ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

Implementing K-Mean clustering method on genes on chromosome1 (Homo sapiens)

Yusuf Talib

Research Student, Dept. of Biotechnology, MGM Institute of Health Sciences, Medical College & Hospital,

Aurangabad. MS. India

ABSTRACT: Statistics is one of the methods which is utilised by the every stream of life sciences, out of which one method is short listed to find out the information hidden in the biology, here in this article I tried to implement the K-mean clustering method to find out the clusters, normal distribution and chi square test, etc.

K-means clustering algorithm is an old algorithm that has been intensely researched owing to its ease and simplicity of implementation. This algorithm has a broad attraction and usefulness in exploratory data analysis. The application of k-mean in various broad areas of artificial intelligence, customer-relationship management, data compression, data mining and image processing provides valuable information for further analysis. This paper presents results of the experimental study of different approaches to cluster means on genes located on chromosome1 using mathematical software STATISTICA. The results are recorded on some performance measures such as normal distribution, K-Mean calculation, frequency distribution, Chi-square test, and dendogram.

KEYWORDS: clustering, K-means, chromosome1, STATISTICA

I. INTRODUCTION

Data mining is classified as search for knowledge from large and huge amount of data using data mining methods we can classify the nature and aspects of any type of data. Data mining and analysis includes lots of protocol, algorithm ans methods, here in this paper I had deal with K-mean clustering method to analyse the data of chromosme1.
Clustering

One of the traditional viewed unsupervised methods for data analysis is clustering, which allows us to group the data instances according to the notions of similarity. Cluster analysis itself is not one specific algorithm it can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and efficiency can be checked. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery. It will often be necessary to modify data pre-processing and model parameters until the result achieves the desired properties. There re lots of various models available for cluster analysis, in this paper tried to focused on implementing K-Mean cluster for data of Chromosome1.

K-Means

K-means clustering (MacQueen, 1967) is commonly used method for partitioning the data sets into groups (k)(1), firstly it creates centers (k) and then iteratively refining them until the last group remain or no change or no more cluster remains, the algorithm converges when there is no further change in assignment of instances to cluster. The clustering method can be given like (picture 1)

DOI: 10.15680/IJIRCCE.2015. 0309050

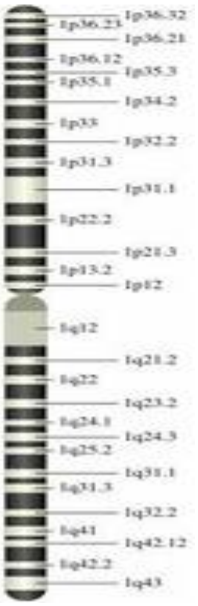
$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

Picture 1 equation of K-mean

Here:- '||xi - vj||' is the Euclidean distance between xi and vj.
'ci' is the number of data points in ith cluster.
'c' is the number of cluster centers.

Chromosome1

Genetic material of each and every organism is located on chromosome, if researcher want to deal with the genes and there information they have to first find out the location on chromosome. In this paper I focused in on genes located on chromosome1(Picture 2) of Homo sapiens species, chromosome1 is largest human chromosome having 249 million nucleotide base pairs and consist of 9% of total DNA in human cells, chromosome1 is currently thought to have 4,316 genes(2).



Picture 2 Chromosome1 structure

STATISTICA

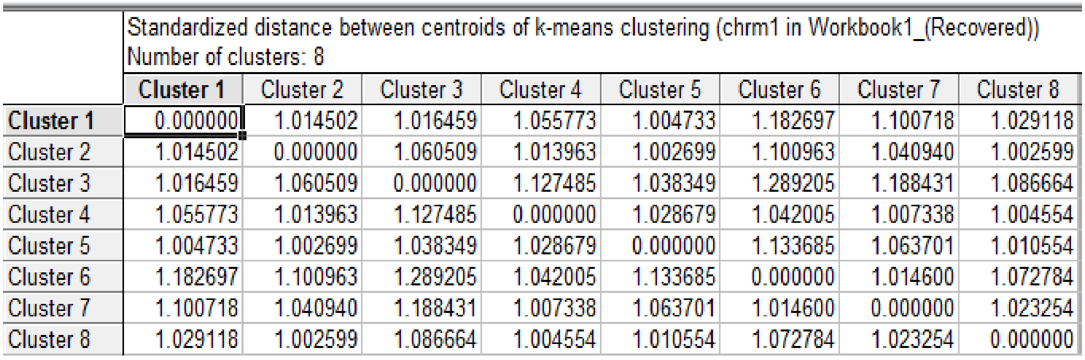
Data mining, data visualization and data management are some of the features provided by STATISTICA, an analytical software products and solution provided by Stat Soft (3). Also some other features like graphical representation of result add more colour to the result which help researcher to interpret the result with ease.

II. Related work

In (1) Authors analysed and implemented the efficiency of K-mean clustering method, there aim was to minimise the mean squared distance from each data point to its nearest centre, author presented the simple and efficient implementation of Lloyd's K-mean clustering method. In (4) Authors suggested the problem of object clustering according to its attributes has been widely studied due to its application in different areas like machine learning, data mining, knowledge discovery, and pattern recognition. There aim was to partition a set of objects which have associated multi dimensional attribute vectors into homogeneous groups such that the patterns within the groups are similar. Authors (5) proposed a method based on information obtained during the k-means clustering operation itself to select the number of clusters. There method employs an objective evaluation measure to suggest suitable values for K, thus avoiding the need for trial and error. In Author (6) suggested that the K-mean methods is one the typical clustering algorithm which aims to partition N inputs (data points). Authors (7) proposed the modified clustering method which created initial set of k partitions, which further utilised for iterative relocation technique that attempts to improve the partitioning of the data points. It is numerical, unsupervised iterative method, author have implemented 3 algorithms and performance is measured on the basis of there defined parameters.

III. Methodology

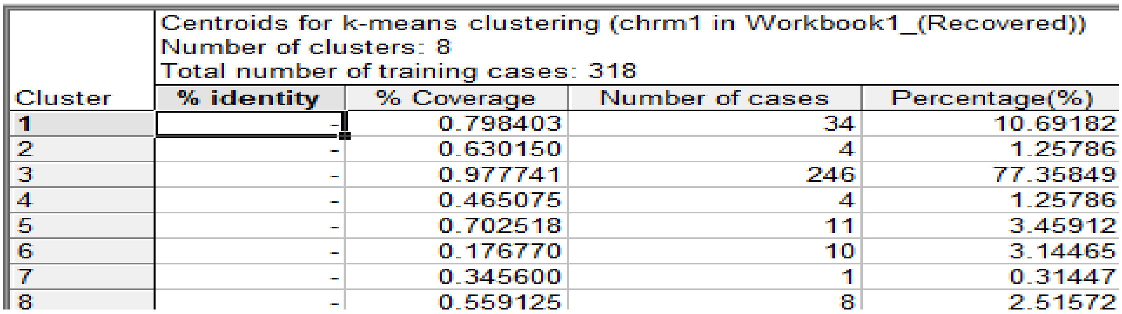
In this paper tried to implement the statistical approach of k-mean to the genes of chromosome1 using the STATISTICA mathematical application. Here in this paper data has been extracted from biological databank GEO (NCBI) 7 (seven) different entries or data has been short listed and saved in datasheets, which is going to be work as an input in STATISTICA for analysis. Emphasis was to get information of clusters in data of chromosome1. Then different clustering methods k-mean (picture1), Dendogram has been implemented to the data of chromosome using %coverage in datasheets (sheet1) which is the centre point as required by clustering methods which has to define earlier and gave 8 different clusters on the basis of centre point.



Standardized distance between centroids of k-means clustering (chrm1 in Workbook1_(Recovered))
Number of clusters: 8

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Cluster 1	0.000000	1.014502	1.016459	1.055773	1.004733	1.182697	1.100718	1.029118
Cluster 2	1.014502	0.000000	1.060509	1.013963	1.002699	1.100963	1.040940	1.002599
Cluster 3	1.016459	1.060509	0.000000	1.127485	1.038349	1.289205	1.188431	1.086664
Cluster 4	1.055773	1.013963	1.127485	0.000000	1.028679	1.042005	1.007338	1.004554
Cluster 5	1.004733	1.002699	1.038349	1.028679	0.000000	1.133685	1.063701	1.010554
Cluster 6	1.182697	1.100963	1.289205	1.042005	1.133685	0.000000	1.014600	1.072784
Cluster 7	1.100718	1.040940	1.188431	1.007338	1.063701	1.014600	0.000000	1.023254
Cluster 8	1.029118	1.002599	1.086664	1.004554	1.010554	1.072784	1.023254	0.000000

Sheet1 Standardized distance between centroids using k-mean clustering



Centroids for k-means clustering (chrm1 in Workbook1_(Recovered))
Number of clusters: 8
Total number of training cases: 318

Cluster	% identity	% Coverage	Number of cases	Percentage(%)
1	-	0.798403	34	10.69182
2	-	0.630150	4	1.25786
3	-	0.977741	246	77.35849
4	-	0.465075	4	1.25786
5	-	0.702518	11	3.45912
6	-	0.176770	10	3.14465
7	-	0.345600	1	0.31447
8	-	0.559125	8	2.51572

Sheet2 K-mean of different clusters and number of cases

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

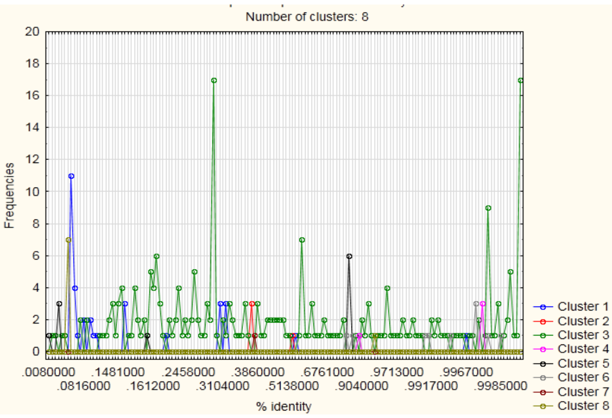


Figure1Frequency graph of cluster

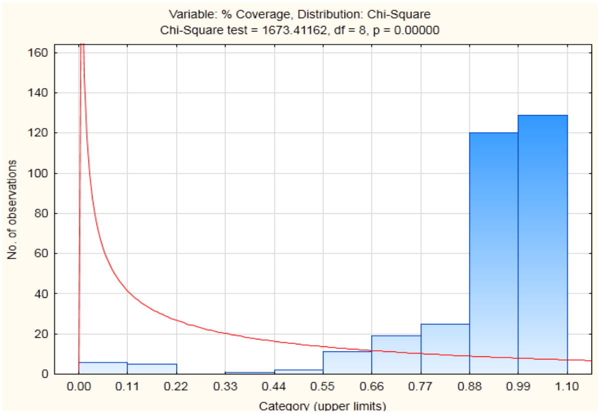


Figure2 Normal distribution of variable % coverage

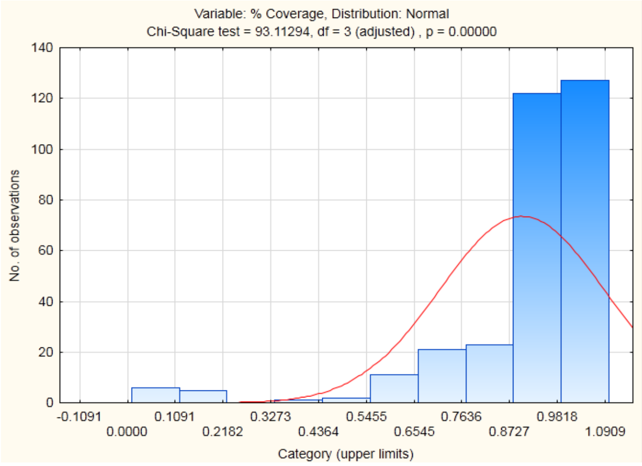


Figure3 Chi square test of variable % Coverage

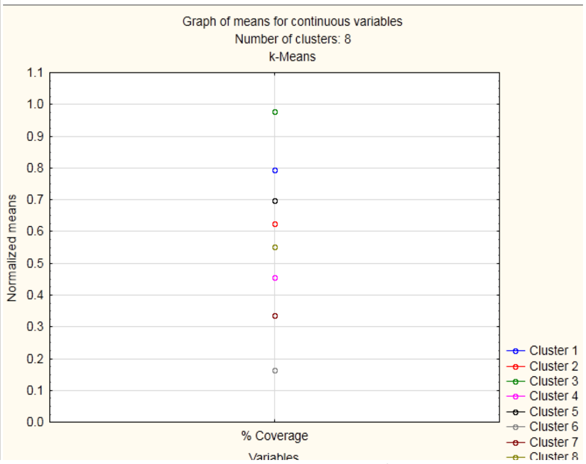


Figure4 Normalized mean of % coverage

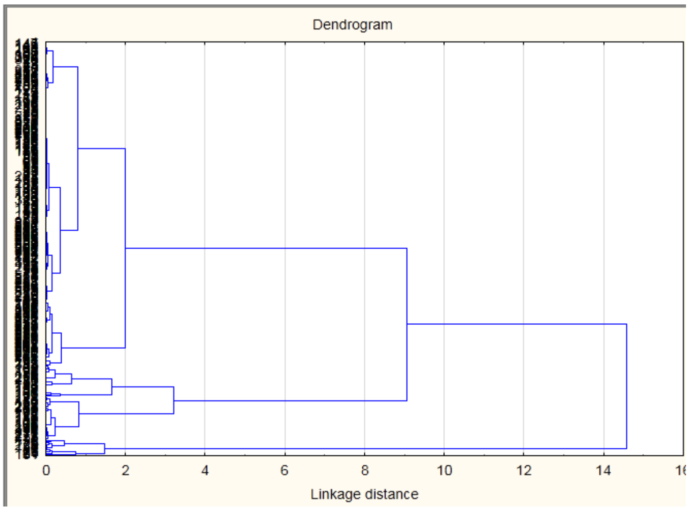


Figure5 Dendogram in STATISTICA

IV. CONCLUSION

Implementing the clustering methods to the gene data of chromosome1 we conclude that the 8 clusters (see sheet 2) which has been obtained after implementing the statistical approach through **STATISTICA** in which Cluster no.3 having many gene involved which can be concluded in it as on the basis of %coverage **0.97,** number of cases **246, Centroids** percentage % **77.35**, also graph frequencies (see figure1) of cluster confirms the information of cluster 3, also chi square test gave the confirmation **1673.41** (see figure 2), also confirmation done by the normal distribution(figure3), frequency (figure4), Dendogram (figure5).In this paper goal was to find out the some interesting mathematical and important clusters in chromosome1 gene data which has been achieved. Conclusion is pulled that mathematical equations algorithms are helpful to find out the interesting information of data, and can implement those methods to smaller and larger scale for analysis. In future plan is to proceed to check the functions and structure of each cluster and this k-mean method can be implemented to get the clusters on large scale data of complete genome of *Homo sapiens and other organisms.*

REFERENCES

1) An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.
2)Table 2.3: Human chromosome groups". Human Molecular Genetics (2nd Ed.). Garland Science. 1999.
3) http://www.statsoft.com/products/statistica-12-new-features/
4) Research issues on K-means Algorithm: An Experimental Trial Using Matlab by Joaquin Perez Ortega, Ma. Del Rocio Boone Rojas and Maria J. Somodevilla Garcia.
5) The k-means algorithm - Notes by Tan, Steinbach, Kumar Ghosh.
6) k-means clustering by ke chen.
7)Ran Vijay Singh and M.P.S Bhatia , “Data Clustering with Modified K-means Algorithm”, IEEE International Conference on Recent Trends in Information Technology, ICRTIT 2011, pp 717-721.

BIOGRAPHY

Talib Yusuf Abbas Hussain is a research student in Department of Biotechnology, MGM University of Health Sciences, Navi Mumbai, India. He received Master of Bioinformatics degree in 2007 from BAMU University, Aurangabad,MS,India. His research interest area is in gene expression, phylogenetic tree construction and analysis, genetic engineering, etc.