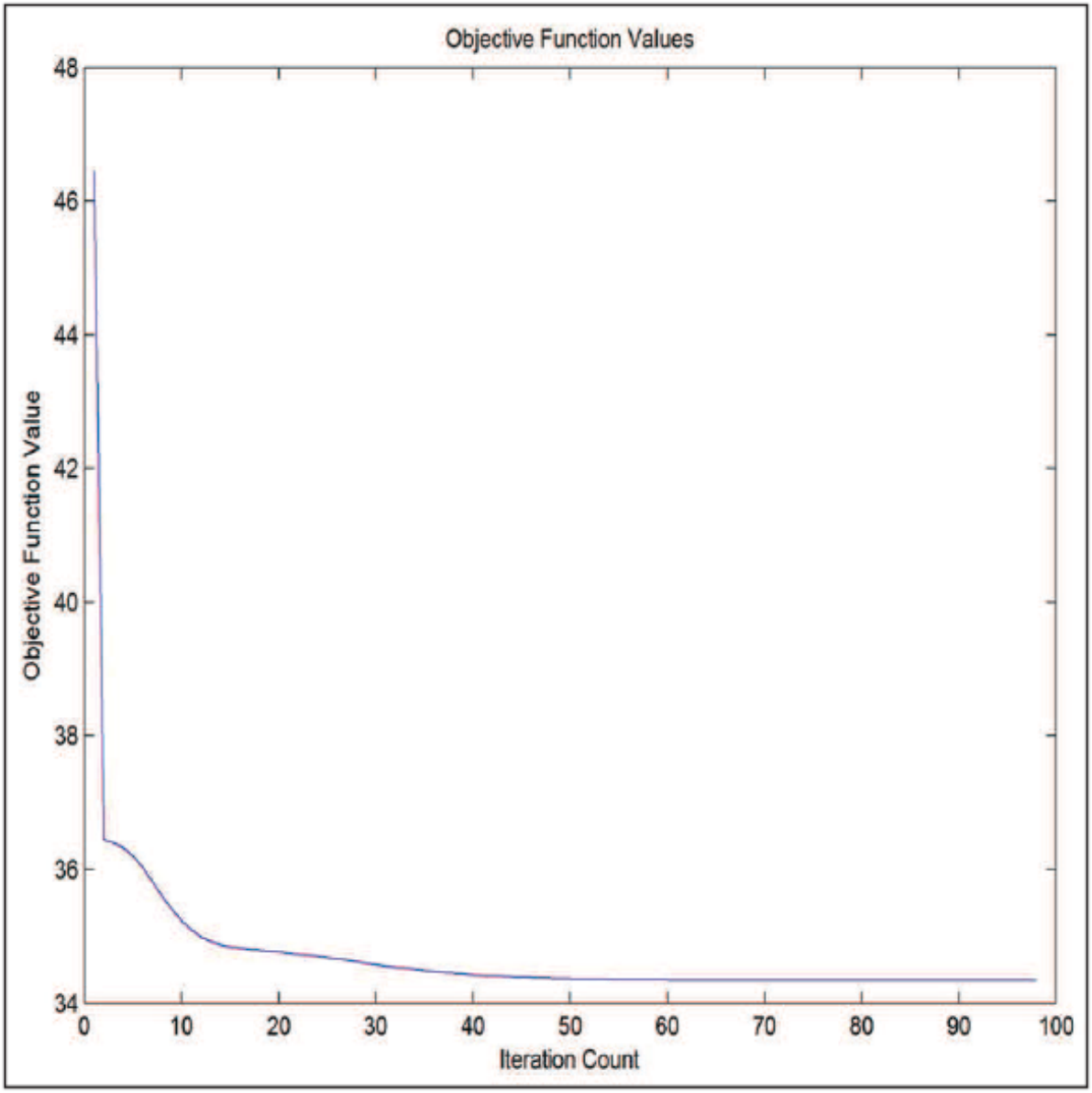

Figure 3: Graph of objective function value for n=3

Graph of clusters 3

Figure 4 shows that data point in the data set a degree of membership to each cluster in fuzzy logic toolbox.

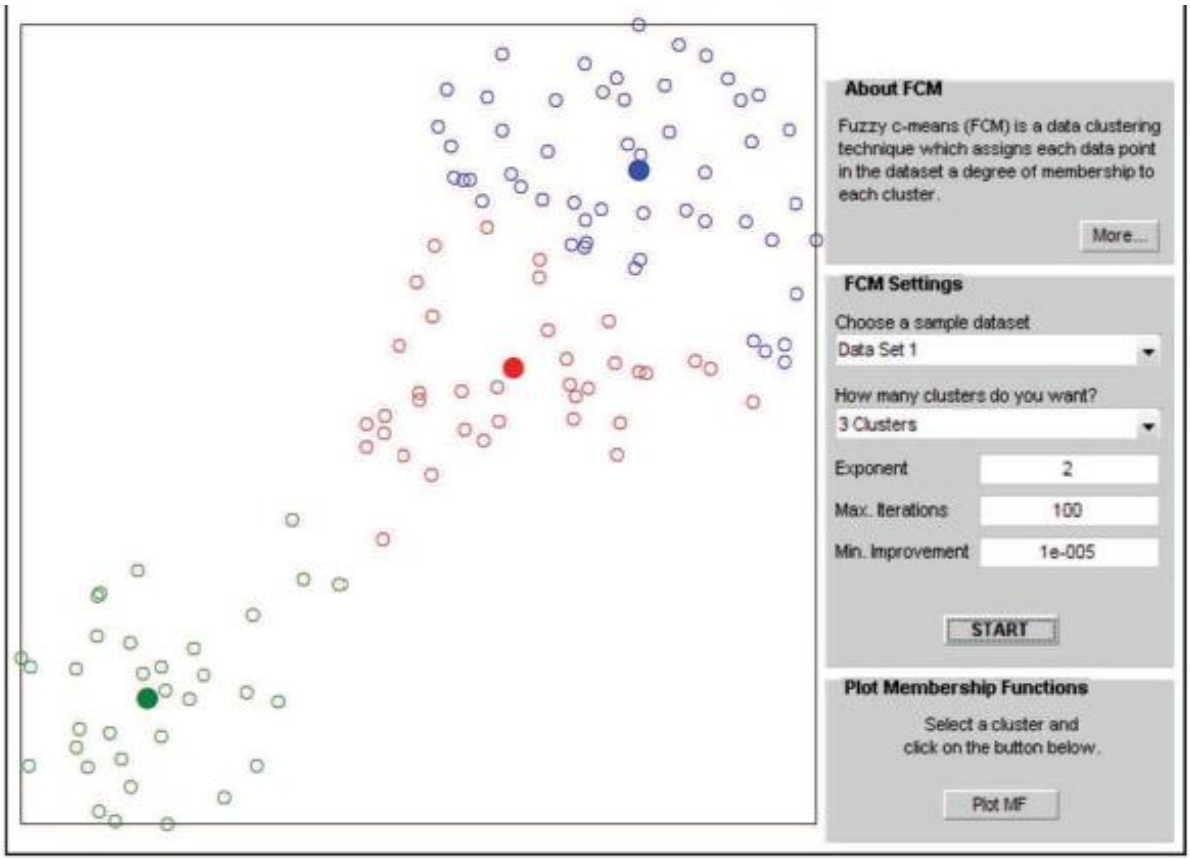

Figure 4: 2D Image data point in the data set a degree of membership to each cluster

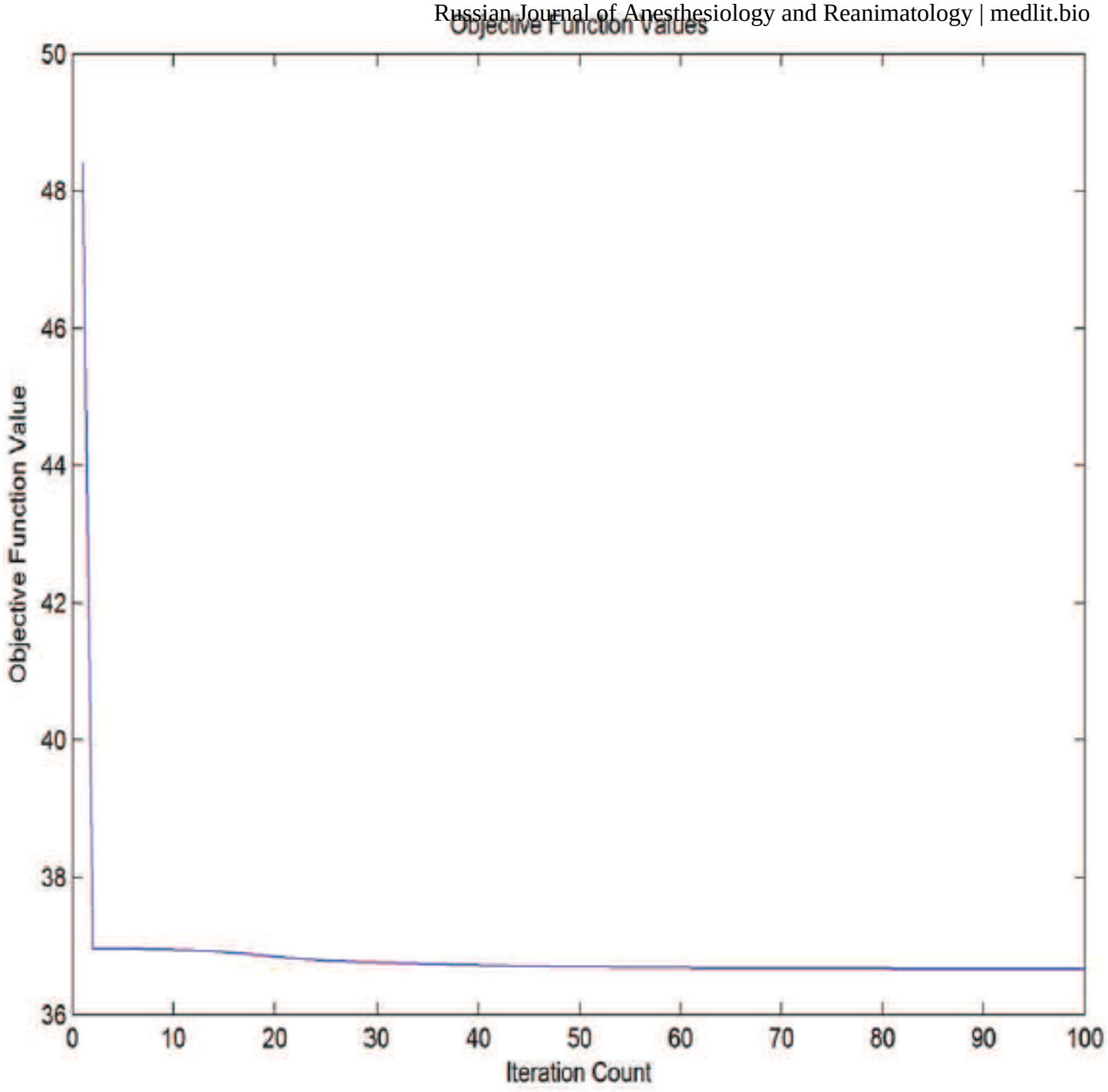

Figure 5: Graph of objective function value for n=4.

REFERENCES

1. *Jain Shreya and Gajbhiye Samta (2012), "A Comparative Performance Analysis of Clustering Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, 2(5):441-445.*
2. *Jenna Reps, Jonathan M. Garibaldi, Jack E. Gibson, Richard B. Hubbard (2013), "Comparing Data-mining Algorithms Developed for Longitudinal Observational Databases" , The 12th Annual Workshop on Computational Intelligence, Heriot-Watt University, 1-8.*
3. *Kalyankar A. Meghali and Prof. Alaspurkar S. J. (2013), "Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Researchin Computer Science and Software Engineering, 3(2):114-118.*
4. *Kanevski M., Parkin R., Pozdnukhov A., Timonin V., Maignan M., Yatsalo B., Canu S. (2004) "Environmental Data Mining and Modelling Based on Machine Learning Algorithms and Geostatistics", Environmental Modellingand Software, 414-419.*
5. *Kavitha K. and Sarojamma R M (2012), "Monitoring of Diabetes with Data Mining via CART Method", International Journal of Emerging Technology and Advanced Engineering, 2 (11):157-162.*
6. *Kumar, Amit, Sabharwa, Yogish, Sen, Sandeep (2004)," A simple linear time (1+_)- approximation algorithm for k-means clustering in any dimensions", In FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04), Washington, DC, USA, IEEE ComputerSociety, 454-462.*
7. *Lavra Nada and Novak Kralj Petra (2013), "Relational and Semantic Data Mining for Biomedical Research", Informatica (Slovenia), 37(1): 35-39.*
8. *Lai J-S and Tsai F.(2012),"Verification And Risk Assessment For Landslides In The Shimen Reservoir Watershed Of Taiwan Using Spatial Analysis And Data Mining", International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXIX(B2): 67-70.*
9. *Yifeng, Ngom, Alioune (2013), "The Non-Negative Matrix Factorization Toolbox for Biological Data Mining", Source Code for Biology &Medicince, 8 (1): 1-18.*