© 2016 IJSRSET | Volume 2 | Issue 4 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Fruit Fly K-Means Clustering Algorithm

D. Gowdham*, K. Thangavel, E. N. Sathish Kumar

Department of Computer Science, Periyar University, Salem, Tamil Nadu. India

ABSTRACT

Clustering is one of the main data mining tasks. It aims to grouping the data objects into significant clusters such that the similarity of objects within clusters is maximized, and the similarity of objects from different clusters is minimized. K-Means algorithm is most commonly used algorithm for unsupervised clustering problem. But it has some problems which make it unreliable. Initialization of the random cluster centers, number of clusters and terminating condition play a major role in quality of clustering achieved. In this paper we proposed Fruit Fly algorithm to select the initial centroids for K-Means algorithm in order to optimize the number of clusters. The experimental analysis is conducted on Cocaine dataset to validate the proposed method.

Keywords: Gene Expression, Microarray Dataset, K-Means clustering, Fruit Fly Optimization Algorithm, Fruit Fly K-Means Algorithm

I. INTRODUCTION

A gene is a small piece of genetic material written in a code and called DNA. Each gene has within it a set of instructions for making molecules that organisms need to survive. Genes themselves cannot be used by an organism. Instead they must be turned into a gene product. Gene expression is the process by which the information contained within a gene becomes a useful product. Then we apply the clustering algorithm on cocaine dataset. The Fruit Fly Optimization Algorithm was invented by Prof. Pan, a scholar of Taiwan. It is a new method for deducing global optimization based on the foraging behavior of the fruit fly. The sensory perception of the fruit fly is better than that of other species, especially the sense of smell and vision. The olfactory organ of a fruit fly can gather various smells from the air, and even a food source 40km away. Afterwards, the fruit fly flies to the food, uses its acute vision to find the food and where its fellows gather, and then it flies in that direction. The "K" in the K-Means algorithm stands for number of clusters as an input and the "Means' stands for an average, location of all the members of a particular cluster. K-Means more efficient and produce good quality clusters. In the K-Means algorithm cluster results are highly depends based on initial centroids. There are different techniques available to generate the initial points which are mandatory at the outset of the process of the algorithm. However, we are hereby focusing on the centroid method. The centroid method consists of choosing all the starting clusters close to the mass centroid of the dataset. The center of each cluster is calculated by adding a small random perturbation to the centroid of the dataset. A centroid represents an average location of the particular cluster. So we can implement the algorithm to select the initial centroid using optimization algorithm. We can combine fruit fly algorithm with K-Means to select the best centroid to initialize.

II. METHODS AND MATERIAL

The methodology discusses k-means clustering with Fruit fly optimization algorithm to find the best centroid for initialize the K -Means.

A. Fruit Fly Optimization Algorithm

The Fruit Fly Optimization Algorithm [1] was invented by Prof. Pan, a scholar of Taiwan. The Fruit Fly Optimization Algorithm is a new method for finding global optimization based on the food finding behavior of the fruit fly. The fruit fly itself is superior to other species in sensing and perception, especially in smell

and vision. The smell organs of fruit flies can find all kinds of scents floating in the air, it can even smell food source from 40 km away. Then, it gets close to the food location. The working concept of fruit fly algorithm is shown is figure 1.

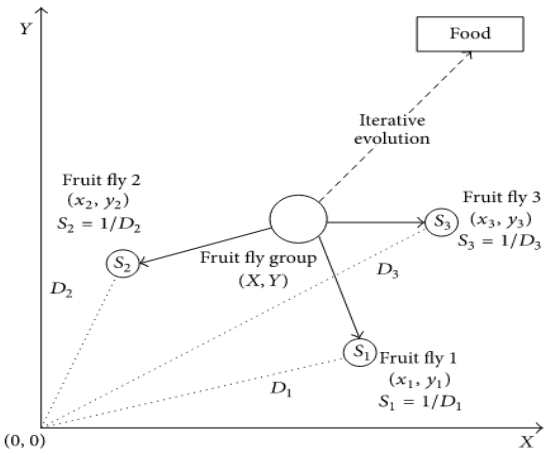

Figure 1. Working of Fruit fly

Algorithm: Fruit Fly optimization algorithm

Step 1: Random initial fruit fly swarm location.

Init X_axis

InitY_axis

Step 2: Give the random direction and distance for the search of food using smell by an individual fruit fly.

X_i = X_axis + Random Value

Y_i =Y_axis+Random Value

Step 3: Since the food location cannot be known, the distance to the origin is thus estimated first (), then the smell concentration judgment value ( ) is calculated, and this value is the reciprocal of distance:

$Disti = i_2 + Y_{i2}$

$S_i = 1/Dist_i$

Step 4: Substitutes smell concentration judgment value (S) into smell concentration judgments function (or called Fitness function) so as to find the smell concentration (Smelli) of the individual location of the fruit fly.

$Smell_i$ = Function (S_i)

Step 5: Find out the fruit fly with maximal smell concentration value and then let the fruit fly move towards the best one

[BestSmell BestIndex = Max (Smell)]

Step 6: Keep the best smell concentration value and X, Y coordinate, and at this moment, the fruit fly swarm will use vision to fly towards that location.

SmellBest = BestSmell

X_axis = X (Best Index)

Y_axis = Y (Best Index)

Step 7: Enter iterative optimization to repeat the implementation of step 2 to step 5, then judge if the smell concentration is superior to the previous iterative smell concentration, if so, implement step 6.[2]

B. K-Means Clustering Algorithm

K–Means clustering is a method of cluster analysis which aims at portioning of n observations into k clusters. Each of the observation belongs to a cluster with the minimum distance between cluster centre and the observation point. It is done iteratively so that the observation point is at least distance from the centre of cluster. The mean distance between the cluster centre and observation is minimized during this iteration process. [3]

Algorithm: K-Means

Require: D = {d_1, d_2, d_3,..., d_n } // Set of n data points.
K - Number of desired clusters
Ensure: A set of K clusters.
Steps:

1. Arbitrarily choose k data points from D as initial Centroids;
2. Repeat

Assign each point di to the cluster which has the Closest centroid;

Calculate the new mean for each cluster;

Until convergence criteria is met

III. RESULTS AND DISCUSSION

A. Data Set

Drug addiction exacts a staggering toll on affected individuals and society as a whole. Chronic drug abuse, craving, and relapse are thought to be linked to long-lasting changes in neural gene expression arising through transcriptional and chromatin-related mechanisms. The key contributions of midbrain dopamine (DA)-synthesizing neurons throughout the addiction process provide a compelling rationale for determining the drug-induced molecular changes that occur in these cells. Yet our understanding of these processes remains rudimentary. The postmortem human brain constitutes a unique resource that can be exploited to gain insights into the pathophysiology of complex disorders such as drug addiction. In this study, we analyzed the profiles of midbrain gene expression in chronic cocaine abusers and well-matched drug-free control subjects using microarray. A small number of genes exhibited robust differential expression; many of these are involved in the regulation of transcription, chromatin, or DA cell phenotype. Transcript abundances for approximately half of these differentially expressed genes were diagnostic for assigning subjects to the cocaine-abusing vs control cohort. Identification of a molecular signature associated with pathophysiological changes occurring in cocaine abusers' midbrains should contribute to the development of biomarkers and novel therapeutic targets for drug addiction.DNA microarray repository: Gene Expression Omnibus, from the National Institutes of Health: http://www.ncbi.nlm.nih.gov/geo/[4]. Dataset Title is Cocaine abuse effect on the midbrain The Summary of the Dataset is Analysis of postmortem midbrain specimens from individuals who died from cocaine abuse. Midbrain dopamine-synthesizing neurons play a important role in the addiction process. Results provide insight into the molecular pathophysiological changes in the midbrain associated with cocaine abuse. In this Dataset the total number of genes is: 48761. Organism: Homo sapiens

B. Evaluation Measure

Evaluation measure which used to compare the actual result and the predicted result analysis. In the evaluation we have to compare the results with Accuracy, Sensitivity, FPrate, Precision, Recall F-Measure, and Specificity. Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n. Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. That is, the accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. To make the context clear by the semantics, it is often referred to as the "Rand accuracy" or "Rand index". The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease. The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease. Performing multiple comparisons, the term false positive ratio, also known as the false alarm ratio, usually refers to the probability of falsely rejecting the null hypothesis for a particular test. F-measure combines the precision and recall concepts from information retrieval. We treat each cluster as if it were the result of a query and each class as if it were the desired set of documents for a query. Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined to make the context clear by the semantics; it is often referred to as the "Rand accuracy".

Table 1. Evaluation Measures

Metric	Formula
$Accuracy$	$\frac{TP+TN}{P+N}$
$Sensitivity$	$TPR = \frac{TP}{TP+TN}$
$FP\ rate$	$\frac{FP}{N}$
$Precicion$	$\frac{TP}{TP+FP}$
$Recall$	$\frac{TP}{P}$
$F\ measure$	$\frac{2}{\frac{1}{precision}+\frac{1}{recall}}$
$Specificity$	$\frac{TN}{FP+TN}$

In this Figure: 2 Shows that the comparison between the K-Means and the proposed Fruit fly K-Means. The Accuracy of the existing Algorithm is 0.35 and our proposed is 0.6, Sensitivity of existing algorithm is 0.4 and our proposed is 1. FPrate of existing algorithm is 0.7 and our proposed is 0.8. Precision of existing algorithm is 0.33 and our proposed is 0.55. F-Measure of existing algorithm is 0.3 and our proposed is 0.7.Sensitivity of existing algorithm is 0.4 and our proposed is 0.7. Specificity of existing algorithm is 0.3 and our proposed is 3. In this Comparison our proposed Fruit fly K-Means algorithm gives the high Accuracy While Compare to the Existing Algorithm, That mean our Algorithm is Better than the existing Algorithm.

Table 2. Evaluation results of actual and predicted classes

S.No	Runs	Accuracy	Sensitivity	FPrate	Precision	Recall	F-Measure	Specificity
1	KM_R1	0.3	0.3	0.7	0.3	0.3	0.3	0.3
2	KM_R2	0.2	0.2	0.8	0.2	0.2	0.2	0.2
3	KM_R3	0.45	0.3	0.4	0.4	0.3	0.3	0.6
4	KM_R4	0.2	0.2	0.8	0.2	0.2	0.2	0.2
5	KM_R5	0.6	1	0.8	0.5	1	0.7	0.2
6	KM_AVG	0.3	0.4	0.7	0.3	0.4	0.3	0.3
7	FFKM	0.6	1	0.8	0.5	1	0.7	0.2

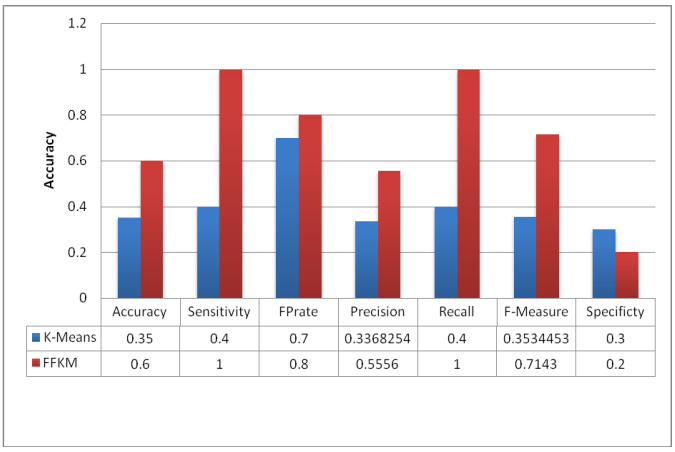

Figure 2. Evaluation Accuracy of Existing and Proposed Method

IV. CONCLUSION

In this paper we have proposed a simple method for selection of initial centroid to make K-Means more efficient and produce good quality clusters. In the K-Means algorithm cluster results are highly depends based on initial centroids. The proposed Fruit fly K-means algorithm has been compared with K-means algorithm by using cocaine gene expression dataset and observed that the proposed algorithm reduced the best quality of cluster.

V. ACKNOWLEDGMENT

The second author immensely acknowledges the UGC, New Delhi for partial financial assistance under UGC-SAP (DRS) Grant No. F.3-50/2011.

VI. REFERENCES

[1] W.-T. Pan, "A new fruit fly optimization algorithm: Taking the financial distress model as an example", Knowledge-Based Systems, vol. 26, (2012), pp. 69-74.

[2] W.-T. Pan, "fruit fly optimization algorithm", Taipei: Tsang Hai Book Publishing Co., (2011), pp. 10-12.

[3] E. N. Sathishkumar, K. Thangavel, T. Chandrasekhar" A New Hybrid K-Mean-Quick Reduct Algorithm for Gene Selection" International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol: 7, No:2, 2013.

[4] Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/

[5] K. Krishna and M. Murty (1999), 'Genetic K-Means Algorithm', IEEE Transactions on Systems,Man, and Cybernetics vol. 29, NO. 3, pp. 433-439.

[6] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown (2004), FGKA: A Fast Genetic K-means Clustering Algorithm', ACM 1-58113-812-1.

[7] J. A. Lozano J. M. Pena and P. Larranaga, \An empiri al comparison of four initialization methods for the k-means algorithm," Pattern Re ognition Letters, vol. 20, pp. 1027{1040, 1999