IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

3 WAY SHIELD ENHANCED K-MEAN NORMALISATION BY INTELLIGENT INIITIALISATION OF CENTROID

Swati Nenava*, Prof. Manoj Chouhan

[*] Information of technology department,Shri Vaishnav institute of technology and science,India
Information of technology department,Shri Vaishnav institute of technology and science,India

ABSTRACT

"Learning to learn" is a notion of full excitement in advance research within machine learning .K-means is a most widely used approach in unsupervised machine learning algorithms though it was proposed 50 years ago,but undoubtedly is usually very fast,but its efficiency is highly depends on the placement of initial centroids.Thus,the proposed study is based on previously made the effort and improvements over traditional K-mean clustering scheme by introducing the impact of one the best initialization method named as PCA -PART,inputting normalized linear transform geometric data-sets over 2D hyper plane.This amalgamation in k-means clustering not only give provision for noise found to be the most issue factor in centroid based clustering but also improve efficiency of k-means by airing the data-set and removing hurdles leading to fine-grained initial centroids.

KEYWORDS: K-mean clustering,initialization methods,preprocessing steps,transformations.

INTRODUCTION

Cluster analysis is everywhere in any discipline that involves analysis of multivariate data. A Google Scholar search found 1,660 entries with the words data clustering ,appeared in 2007 alone. This immense literature display enormous importance of clustering in data analysis.Though listing the numerous scientific fields and applications that have utilized clustering techniques as well as the thousands of published algorithms is difficult to handle .An Image segmentation is an significant clustering problem hierarchies for efficient information access, the most outstanding technique is Documents clustering .

Clustering have various usage from grouping customers into different types for efficient marketing to grouping services delivery engagements for management and planning also to study genome data in biology.with the consequences that the basic problems and method of clustering become well known in a broad scientific community,in statistics,data analysis and in particularly in applications.

Data mining process involves cluster analysis as a unsupervised cluster learning algorithm for knowledge discovery,used in various applications.When a user input data ,it needs to to mined in order make it useful for a particular purposes. Inputted data is processed into a no. of data mining or KDD process preprocessing,which includes various methodologies in order to convert raw data into useful forms since the data sets is very complex and with the high dimensional space it becomes difficult to map the data and finding the hidden patterns,after preprocessing step various techniques and algorithm role take place.This includes learning algorithms,which are of two types supervised and unsupervised ,supervised is also called as clustering and unsupervised are clustering[9] Clustering and classification concept have its own importance and impact on different applications.In classification the training and testing set are given ,on the basis of it simulation done of new data set deciding which category it belongs.Since in classification the dataset is previously categorized in the training data sets and this dataset is trained on which it is categorized.Thus in classification the characteristics of similarity of data is already fixed or known thus this data set is trained algorithm decide the category of new data set.whereas in clustering the characteristics of similarity is unknown and the training data set is used but doesn’t require training since the process of algorithm is finding hidden patterns the similarity in the basis of characteristics /attribute value is unknown.Educational data mining is a research field which has the potential to lower the barriers from the mining process from the useful educational data and methods emerge increases the prediction and discovery with models .[7]

REVIEW OF RECENT WORK

Ahamed Shafeeq b.m et.al [2]find the procedure in order to find the optimal number of clusters,by examining the problems in k means clustering algorithm related to fixing number of clusters.It is found that on fixing the number clusters,its not necessary to get quality clusters in traditional k-means,therefore they concerned about finding a ways when clusters are not fixed on prior,the algorithm works for two cases first when the cluster inputted from the user is known that is when it is fixed and secondly, when it is unknown.Usually for finding cluster rank level validity threshold for fixing the cluster,the least number of clusters is taken as a input from a user,when the traditional algorithm is run and repeated, expanding the number of clusters at each iteration by 1, and at a point when validity threshold is calculated by finding the number of clusters on run time,the algorithm find the no. Of cluster and fixed.The proposed work is entirely focuses on improving data clustering for unknown hidden patterns in data set,when the number of cluster is not fixed in traditional k-means.Finding a fact that, the number of clusters when small, there is chance of forming a clusters with more dissimilar objects in the same group and on the same similar objects in different group when the number of clusters are large.

Fan .Cai et.al[3] evaluate different clustering algorithms for analyzing different financial datasets varied from time series to transactions and discuss the pros and cons of each method to raises the understanding of inner structure of financial datasets and other large size database as well as the capability of each clustering method in this context. Show how that density-based clustering does not suit financial dataset. Normalised centroid-based clustering with higher DI or lower DBI gives the best number of clusters to help understanding financial data classification. Original attribute scales do not reflect the behaviour similarity since Euclidean distance is controlled by large scaled attributes, best average tightness does not indicate the best case by departing the worst case.Some constrains are found e.g., K-means clustering tends to find spherical clusters, centroid-based clustering does not handle the noise, etc.

Madhu Yedla et.al[6] shows a new method is proposed for finding the better initial centroids and to give an effective way of assigning the data points to the clusters which are suitable with reduced time complexity. k-means algorithm is one of the often used clustering

method in data mining, its because of the staging in clustering huge data sets. The result of the final cluster in k-means algorithm substantially rely on the accuracy of the initial centroids,which are selected randomly.More importantly original k-means algorithm converges to local minimum, not the global optimum.

M. Emre Celebi et.al [4]present an overview of initialisation methods with an prominence on their computational efficiency and differentiate eight commonly used linear time complexity initialization methods and vast accumulation of data sets using various performance criteria, analyzing the experimental results of using non-parametric statistical tests and also provide recommendations for practitioners by demonstrating popular initialization methods.They also proved that traditional initialisation methods often perform poorly and in fact there are strong substitutes to these methods.As preprocessing steps is the one of the important step in finding good clusters by identifying structures,attribute normalization is highly preferable preprocessing step which can give overwhelming results. Deterministic methods,generally outperform the non-deterministic ones,this fact evolve because non-deterministic methods can produce highly variable results during multiple runs. In time-critical applications when considering large dataset where determinism is important ,methods V or P is preferred. These methods have a very good impact in a way that executed only once and resulting in very fast k-means converge.

anil k jain[5] concluded the summons in data clustering is to assimilate domain knowledge in the algorithm, finding meaningful representation and finding similarity measure,calculating accuracy of cluster,planning a rational basis for comparing methods,and develop efficient algorithms for clustering large data sets.They explain overview of clustering,explained well known clustering methods,and discusses the major dispute also key issues in designing clustering algorithms, and pinpoint about various methodologies,such as ensemble clustering, simultaneous feature selection, data clustering,semi-supervised clustering and large scale data clustering

Fahim A.M. Et.al[8]benefit from previous iteration of k-means algorithm, For each data point, they can keep the distance to the nearest cluster. At the next iteration, and calculate the reachability to the previous nearest cluster.Through this the time required to compute distances to k−1 cluster centers is saved.This concept makes the center closer to some points and distant apart from the other points, whatever points get closer to the center it will stay in that particular cluster,making no necessity to find its distances to other cluster centers. The points far apart from the center may change the cluster, so only for these points their reachability to other cluster centers are determined, and reallocate to the nearest center

TRADITIONAL K-MEANS CLUSTERING

K-means clustering algorithm is very popular clustering technique used in various applications like in pattern recognization and document clustering,in this approach clusters are represented by mean(weighted average).It has a good sense for numerical attributes with a sense of statistics and geometrical way but not good in categorical way. In order to make clusters of the same type,the difference between the point and centroid is calculated through distance function which is represent as a objective function which has to be converge,that is the at the point when difference is found to be null the algorithm stops.[17]The sum of variations between point and centroid is a concern point.

Objective function based on the L2-norm,the total intra cluster is identical to SSE errors between points and corresponding cluster.With the logical reason SSE can be justify as a negative of log probability of normal distribution model used in statistics.Thus k means algorithm concept relate with the probabilistic distribution function,An objective function based on L2 norm ,normalised individual errors of points by cluster standard deviation

The pseudo code of classical iterative K-means algorithm are as follow:

1. Initializing for cluster ***n*** ,numbering the clusters 1 through ***n***.
2. Compute cluster distance ***D***(p, q) as the between two objects for p,q, =1, 2, **...**, ***n***.
3. For quantitative vectors as a Euclidean distance D = (D(p,q))
4. Find Ci similar pair of clusters ***r*** and ***s***, such that the distance, ***D***(***r***, ***s***) < Ci
5. Merge p and q to a new cluster t and compute such that distance ***D***(***t***, ***k***) for any existing cluster ***k*** ≠ ***r***, ***s*** .
6. Delete the rows and columns of p[] and q[]
7. Add a new row and column in ***D*** corresponding to cluster ***t***.
8. Repeat Step 3 until it converges.

PROPOSED WORK

Data mining is a huge workspace for the analysis of large volumes of data which integrates techniques from several research fields such as machine learning,statistics,pattern recognition,artificial intelligence,and database systems.Large number of data mining algorithms implanted in these fields to accomplish different data analysis tasks.[9]k-means clustering algorithm plays an important role in data mining and in various application from information retrieval to CRM.The proposed study emphasis on finding the a way to make right place to initial centroids,since it is found that the efficiency of algorithm is totally depends upon the initial clusters made.[5]Traditional k-means is 50 years beyond in data clustering ,an oldest algorithm to data but very efficient and no doubt is fast as well can be use in different initial conditions,and thus it is a subject of research study which matters a lot in making difference among data miming clustering algorithm and make a use of it in a benchmark level.

The process includes three way shield in order to make the right placement of initial centroids that is normalisation in preprocessing ,we have used on e of the best noramlisation technique called as mini-max and then geometric transformation in order to remove the normal values residuals occur and lastly PCA-PART,it is called as principal component analysis, this analysis is based on finding the principal components which do not change their direction It is the one of the best initialization method to find the initial means correctly.[4]

Proposed k-means algorithm

The proposed algorithm includes calculating difference between each instances to other ,sum the differences and square it,calculating the average values of each instances,finding the normalized value using mini max formula withing range of [0,1]Simple data (Normalized data) ,forming arrays for vectors in 2 dimensional space.

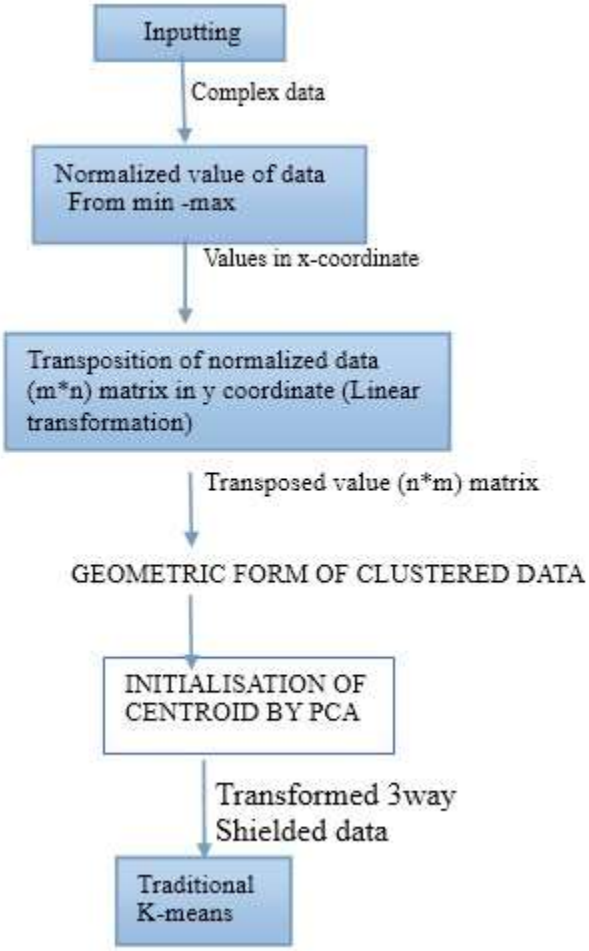

Inputting
Complex data
Normalized value of data From min -max
Values in x-coordinate
Transposition of normalized data (m*n) matrix in y coordinate (Linear transformation)
Transposed value (n*m) matrix
GEOMETRIC FORM OF CLUSTERED DATA
INITIALISATION OF CENTROID BY PCA
Transformed 3way Shielded data
Traditional K-means

Fig. Flowchart of enhanced K-means process

Process in brief

Proposed work is focuses on the removing such problems any finding a way towards it.from fig we can see that the preprocessing step is making a head towards normalisation,since it is required to map low-level data in the forms to make it compact, abstract and useful.Being seeing the importance of data preprocessing,an urgent need for a new computational theories as well as tools to assist humans in extracting useful knowledge from the multivariate digital data[10]In addressing issues of mapping of data it is required data to be uniformly representing and proper check of handling missing data, and handling noise and error as well.

By analysis of performance metrics and variance using multifactorial and parametric method efficiently able to assess the impact of real data set used in direct marketing ,this can be mainly achieve by choosing appropriate preprocessing choices in order to apply it in various fields [11]. When following KDD(Knowledge discovery in database)process it becomes prominent need for many organizations or companies to take a keen eyes on the data and process it mainly because of low-quality information in various data sources on the Web[12]

Normalization

The process of normalization can be understood through this forms,once the data is converted into mathematical expression/representation,the same objects of a data can be represent in different forms. standard representation represent every object uniquely.

The role of normalization is a very vast fields of data as for eg. Audio to imaging data.In order to reduce redundancy and dependency of relational database.It is noted that input validation is a very prominent requirement for software security and for the malicious input vulnerability,generally it is very is high if the input data is not preprocessed well.In order to prevent such security problems,the data when inputed need to be normalized removing encoded format and converting into a common scale within a range.In proposed work normalisation plays a vital role in doing preproceesing of data ,since the data is in the form of complex one is difficult to process further in algorithm,it is required to make the data desirable without losing any information,in other techniques if we see in order to make a data efficient a valuable data points also get removed.

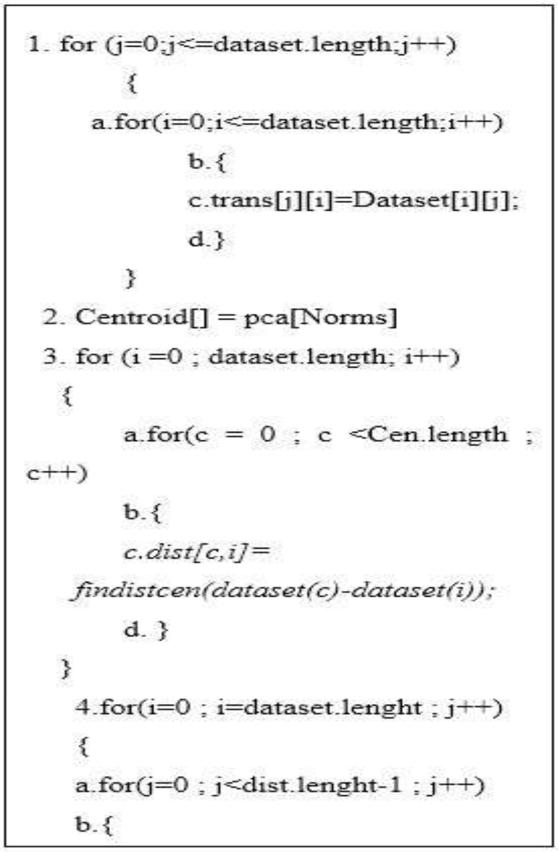
Thus it is required to find such way to prevent it ,several facts surveyed and attribute normalisation found to be the best first step applied on the data which make the data in normal ,simple scale.The formulation which is used are as follow:

Mini max formulation

max-i ⁄ max-min, where max is the maximum value of instance and min is the mimimum value ,i_1 is an instance.
Since dataset and the choose of algorithm is an important factor and related to each other.It is required to optimize the dataset and choose appropriate algorithm in order to extract useful knowledge in order to select an algorithm that will give desired results[1]..The proposed work uses transposition keeping purpose and usefulness in mind,it is mainly done to make the matrices order in a right form since it saves computational time,At the starting of calculation the algorithm usually takes usually pivot values which are least distant from zero.the concept includes the reflecting the matrix over its main diagonally from top left to bottom right such that all the pivot values get fixed and elements put into right place in a vector subspace.

For Example: A=[12134;12134;12134]
A’= [111;222;111;333;444]
[A’]’=[12134;12134;12134]

Thus we can see the matrices elements interchange it rows and columns such that it possesses symmetry removing round off error.



```
1. for (j=0;j<=dataset.length;j++)
        {
      a.for(i=0;i<=dataset.length;i++)
            b.{
            c.trans[j][i]=Dataset[i][j];
            d.}
        }
 2. Centroid[] = pca[Norms]
 3. for (i =0 ; dataset.length; i++)
   {
        a.for(c = 0 ; c <Cen.length ;
c++)
        b.{
        c.dist[c,i]=
   findistcen(dataset(c)-dataset(i));
        d. }
   }
   4.for(i=0 ; i=dataset.lenght ; j++)
   {
   a.for(j=0 ; j<dist.lenght-1 ; j++)
   b.{
```

Fig. Pseudo code of proposed k-means algorithm

PCA(INITIAL INTELLIGENT CENTROID)
uses a divisive hierarchical approach based on
PCA (Principal Component Analysis). The process of PCA start from an initial cluster that contains the entire data set, firstly,the method recursively selects the cluster with the largest SSE and divides it into two subclusters ,this is achieve by using a hyperplane that passes through the cluster centroid and is orthogonal to the principal eigenvector of the cluster covariance matrix. This procedure is repeated until K Given a set of N points in a metric the clusters are obtained. The centers are then given by the initial centroids of these clusters.

EXPERIMENTAL AND EVALUATION RESULTS

There are 4 datasets are used in experiments particularly choosing large scale,multivariate datasets considering the problem domain of kmeans clustering . This dataset are taken from UCI repositery The Machine Learning (ML) Repository is commonly used by industrial and academic researchers. It is widely cited in the artificial intelligence literature and the \UCI data sets" are the most widely used benchmark for empirical evaluation of new and existing learning algorithms. the dataset named are laustralian dataset30,lbreast cancer wisconsin dataset,lpima indian diabetes dataset,nlbupa dataset.These dataset are chosen in order to keep in mind about problem domain of research ,since the research part chosen is Kmeans clustering data mining algorithms as this algorithm failed to give good results for multivariate large dataset though it is fast in nature and can run in several conditions ,the enhanced k-means algorithm experimental set up is conducted .The parameters on which the performance is evaluated is accuracy and run time which is based on memory usage.

Fig (a),(b),(c),(d)The two algorithms classical and proposed enhanced algorithm is tested and evaluated the results comparison between showing how proposed work outperforming the classical one.
Proposed enhanced Kmeans algorithm: It includes the 3 way process to give the fine grained dataset (a large dataset) which kmeans fails to perform on to give initial centroids to make algorithm work effectively.A dataset processed various transformations from normalisation to linear transformation and lastly and importantly the PCA for outcoming of initial centroids.

Comparison of proposed and classical k-means clustering algorithm

Clustering is the process of grouping the object belongs to the similar groups ,this groups called as clusters .No. Of accurate clusters in less time inidicate decreasing the np hard problem found in K-means clustering algorithm.The membership of object existence which decides on which cluster it belongs is decided by the the variance between the points and distance calculated ,in which initial centroid plays a major role in begining the process.The results below show the cluster formation with the initial centroid calculated with accuracy and time by classical and enhanced kmeans clustering algorthm on different datasets and summarized the observations from experimental results.

Computation efficiency

By calculating computation efficiency we study the relation of number of time algorithm run (tested) over accuracy and time. Computation efficiency is another important performance indicator of the data mining algorithms . In this section, we evaluate the computation the efficiency of enhanced and classical k means algorithm in terms of running time ,different datasets are examined and compared in over number of test run for accuracy and time efficiency.

For k=3 ,the initial centroids of proposed algorithm for different data sets is calculated by PCA whereas the for k-means random initialisation is done.Thus from the figures we can see that the over different dataset the proposed k-means is outperforming the classical k-means.Since the dataset taken to be large datasets as per considering the problem domain of k-mean ,it is observe that different size of dataset gives different accuracy and run time.In case of laustralian dataset (No. Of instances =690,attributes 14) fig (a) classical k-mean have accuracy of 76.844 in time 0.06,whereas enhanced k-means having accuracy of 79.92 in time 0.03 which is dramatically improved ,in lbreast wisconsin dataset (No. Of instances=699,attributes =10) fig (b) in this accuracy is 82.20 in time 0.05 and enhanced k means accuracy is 89.72 in time 0.034,though in this dataset the k-means improved as compared to previous dataset but thr proposed results are overwhelming for this dataset .lp indian diabetes dataset having instances =768 and attributes =8 this datset is having the least size as comparable to other datasets,accuracy is 73.6 in time 0.08,whereas in proposed accuracy is 85.21 in time 0.039.It should be noted that the accuracy showing the no. of clusters create in % and time in ms. fig (a) shows the accuracy (%) of classical k-mean from the fig it can be seen that the at every no.

Of test run the accuracy changes this effect the efficency drastically for every datasets.whereas from fig (b) it can be seen that the accuracy of proposed k mean is constant and at every no. Of test run for different datasets.

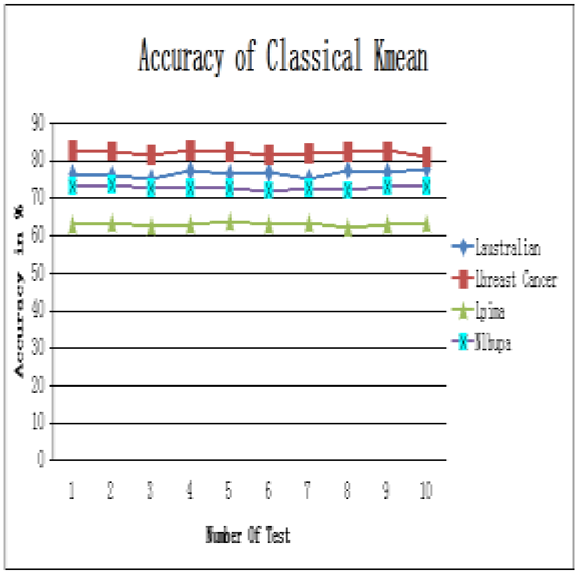


Fig.(a) Line chart showing datasets accuracy of classical Kmean

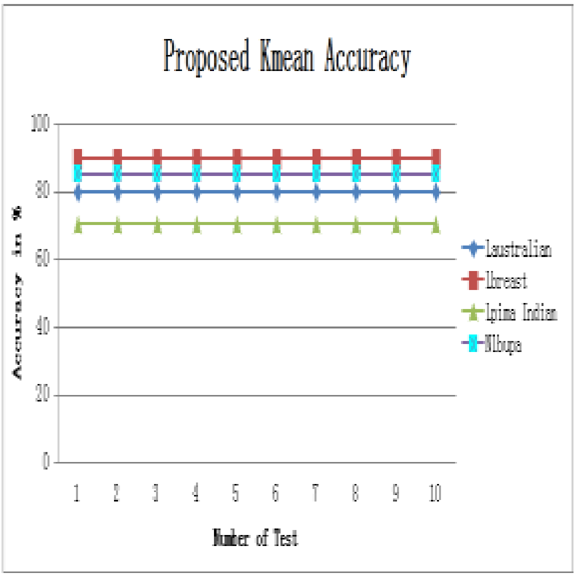


Fig.(b) Line chart showing datasets accuracy of proposed Kmean on n times of run

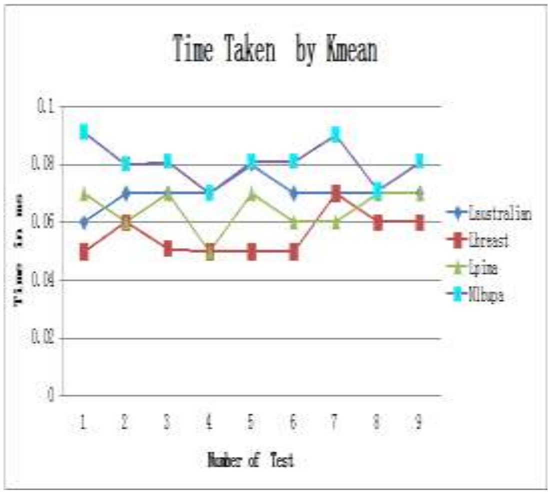


Fig.(c) Linechart showing datasets time taken by Classical K mean on times of run

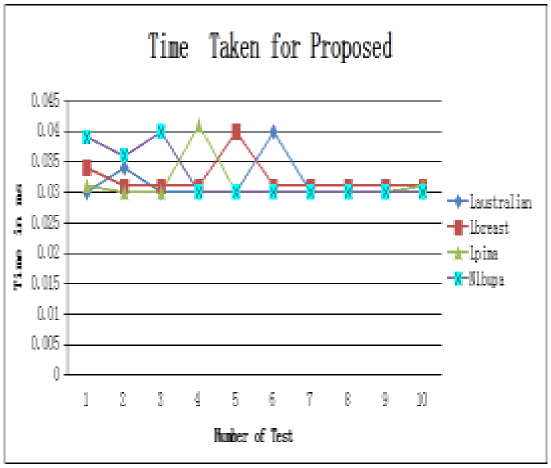


Fig (d)line chart showing datasets time taken by proposed enhanced Kmean on times of run

In case of time achieved in every number of test run ,the running time exhaustively changes whereas from fig (d)on the number of test run for proposed algorithm the time take by proposed is very minor and after that it follows the constant path in terms of time.

CONCLUSION

In this study we examined 3 way enhanced k-mean clustering algorithm and examined its results with classical K-means ,as we have seen thousands of enhanced clustering algorithms and new ones continue came into exist.To find the best algorithm is again the part of research with the motivation of the surveyed on k-means and its weakness and strong points and research implementation is proposed and evaluated for appropriate solution finding.From the experimental study it has been seen that the proposed 3 way enhanced k mean algorithm outperforming the classical k-means.The proposed work is the 3 way shield protection enhanced k-mean algorithm which make the dataset in useful form by reducing SSE error through the process of transformation,firstly by normalizing,linear transformation in order to find pivot entry by exchanging rows and column and finally by creating intelligent centroids by PCA method. Proposed study focuses on the main problem of k-means clustering and making the best use of it and can be used in different application such as in document clustering or in pattern analysis.Np HARD problem which is key problem domain of K-means clustering algorithm which drastically reduced its efficiency by keeping this in mind the research is done.The challenges of finding the accurate clusters in less time as compared to K-mean algorithm has been achieved leading this experiment to be successful.

REFERENCES

[1] A Study of the Factors Considered when Choosingan Appropriate Data Mining Algorithm,TeressaT. Chikohora,ISSN: 2231-2307, Volume-4, Issue-3, July 2014

[2] Ahamed Shafeeq B M and Hareesha K S,Dynamic Clustering of Data with Modified K-Means Algorithm IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore

[3] Fan Cai, Nhien-An Le-Khac, M-Tahar Kechadi, Clustering Approaches for financial data analysis: a survey,2012

[4] M. Emre Celebi,Hassan A. Kingravi,Patricio A. Vela ,Expert Systems with Applications, 40(1): 200–210, 2013

[5] Anil K. Jain, Data Clustering: 50 Years Beyond K-Means,2010

[6] Madhu Yedla, Srinivasa Rao Pathakota , T M Srinivasa,Enhancing K-means Clustering Algorithm with Improved Initial Center(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, 121-125

[7] Ryan S.J.D. Baker, Kalina YACEF,The State of Educational Data Mining in 2009: A Review and Future visions

[8] Fahim A.M, Salem A.M, Torkey F.A, Ramadan M.A, An efficient enhanced k-means clustering algorithm,1626 al. / J Zhejiang Univ SCIENCE A 2006 7(10):1626-1633

[9] XindongWu • Vipin Kumar • J. Ross Quinlan • Joydeep Ghosh • Qiang Yang •Hiroshi Motoda • Geoffrey J. McLachlan • Angus Ng • Bing Liu • Philip S. Yu •Zhi-Hua Zhou • Michael Steinbach • David J. Hand • Dan Steinberg, Top 10 algorithms in data mining,Knowl Inf Syst (2008) 14:1–37

[10] The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing,Sven F. Crone , Stefan Lessmann , Robert Stahlbock,European Journal of Operational Research 173 (2006) 781–800

[11] DATA PREPARATION FOR DATA MINING,Applied Artificial Intelligence, 17:375–381, 2003

[12] Pavel Berkhin,A Survey of Clustering Data Mining Techniques,2003

[13] G.Milligan, M. C. Cooper, A Study of Standardization of Variables in Cluster Analysis, Journal of Classification5 (2) (1988) 181–204

[14] scholar.google.co.in