Indian Journal of Science and Technology, *Vol 8(22), DOI: 10.17485/ijst/2015/v8i22/79106, September 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Yoking of Algorithms for Effective Clustering

K. Mohana Prasad[1*] and R. Sabitha[2]

[1]Department of CSE, Sathyabama University, Chennai - 600119, Tamil Nadu, India; mohanaprasad1983@gmail.com
[2]Department of IT, Jeppiaar Engineering College, Chennai - 600119, Tamil Nadu, India; sabitha-ramadoss@yahoo.com

Abstract

Cluster plays a vital and very important in data mining. Cluster is a main and absolute part of real time applications. Grouping an object with its own class is known as Cluster. It has two different segments, Similar and Dissimilar objects. K Mean (KM) is one of the exclusive clustering algorithms. K Mean algorithm is introduced by cluster, which forms an easier and simpler way of classifying a given set of data. This paper is clearly based on Gravitational Search Algorithm (GSA) and KM algorithm. The main advantage of GSA and KM algorithm is to escape local optima and make convergence motions in rapid progression. A main five data sets in an UCI repository is used to bring the results and solutions in an excellent way using these algorithms. This paper aims to bring an exclusive and efficient result from both the algorithms compared to other algorithm and also gives perfect solution for the existing set of data.

Keywords: GSA, K Mean, UCI Repository

1. Introduction

Data mining is used to draw data out and knowledge by looking inwardly into a data structure. Cluster has an important role in Data mining[1-3]. The Cluster will form groups by critical action on data[3]. The algorithm differentiates the two different segments i.e. similar and dissimilar data. Unsupervised learning methods are used in cluster analysis when the data are not understood[10]. It has an absolutely necessary condition which is used for a data formation[5,14]. It is an inseparable part for medicine and biology. In management construction process the data clusterisation plays a vital role on clustering[2]. Marketing and analysis on customer satisfaction are important during a data structure process[12,15]. Too many other fields depend only on a Data clustering. An initial process is data to be structured and final process is based on clustering algorithm for a cluster formation[11]. Algorithm used for application and data types are unavoidable[4-7]. Algorithms are unable to fulfill the requirements. Generally cluster algorithms are differentiated accordingly into two types one is hierarchical and another one is partition algorithm[13]. Partition algorithm will partition the data when a structure for a hierarchical is fully obtained. Partition algorithm is completely based on center clustering method[8]. K-means clustering algorithm is more compatible and effective[11]. K-means clustering algorithm brings various inter cluster to a possible extent. For a past 2 decade a large amount of heuristic algorithm are proposed[9]. Heuristic algorithm is used to evaluate the cluster performance[5]. A main drawback in clusters is the inputs. The inputs should be given as prior as information to the application[6] . All the drawbacks in the cluster process are solved through the algorithms.

1.1 Demands of an Algorithm

Clustering Algorithm demands for a huge number of important test .They deals with some attributes like scalability, cluster discovery, arbitrary shapes, and some input parameters which are purely based on domain knowledge of data mining. It also deals with data outside a system that requires low amount of necessary noise, and also checks its arrangements.

1.2 Limitations of Algorithm

A countless limitations that are found in clustering are the algorithm doesn't cater on currently progress techniques.

**Author for correspondence*

The main drawbacks are the time compulsion during countless dimensions and various data process. There are huge numbers of time and energy consumption that occurs. On distance based algorithm the effectiveness is not very well defined. The distance is a powerful task during definition and sometimes distance is made to define but it is very tougher one. The result will be on various explanation and also some possible during correct time.

1.3 Classification of Clustering Algorithm

By the validity and formation of data structure, algorithm is classified into different kinds. K-means clustering algorithm is an exclusive clustering algorithm. A fuzzy c-means algorithm is an overlapping algorithm which is used to minimize the given data set; the hierarchical algorithm is a custom and mixing of some Gaussian algorithm. Gaussian algorithms are used in probabilistic clustering algorithm for a medical, training and other sector.

2. Background on Clustering and K-Means Algorithm

We can boldly discuss clustering from an example by calculating 8 out of 150 data based on students. (With (x,y)) via (age, marks of the students)

S1 (5,10), S2(6,8), S3(4,5), S4(7,10), S5(8,12), S6(10,9), S7(12,11), S8(4,6)

(I)Initially the center of clusters are S1 (5, 10), S4 (7, 10), S7 (12, 11)

Distance function of two points a = (X1, Y1) & b = (X2, Y2) is defined.

$$P(a,b)=\sqrt{(X_2-X_1)^2+(Y_2+Y_1)^2} \quad (1)$$

2.1 Progressions

Table 1. Iteration process

Roll no	Age	Marks	Dist mean 1 (5,10)	Dist mean 2 (7,10)	Dist mean 3 (12,11)	Cluster
S1	5	10	0	2	8	1
S2	6	8	3	3	9	1
S3	4	5	6	8	14	1
S4	7	10	2	0	6	2
S5	8	12	5	3	5	2
S6	10	9	6	4	4	3
S7	12	11	8	6	0	3
S8	4	6	5	7	13	1

Initial center cauterization means are (5, 10), (7, 10) and (12, 11) choose randomly.

2.1.1 Progression 1

We would calculate a distance function from a very first point (5, 10) to each three centroids by use of distance function process.

1.

FACTOR	MODE1
(x1,y1)	(x2,y2)
S1(5,10)	(5,10)

$$P=\sqrt{(X_2-X_1)^2+(Y_2+Y_1)^2}$$
$$=0 \quad (2)$$

2.

FACTOR	MODE2
(X1, y1)	(X2, y2)
S1 (5, 10)	(7, 10)

$$P=\sqrt{(X_2-X_1)^2+(Y_2+Y_1)^2}=0$$
$$=\sqrt{(7-5)^2+(10-10)^2}$$
$$=2 \quad (3)$$

3.

FACTOR	MODE3
(X1, y1)	(X2, y2)
S1 (5,10)	(12,11)

$$P=\sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$$
$$=8 \quad (4)$$

So that the point (5, 10) to be placed at the closest place for mean, i.e.

Mode 1(cluster 1) since the distance is 0

Cluster one Cluster two Cluster three (5, 10)

2.1.2 Progression 2

Next second point (6,8) i.e. each three mean should be progressed by distance function.

1.

Point	mean 1
(X1, y1)	(X2, y2)
S2 (6, 8)	(5, 10)

P = 3

2.

Point	mean 2
(X1, y1)	(X2, y2)
S2(6,8)	(7,10)

P = 3

3.

Point	mean 3
(x1,y1)	(x2,y2)
S2(6,8)	(72,11)

P = 9

On which cluster the point (6,8) should be placed?

The point have a shortest distance and mean1 (cluster 1) so distance value is 3.

Cluster one	Cluster two	Cluster three
(5,10)		
(6,8)		

So, rest of table should be filled by analogically,

Cluster one	Cluster two	Cluster three
(5,10)		
(6,8)		
(4,5)		
(4,6)		

At next we should recompute a new centers for a cluster (mean) for an cluster 1, we have (5+6+4+4)/4=> 4.75, (10+8+5+6)/4 = (7.25) For cluster 2 => (7+8)/2, (10+12)/2 = (7.5, 11) For an cluster 3 => (10+12)/2, (9+11)/2 = (11, 10)

New clusters: 1{S1, S2, S3, S8}, 2{S4, S5}, 3{S6, S7}

3. Experimental Setup and Evaluation

A K-means clustering algorithm performs a fast clustering on document and avoids catching from a local solution. K-means clustering algorithm and two other hybrid algorithms are to be on datasets of a different document. The datasets documents ranged from 204 to 800 and the terms are also ranged from over 5000 to 7000. It generates good results than all other clustering algorithms. It avoids the main drawback of cluster i.e. high consumption. A process generates a close clustering than either K-means clustering algorithm or hierarchical. By simulation the observation on Gravitational Search Algorithm (GSA) converges best optimum solution for a given data set in concurrence with result of genetic algorithm, which finds a best optimum solution corresponding to the given data set converges the best result.

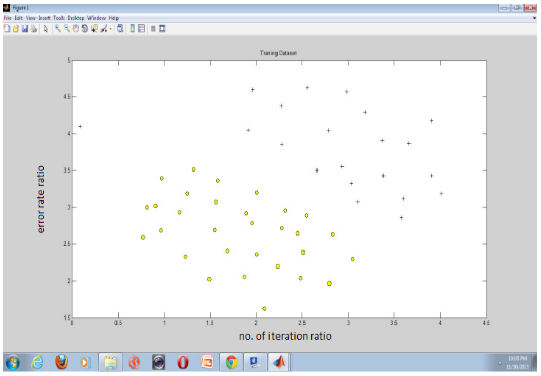

Figure 1. Training dataset.

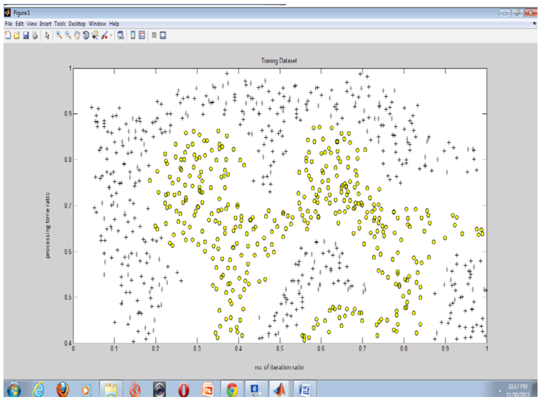

Figure 2. Training dataset.

K-means clustering algorithm is fast on searching while compared to other algorithm for a clustering process.

The non-linear and gradient methods converges a best suboptimal solution. Quality degradation for their approximate value to handle huge data sets and then it produces a forward implementation on parallel process based upon small data set.

4. Outcome

An order and relation between variables are measured using successions with ratio proportions. On using linear regression the coefficients are treated as same. It starts with the value negative 1 and it progress to a positive

- Capacity of positive 1 indicates a friendly relationship with various variables. Negative 1 represents a dissatisfaction for a variable. A Pearson coefficient on non-linear equation can't be on a higher level.

x=500/5=100; y=400/5=80

$$sx = \sqrt{2};\ \ sy = \sqrt{200}$$

$$\begin{aligned} r &= \left[\left[x - x/x\right]\left[y - y/y\right]\right]/n \\ &= \left(100/\sqrt{400}\right)/5 \\ &= \left(100/20\right)/5 = 5/5 = 1 \end{aligned} \quad (5)$$

Table 2. Pearson's coefficient

Roll No	Mark 1	Mark 2	$\left[x-\bar{x}/\sigma x\right]$	$\left[y-\bar{y}/\sigma y\right]$	$\left[x-\bar{x}/\sigma x\right]\left[y-\bar{y}/\sigma y\right]$
A	102	100	$102-80/\sqrt{2}=2/\sqrt{2}$	$100-80/\sqrt{200}=20/\sqrt{200}$	$40/\sqrt{400}$
B	101	90	$101-100/\sqrt{2}=1/\sqrt{2}$	$90-80/\sqrt{200}=20/\sqrt{200}$	$10/\sqrt{400}$
C	100	80	$101-100/\sqrt{2}=0/\sqrt{2}$	$80-80/\sqrt{200}=0/\sqrt{200}$	$0/\sqrt{400}$
D	99	70	$99-100/\sqrt{2}=-1/\sqrt{2}$	$70-80/\sqrt{200}=-10/\sqrt{200}$	$10/\sqrt{400}$
E	98	60	$98-100/\sqrt{2}=2/\sqrt{2}$	$60-80/\sqrt{200}=-20/\sqrt{200}$	$40/\sqrt{400}$
TOTAL	500	400	0	0	$100/\sqrt{400}$

5. Conclusion and Future Work

From a divided work we discuss the various factors which affect a clustering algorithm such as significance level, distance, selection of a different data, and different similarities on measures. A probe on k-means clustering algorithm and their steps are used, so the reduction of time complexity occurs and accuracy of data also evolved. The more effective data is made by the response that will make research related to a marketing process and other etc…. and there is a very strong formation used to reduce an error rate in very successful way.

6. References

1. Mohana Prasad K, Sabitha R. Evolution of an Algoritm for Formulating Efficient Clusters to Eliminate Limitations. International Journal of Applied Engineering Research (IJAER). 2014; 9(23):20111–8.
2. Akay B, Karaboga D. A modified artificial bee colony algorithm for real-parameter optimization. Information Sciences. 2012 Jun; 192(1):120–42.
3. Babu G, Murty M. A near-optimal initial seed value selection for K-means algorithm using genetic algorithm. Pattern Recognition Letters. 1993 Oct; 14(10):763–9.
4. Bagirov. Modified global K-means algorithm for sum-of-squares clustering problem. Pattern Recognition. 2008 Oct; 41(10):3192–9.
5. Che Z, Unler A. Clustering and selecting suppliers based on simulated annealing algorithms. Computers and Mathematics with Applications. 2012 Jan; 63(1):228–38.
6. Jimenez J, Mares J, Torra V. An evolutionary approach to enhance data privacy. Soft Computing. 2011 Jul; 15(7):1301–11.
7. Li T, Li N, Zhang J, Molloy I. Slicing: a new approach for privacy preserving data publishing. IEEE Transactions on Knowledge and Data Engineering. 2012; 24 (3):561–74.
8. Mogre NV, Agarwal G, Patil P. A review on data optimization technique for Data publishing. International Journal of Engineering Research and Technology. 2012 Dec; 1(10).
9. Hsu CC, Chen YC. Mining of mixed data with application to catalog marketing. Expert Syst Appl. 2007 Jan; 32(1):12–27.
10. Chen W, Feng G. Spectral clustering: A semi-supervised approach. Neuro-Computing. 2012; 77(1):229–42.
11. Li P, Chen C, Bu J. Clustering analysis using manifold kernel concept factorization. Neurocomputing. 2012; 87(15):120–31.
12. Ahmadi, Karray F, Kamel MS. Model order selection for multiple cooperative swarms clustering using stability analysis. Information Sciences. 2012 Jan; 182(1):169–83.
13. Che Z, Unler A. Clustering and selecting suppliers based on simulated annealing algorithms. Computers and Mathematics with Applications. 2012; 63(1):228–38.
14. He H, Tan Y. A two-stage genetic algorithm for automatic clustering. Neurocomputing. 2012; 81(1):49–59.
15. Li P, Chen C, Bu J. Clustering analysis using manifold kernel concept factorization. Neuro computing. 2012; 87(15):120–31.