Analysis and Optimization of Data Classification using K-Means Clustering and Affinity Propagation Technique

Srinivas Vadali[1], G.V.S.R. Deekshitulu[2] and J.V.R. Murthy[3]

[1]Research Scholar, Department of CSE, Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, India

[2]Department of Mathematics, University College of Engineering Kakinada (UCEK), Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, India

[2] Department of CSE, University College of Engineering Kakinada (UCEK), Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, India

Abstract

Twitter provides an enormous platform to perform analysis on data consisting of events, trends, personalities. It enables to determine the inclination, likes of the people in real time independent of size. There several techniques to retrieve the data and the most efficient technique to retrieve the data is the clustering technique. There are many approaches in clustering to group and analyze the data. This paper provides an overview on various algorithms and their effectiveness in determining the trending pulses efficiently. Once the data are clustered, they could be classified based on the topics for real time analysis on the huge collection of data set which is very dynamic. In this paper classification of data is performed and analyzed to determine the flaws. The classification is again performed on the same dataset using an optimized technique and analysis is performed on the clustering of the data.

Keywords: K-means Clustering, Affinity Propagation Technique, Centroid, Euclidian Distance

1. INTRODUCTION

There are some set of data which could be available. These data sets may not be readily classified or labeled. Such data could be obtained from newspapers, magazines, blogs, twitter, etc. The texts obtained from such wide range of open sources would be raw and unformatted. The amount of data obtained would range from seconds, minutes, hours, day, months, years and so on which lead to storage of millions of racks of data. To store the data and efficiently would require some asserted way to sort the data and arrange the data to determine the type of data. Data storage is very critical as it is required to be retrieved some point of time later efficiently. It is not only the amount of time that matters but also the type of data retrieved would also be essential. The right data retrieval would be the order of the day as wrong and unwarranted data retrieved would mean wastage of time and energy even with quickest time of data retrieval.

The K-Means clustering[1] provides a quick and reliable means to classify the streaming data into several groups based on the features of the data available. The most important aspect of K-Means clustering method is that the technique can be applied to raw and unformatted data which could be grouped into multiple groups called as clusters. The number of clusters would be initially being determined based on the number of divisions required by the problem definition. The number of cluster groups obtained would be the number equivalent to types of datasets that correspond to the subject of interest. The unsupervised learning requires no prior processing of data with metadata tags such as labels. The data that would be input is raw and stream from the user source. The basic preprocessing of data could be done remove unnecessary whitespaces and irrelevant data patterns.

The Centroid[2] is randomly placed somewhere inside the group. The Centroid is the point of concentration of the data group. Every input sentence is represented as a data point associated with certain mathematical value calculated using vectorization[3] techniques term frequency, inverse document frequency and counting to determine the vector form of a given data text. The vector would be the dot product representation of scalar quantities such as the mathematical value of the data text, the parameter value considered if any during the process of classification.

Every data point is a vector and each vector can be defined as the set of features. The distance between each of the vectors and the Centroid of each cluster is calculated using technique such as Euclidean algorithm[4]. The distance determined between the Centroid and each data point is measured in terms of Euclidean distance[5]. The Euclidean distance calculated to determine the nearest Centroid for each of the data point or vector. The vector with shortest Euclidean distance to the Centroid of each group is assigned to corresponding cluster containing the Centroid. Thus the Euclidean distance plays a major role in assigning the data points to individual cluster groups and hence classification of datasets.

The Twitter is the world's largest social platform in the world to update the current affairs, events, news, beliefs and perspectives as tweets. The updates from the events are obtained in real time which is made through the series of re-tweeting of the tweet posted through the original handle. The Twitter allows its users to follow and un-follow their favorite personalities, friends, media through a twitter handle. The popularity of smart-phones due to its portable size and affordable prices has increased the social media users especially the twitter by enormous amount. Any event occurring in any corner of the world is updated on the twitter instantaneously. It has been said the news media channels break the news first on twitter through its twitter handle and subsequently take them on to the television channels.

The best thing about the platform is that it exposes APIs to retrieve the data based on keywords. There are third party applications which expose the twitter data in different custom approaches using OAuth(Open Authentication) mechanisms. The tweets with the matching keyword would be retrieved seamlessly in a size and time independent manner. The number of tweets retrieved could more or less depending on the number of times the searched keyword, the topic is tweeted or re-tweeted. Generally the topics would be tweeted as hash-tags, i.e. the words beginning with the symbol '#' or uppercase letters in case of personalities. If the personality has a twitter profile then that person would be addressed beginning with '@' symbol. These conventions throw away a lot of options for data mining capabilities. It becomes relatively easy for data miners to find all the data abundantly in the world on a single platform.

There are several approaches to retrieve the content, process it into information in required format and then analyze the information for marketing, sentiment determination, politics, trends, etc. The most effective technique is by clustering the data and then classifying the data to determine the trending topics across the world. The number of data sentences to be required to perform classification is randomly determined and is dependent on the performance of the computing machine. If millions of data text are to be classified it is necessary to have good computing machine. In the current project 8, 22 and 50 sentences are used to perform classification and analyze its result. The number of sentences could well be extended to many hundreds based on the computing power of the machine.

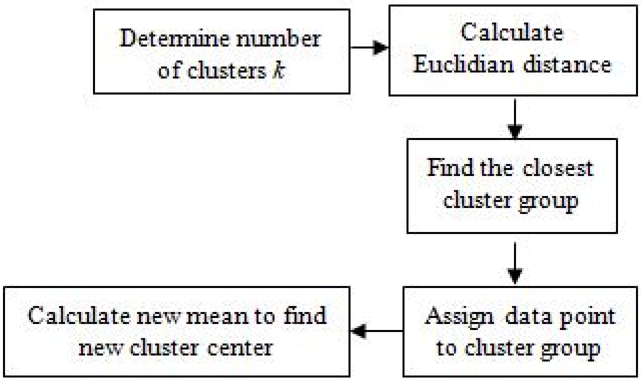
Determine number of clusters k
Calculate Euclidian distance
Find the closest cluster group
Assign data point to cluster group
Calculate new mean to find new cluster center

Fig1: K-Means Clustering Flow Diagram

2. IMPLEMENTATION

The tweets retrieved would be highly unstructured and informal. The data obtained would be ineligible for performing processing on it. This is because of all the metadata that is associated with the tweet. It is required to identify the essential words in the tweet content and removal of unnecessary text such as proper nouns beginning with '@' symbol followed by the twitter handle user name. The other invaluable content would include digits, whitespaces, smileys, and stop-words which do not have any information and are used for imparting meaning into the data. These texts have to be removed. There have to spell checks in place to correct in case of spelling errors found in the content. It also would be a good convention to convert all the upper case into lower case letters so as to decrease the time required during processing and reduce redundant checks. The normalization[6] is not mandatory step to perform during processing stage but it is an internationally accepted approach to follow before data analysis. This step also reduces the randomness coefficient associated with the data. It also decreases the error percentage gradually in the result during each of the iteration.

Now there are essential words in the tweet that undergo the processing after normalization. These words are bifurcated into nodes. Multiple nodes with similar meaning are group into same cluster. Words with different meaning will not be assigned into the same group. The relative distance between the words with different meaning or words referring to different context would be high. Words referring to similar context would be placed with minimal separation. The clustering would occur in a manner that given any words from different cluster groups would definitely be imparting different meanings or might be referring to a totally different context altogether. The intra-cluster distance[7] refers to distance between the words in the same cluster which should be always less. The inter-cluster distance[8] refers to distance between different clusters. This distance should be greater than the mean highest distance between the words within the same cluster. This is the most important property to determine the efficiency of a clustering process. There would be occurrences where in some of the text data might get assigned to a irrelevant cluster group. Over the iterations the mean center in a cluster would be nominated in such a way that all the relevant text data would be surrounding the mean center in the cluster. In this way the error rate would be lower than that found using keyword based search technique. Cluster homogeneity[9] computation determines the proximity rate or the interchangeable ability among the words in the cluster. Higher the rate for the cluster, easier it is to substitute the word for another belonging to the same cluster and vice versa. This is because of the alikeness between the words. It is measured as a function of distance between the words. There are numerous techniques to measure these distances such as Euclidean distance, Manhattan distance[10], Chebychev distance[11], Spearmandistance[12], Euclidean square distance[13], Pearson square distance[14] and Pearson correlation distance[15]. All these measure the distance between the text data within each of the cluster. The most widely and frequently used technique is the Euclidean distance method due to its ability to be able to easily be applicable on most of the given real world data set objects and technical simplicity. In case of inter-cluster distance the approaches such as single link, average and complete link are used. The single link denotes the distance between the most adjoining data points across the cluster. The average link denotes the mean distances of all the data points between the clusters. The complete link is the exact opposite to the single link distance measuring technique. It is the distance between two most significantly farthest located data points in the two cluster groups.

There are many clustering techniques such as K-Means clustering, Agglomerative Hierarchical clustering[16] and Jarvis-Patrick clustering[17] which could be applied onto data points in the cluster obtained from the twitter data. There is also Naives-Bayesian[18] technique which is a popular technique for classification of data. But the clustering method offers more flexibility and extensibility to the data points in the cluster sets.

The K-Means clustering is numerical based grouping of data points. The number of clusters is always fixed. It should be ‘K’. There would be at-least one data item in each of the cluster group. The cluster groups do not converge in K-Means clustering. In other words there is no data which is belonging to more than one cluster group at a given space and time. The main advantage of K-Means clustering is that since it uses numerical values on data content performs mathematical deductions it is more fast and efficient. The loop hole in this technique is the ambiguity of choosing number of clusters ‘K’ to initiate the clustering process. It is tedious to determine the standard and features of the cluster groups produced in K-Means clustering.

The Jarvis-Patrick clustering is based on the alikeness of the data content points determined by the distance measure between the content objects. It is a natural algorithm which does not involve lot of mathematical computations unlike K-Means clustering algorithm. It is also deterministic in the number of clusters and it is not required to provide the number of clusters as input at the beginning.

The Agglomerative Hierarchical clustering is similar to Jarvis Patrick clustering as it is also a natural clustering based on the gene merging process in living beings. It starts with each data point assigned with a separate cluster and then grouped into same clusters when conducive data partner points are found. Here the number of clusters produced finally is very less number and hence the data content look up is very easy. The disadvantage is that once a data point is merged with a particular cluster, there is no re-grouping or re-clustering in case of data point being merged with wrong cluster group.

The following code snippet transforms WordCorpus into a Term Document Frequency matrix which is used to perform classification using K-means clustering technique. The distances between datapoints are calculated using Euclidean distance algorithm.

```
tdm <- TermDocumentMatrix(wordcorpus)

tdm <- as.matrix(tdm)

distMatrix <- dist(tdm, method="euclidean")
```

3. CASE BASED ANALYIS

Test Case 1:
Consider the following eight sentences with k=3for the word accident:

[1] "TheCitizen_in: There is nothing accidental about it, these threats, and horrendous efforts to intimidate artists.",

[2] "I spend half of my sleepless night trying to find the sleeping position from my accidental afternoon nap. #SleeplessNights",

[3] "Campfire safety is key to preventing injuries and forest fires. Keep these #campfire #safety tips in mind during ",

[4] "rhettandlink talk about their visit to Aussie and accidental HCFC cameo!\n\nOriginally shared by @Bluzae ",

[5] "The accidental hilarity of the self own is just too perfect.\nCohen worked for Trump for over a decade.",

[6] "Congressman Louis T. MacFadden, Chairman of the House Banking &amp; Currency Committee: It was not accidental",

[7] "There is Nothing Accidental About These Threats, It Is a Pattern: T.M.Krishna",

[8] "I met her accidentally and it was all fun"

Table1: Clustering 8 sentences with k=3

Cluster #	Sentence #
1	3
2	1,2,4,5,6,7
3	8

Test Case 2:

By increasing the sample sentences to 50, with k=5 for the word *accident*:

Table2: Clustering 50 sentences with k=5

Cluster #	Sentence #
1	49,50
2	1,7,23,25,31,,32,37,39,45
3	48
4	43
5	2,3,4,5,6,8,9,10,11,12,13,14,15,16,17, 18,19,20,21,22,24,26,27,28,29,30,33, 34,35,36,38,40,41,42,44,46,47

Test Case 3:

Similarly, by performing clustering to sentences containing word *collapse* with k =4, we have:

Table3: Clustering 22 sentences with k=4

Cluster #	Sentence #
1	1,2,3,4,5,6,8,9,10,11,12,13,14,15,16,17,18
2	19,20
3	21,22
4	7

4. Results

The word *accident*, which is used in various contexts, has resulted in unexpected cluster formation. The sentences 1,7, 23, 25, 31, 32, 37, 39, 45 are classified under cluster 2. The intention of these sentences though with the usage of the term *accident* does not have any physical damage to the subject of the sentences. Most of these sentences included in cluster group 2 intend to be philosophical in nature. The presence of phrases such as *crossing of paths*, *ending of life*, *catching up with people* alongside the word *accident* changes the sense of the sentence from being something physically harmful to be potentially philosophical. The cluster group 2 could be labeled as the group in which sentences tend to be philosophical. In sentences 1 and 7 even with occurrence of word *threat* in the sentences alongside *not*, negate the context of the sentence which justifies the inclusion of these both sentences in cluster group 2.

The sentence 48 is in separate cluster group of 4. The context intends to convey about *accident* leading to extreme case of tragedy like death.

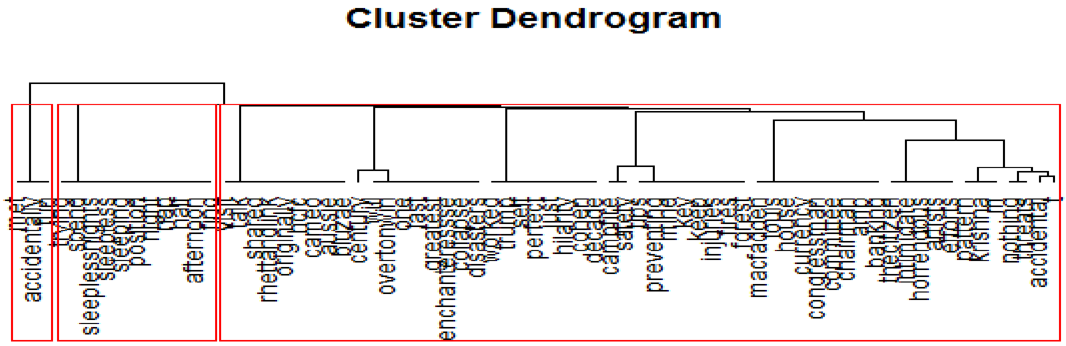

Figure 2: Clustering 8 sentences with 3 cluster groups containing word ‘accident’

The sentence grouped under cluster 5 intends that there is some physical harm involved in the presence of term *accident*. The accident in normal concourse refers to getting harmed physically though unintentionally. The occurrence of phrases such as *sleepless nights, injuries, hurt* refers to getting harmed physically.

The sentences 49 and 50 with intention of meeting with some object are grouped under cluster 1. The cluster 1 could be labeled for *meeting*. Although the sentence 48 which tells about meeting with an object. It also tries to be philosophical which made it to get included in separate cluster group.

It although cannot be said that the classification has occurred perfectly as there are sentences which are not intended to physical harming yet included under cluster 5. The sentences which are ranging from 4 to 6 and 8 to 13 shows that they are not harmful but yet grouped under cluster 5.

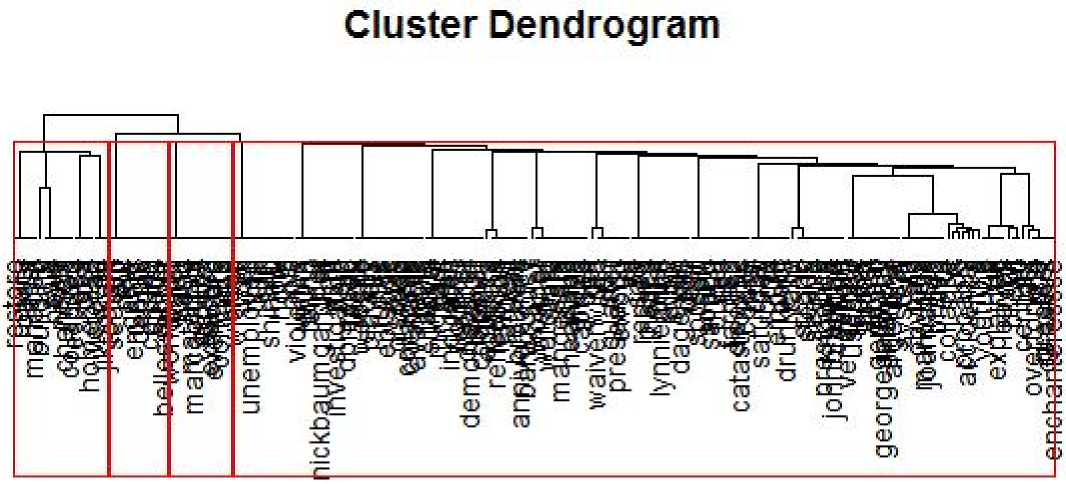

Figure 3: Clustering 22 sentences with 4 cluster groups containing word ‘collapse’

Overall, the classification provides a brief overview of the nature of the context of the sentences but classification cannot be heavily depended to determine the exact sense of the sentence. Hence the classification could be used as the first step during processing of the sentences.

For the sentences involving word *collapse*, the sentences involving the word *collapse* are classified into 4 cluster groups. Majority of sentences are grouped under first cluster which involve - suddenly, to fall down, to give way or to subside. All these sentences have negative intent of going down and seem to be grouped together. The 7th sentence (the occurrence of the word *collapse* alongside the word *rescue*) provides a positive effect to the sentence and grouped in the separate fourth cluster. The 21stand 22nd sentence tells about the intent of something that never went wrong with occurrence of the words ‘not’ and ‘never’ respectively. The 20th sentence is classified into separate second cluster group.

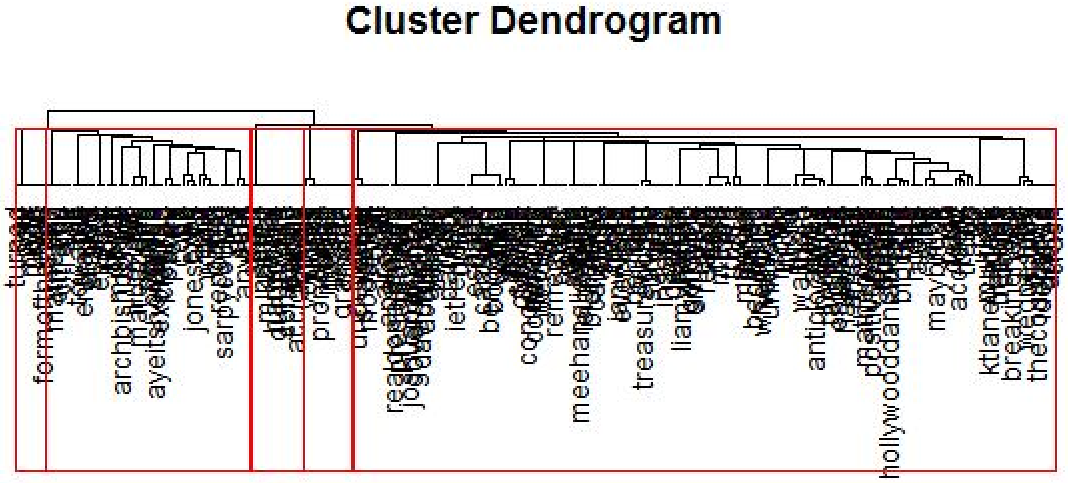

Figure 4: Clustering 50 sentences with 5 cluster groups containing word *accident*

The words such as *not*, *never*, *about*, and *rescue* occurring along with the word *collapse* which intends impart a very negative emotion, such words tends to combat the negative intent and turns the sentence to nullify the negative effect or brings in the positive effect to such sentences.

From the classification performed above it can be inferred that the emotions which has occurred, which are about to occur and which are bound to occur or which never occur can be classified into separate cluster groups.

1. Optimization Technique

In the K-means clustering technique it is required to choose the number of Centroids initially. The number could be chosen suitably provided the user has good knowledge about the application domain else the number of clusters obtained as a result would be too many or too less which might not satisfy the problem definition. If the number of clusters formed is high then it would consume more time iterating the clusters formed and hence efficiency would decrease gradually.

Affinity Propagation[19] is an algorithm that which identifies the exemplars[20] among the datapoints. The number of clusters need not be chosen by the end user. The algorithm by itself has the potential to identify the number of clusters to be formed. The exemplar would be the mean value of the cluster but not exactly the Centroid as obtained in K-means clustering. The exemplar is not a space value instead it is also a datapoint. Initially each of the datapoints stake claim to be the exemplar of the cluster group. The messages are exchanged between the datapoints until good set of exemplars and clusters are formed. There are two kinds of messages are passed across the datapoints.

1. The messages are sent from datapoints to exemplars called responsibility *r(i,k)* where *i* is the datapoint and *k* is the exemplar. This indicates how well the datapoint is suited to be the member of the exemplar's cluster.
2. The message sent from exemplar to datapoint is called availability *a(i,k).* This indicates how appropriate *k* would be an exemplar to the datapoint *i*.

Summing up the responsibility and availability gives us the clustering information for a given datapoint called the preference value *p(i,k)* where *i* is the datapoint and the *k* is the exemplar.

p(i,k) = r(i,k) + a(i,k) is the preference value of exemplar *k* for i^{th}datapoint. The higher the preference value of exemplar *k* to the datapoint *i*, higher is the chance of *k* to become the exemplar for the cluster group provided other datapoint's preference values be lower than the k^{th}preference value.

If the numbers of clusters are not optimal and the user wants to change it one can do so using the preference values. By adjusting the preference values we can lower or raise the number of clusters from the given set of datapoints. The higher is the preference values, the higher is the number of cluster groups formed as each point is more certain to be an exemplar and form a cluster. The lower the preference value, the lower the number of cluster groups formed as it would prefer to join another cluster with high preference value.

The following code snippet shows a Term Document Frequency matrix which is used to perform classification using Affinity Propagation technique.

```
tdm <- as.matrix(tdm)
```

Table4: Classifying 8 sentences using Affinity Propagation technique resulting in 3 clusters

Cluster #	Sentence #
1	3
2	1,2,4,5,6,7
3	8

The classification performed by Affinity Propagation algorithm on eight sentences with context of word ‘accident’ without specifying the number of clusters during initialization produced the result of three clusters with third and eight sentence in the different cluster groups. If the result is analyzed closely, the result obtained is similar to that obtained using K-means clustering technique using three clusters. There seems to be no change in the result obtained using the Affinity Propagation technique and the K-means clustering algorithm for the eight sentences for the context word ‘accident’.

Table 5: Classifying 22 sentences using Affinity Propagation technique resulting in 6 clusters

Cluster #	Sentence #
1	1
2	7
3	2,3,4,5,6,8,9,10,11,12,13,14,15,16,17,18,19
4	20
5	21
6	22

The classification for twenty two sentences involving the context word *collapse* using Affinity propagation technique produced six clusters with sentences 1, 7, 20, 21 and 22 forming a separate cluster group in other words as distinct exemplars. The K-means clustering technique was performed with initial number of clusters initialized to four. The results obtained using K-means clustering and Affinity seems to match as the sentences 7, 20, 21 and 22 are part of separate cluster group.

Table 6: Classifying 50 sentences using Affinity Propagation technique resulting in 3 clusters

Cluster #	Sentence #
1	1,3,4,5,6,7,8,9,11,12,13,14,15,17,18,20,21,22,23,24,25,27,28,30,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,50
2	2,10,16,19,26,29,31,49
3	32,33

The classification for the fifty sentences involving the context word *accident* using Affinity Propagation technique yields a totally different result than that obtained using the K-means clustering using five different cluster groups. The Affinity propagation technique produces only three different cluster groups.

Conclusion

In K-means clustering the Centroids are picked initially and provided if the choice of the Centroids is close to good solution then it works the best. It is suitable if the clusters are small otherwise it will consume time iterating through every data item in each of the cluster group. The initial number of clusters needs to be specified for which the Centroids are initialized randomly. The Centroid may be either the data value or the space value which the end user has no control over the Centroid scalar value. There is also the possibility of too many clusters more than the required number leading to decrease in the specification of outrage values.

The Affinity Propagation technique helps in identification of exemplars among the data points and forms the clusters of data points around the exemplars. The exemplar is a data value but not a space value. The cluster group number specification would be taken care by the algorithm using the preference value to distinctly specify the outrage values for a given set of data points by determining the number of cluster groups intrinsically.

References

[1] Quing Yang, Ye Liu, Dongxu Zhang, Chang Liu, Improved k-means algorithm to quickly locate optimum initial clustering number K, Proceedings of the 30th Chinese Control Conference, 2015

[2] Md. Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar, Improvement of K-means clustering algorithm with better centroids based on weighted average, 7th International Conference on Electrical and Computer Engineering, 2016.

[3] Farid Bourananni, Mouhcine Guennoun, Ying Zhu, Clustering Relational Database Entities Using K-means, Second International Conference on Advances in Databases, Knowledge, and Data Applications, 2015.

[4] H.H. Crokell, “Specialization and International Competitiveness,” in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)

[5] K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, “A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II,” KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)

[6] J. Geralds, "Sega Ends Production of Dreamcast," vnunet.com, para. 2, Jan. 31, 2001. [Online]. Available: http://nl1.vnunet.com/news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)

AUTHOR

Vadali Srinivas received M.Sc., degree in Electronics from Andhra Univeristy in 2006 and M.Tech., in Computer Science and Technology from GITAM University (Deemed to be University), Visakhapatnam in 2009. Currently, he is working towards Ph.D. degree in Computer Science and Engineering at **JNTUK (Jawaharlal Nehru Technological University)Kakinada,** Kakinada , India. Over 10 years of teaching experience and fields of interest in Optimization Techniques, Mission Learning and Data warehousing and Mining.

Deekshitulu GVSR received M.Sc., in Applied Mathematics from the Andhra University, Visakhapatnam, India, in 1994 and Ph.D. degree in Mathematics with Specialisation as Differential and Integro differential equation on measure chains from the Andhra University, Visakhapatnam, India, in 1998. In the capacity of Junior Research Fellow done Department of Science and Technology (DST) Project with 9.5 Lakhs from 1995-1997. Currently, he is Professor of Mathematics JNTU College of Engineering, Kakinada. 20 years of experience in teaching engineering graduates and post graduates. He has guided 02 Ph.D., students in Mathematics and guiding 06 Ph.D., in Mathematics and Computer Science . He published many papers in various national and international journals and conferences. His current research interest include Boundary value problems, Integro differential equations on time scales, Fractional differential equations, Fractional difference equations, started working on difference equations involving causal maps.

J.V.R Murthy has received his PhD degree in Computer Science and Engineering from JNTU College of Engineering, Kakinada in 2004. He completed his M.Tech from IIT Khargpur in 1990 and degree of Bachelor of Engineering from JNTU College of Engineering, Kakinada in in 1982. He is currently working as Professor in the Department of Computer Science and Engineering, JNTU College of Engineering, Kakinada, A.P., INDIA. Over 24 years of Teaching, Research and Industrial experience in the field of Computer Science with specialization in Data warehousing and Mining. Started the career as computer programmer and occupied various positions such as lecturer, Assistant professor and professor. More than three years of Industrial Experience in USA as Senior People Soft Consultant (Techno functional) with reputed companies such as William M Mercer, Key Span Energy and AXA Client solutions.