Improve Performances and Efficiency in Clustering by using Single pass Seed selection Algorithm

Shaik Asif Ali[1] Dr K. Venkataramana[2]
[1]Student [2]Professor
[1,2]Department of Computer Applications
[1,2]KMM Institute of PG Studies, Tirupati, India

Abstract— We widely use k-means method for clustering technique for various applications. However, the k -means often converges to nearby superior and the end result depends on the initial seeds. Inappropriate desire of initial seeds may also yield terrible outcomes. K means++ is a manner of initializing k-means by way of choosing preliminary seeds with particular possibilities. Due to the random selection of first seed and the minimal likely distance, the k-means++ additionally outcomes different clusters in special runs in extraordinary range of iterations. In this examine we proposed a technique referred to as Single Pass Seed Selection (SPSS) algorithm as modification to k-means++ to initialize first seed and probable distance for k-means++ based at the point which became near extra wide variety of other factors within the statistics set. We evaluated its overall performance by making use of on various datasets and compare with k-means++. The SPSS set of rules turned into a single skip set of rules yielding particular solution in much less variety of iterations whilst in comparison to ok-means++. Experimental effects on real information sets from UCI established the effectiveness of the SPSS in generating regular clustering effects. Conclusion: SPSS performed well on high dimensional facts units. Its performance improved with features in the data set; specifically while range of capabilities we recommended the proposed method.

Key words: Clustering, K-Means, K-Means++, Local Optimum Minimum Probable Distance, SPSS

I. INTRODUCTION

Clustering is one in all the vital unattended learning in data processing to cluster the similar options. The growing purpose of the cluster is understood as a seed. to pick out the acceptable seed of a cluster is a crucial criterion of any seed primarily based bunch technique. The performance of seed primarily based algorithms square measure smitten by initial cluster center choice and also the best range of clusters in AN unknown knowledge set. Cluster quality And a best range of clusters square measure the vital problems in cluster analysis. During this paper, the planned seed purpose choice formula has been applied to three band image knowledge and 2nd distinct knowledge. This formula selects the seed purpose mistreatment the construct of maximization of the chance of element intensities with the space restriction criteria. The best range of clusters has been selected the idea of the mixture of seven totally different cluster validity indices. we've additionally compared the results of our planned seed choice formula on AN best range of clusters mistreatment K-Means bunch with alternative classical seed choice algorithms applied through K-Means bunch in terms of seed generation time (SGT), cluster building Time (CBT), segmentation entropy and also the range of iterations (NOTK−means). We have conjointly created the analysis of computer hardware time and no. of iterations of our planned seed choice methodology with alternative clump algorithms. clump is that the method of grouping similar information into teams known as clusters, so the objects within the same cluster area unit a lot of almost like {each alternative|one another} and a lot of completely different from the objects within the other cluster it's a helpful approach in data processing processes for distinctive hidden patterns and revealing underlying information from giant information collections The cluster analysis is that the most basic technique in numerous applications like data processing and information discovery (Fayyad et al., 1996), information compression and vector quantization), pattern recognition and pattern classification k-means++ may be a approach of initializing k-means by selecting initial seeds with specific possibilities and is O (log k) competitive. The k-means++ selects initial center of mass and minimum probable distance that separates the centroids willy-nilly. thus {different|totally completely different|completely different} results and different variety of iterations area unit doable in several runs. To get smart leads to less variety of iterations the k-means++ has got to be run variety of times. During this study we tend to propose a way, Single Pass Seed choice (SPSS) formula to initialize initial seed and therefore the minimum distance that separates the centroids for k means++ supported the purpose that is about to a lot of variety of alternative points within the information set. We've evaluated its performance by applying on numerous datasets and compare with k-means++. The experiments indicate that the SPSS algorithmic program converge k-means in less variety of iterations with distinctive resolution and additionally it performs well on high dimensioned knowledge sets compared to k-means++.

II. RELATED WORK

k-means could be a wide used bunch technique thanks to its simplicity, potency and discovered speed and therefore the Lloyds technique remains the foremost fashionable approach in apply it's the drawbacks as A priori fixation of variety of clusters Random choice of initial seeds. Inappropriate selection of variety of clusters and dangerous choice of initial seeds could yield poor results and should take additional variety of iterations to succeed in Holocaust. during this study we have a tendency to ar concentrating on choice of initial seeds that greatly have an effect on the standard of the clusters, {the variety|the amount|the quantity} of iterations and number of distance calculations needed for Holocaust. Fahim et al. (2006) planned a way to attenuate the quantity of distance calculations needed for convergence.

A. Refining Initial Points for K-Means Clustering

Practical approaches to agglomeration use associate degree repetitious procedure (e.g. K-Means, EM) that converges to 1

of various native minima. it's far-famed that these repetitious techniques square measure particularly sensitive to initial beginning conditions. we have a tendency to gift a procedure for computing a refined beginning condition from a given initial one that's supported associate degree economical technique for estimating the modes of a distribution. The refined initial beginning condition permits the repetitious algorithmic rule to converge to a "better" native minimum. The procedure is applicable to a large category of agglomeration algorithms for each distinct and continuous information. We have a tendency to demonstrate the applying of this methodology to the favored K-Means agglomeration algorithmic rule and show that refined initial beginning points so cause improved solutions. Refinement run time is significantly less than the time needed to cluster the total info. the strategy is ascendable and may be including a ascendable agglomeration algorithmic rule to handle the large-scale agglomeration issues in data processing. An interactive approach to mining factor expression information Effective identification of coexpressed genes and coherent patterns in organic phenomenon information is a crucial task in bioinformatics analysis and medicine applications. many agglomeration strategies have recently been projected to spot coexpressed genes that share similar coherent patterns. However, there's no objective commonplace for teams of coexpressed genes. The interpretation of co-expression heavily depends on domain information. moreover, teams of coexpressed genes in organic phenomenon knowledge area unit usually extremely connected through an outsized range of "intermediate" genes. There could also be no clear boundaries to separate clusters. cluster organic phenomenon knowledge conjointly faces the challenges of satisfying biological domain needs and addressing the high property of the info sets. during this paper, we tend to propose associate interactive framework for exploring coherent patterns in organic phenomenon knowledge. a unique coherent pattern index is projected to provide users extremely assured indications of the existence of coherent patterns. To derive a coherent pattern index and facilitate cluster, we tend to devise associate attraction tree structure that summarizes the coherence info among genes within the knowledge set. we tend to gift economical and ascendable algorithms for constructing attraction trees and coherent pattern indices from organic phenomenon knowledge sets. Our experimental results show that our approach is effective in mining organic phenomenon knowledge and is ascendable for mining giant knowledge sets.

B. A Graphical Aid to the Interpretation and Validation of Cluster Analysis

A new graphical show is projected for partitioning techniques. every cluster is painted by a supposed silhouette, that is predicated on the comparison of its tightness and separation. This silhouette shows that objects lie well among their cluster, and which of them area unit simply somewhere in between clusters. The entire cluster is displayed by combining the silhouettes into one plot, permitting associate degree appreciation of the relative quality of the clusters and an outline of the information configuration. the typical silhouette breadth provides associate degree analysis of cluster validity, associate degreed could be accustomed choose an 'appropriate' range of clusters.

III. PROPOSED ALGORITHM

In this planned system, a way referred to as Single Pass Seed choice (SPSS) rule as modification to k-means++ to initialize primary seed and probably distance for k-means++ supported the issue that turned into close to a lot of vary of different points within the records set. Result: We evaluated its overall performance via making use of on various datasets and evaluate with k-means++. The SPSS algorithm was a unmarried skip set of rules yielding particular solution in less quantity of iterations when compared to k-means++. Experimental effects on real statistics units from UCI tested the effectiveness of the SPSS in producing constant clustering effects. SPSS finished well on high dimensional statistics sets. Its efficiency increased with the boom of capabilities in the facts set; particularly whilst quantity of features greater than 10 we recommended the proposed method.

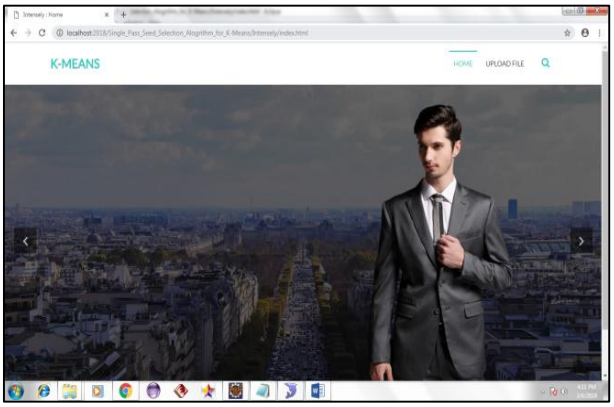
A. Algorithm

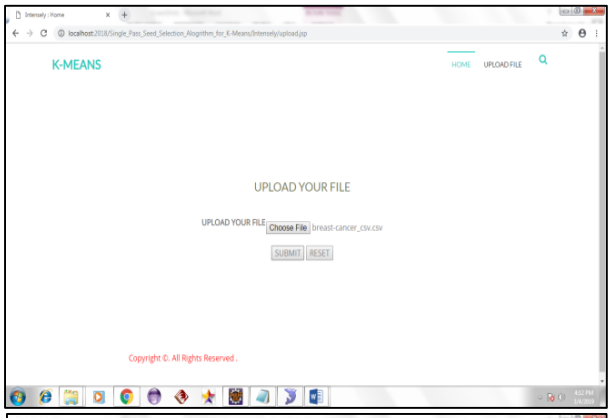
1) The SPSS Algorithm

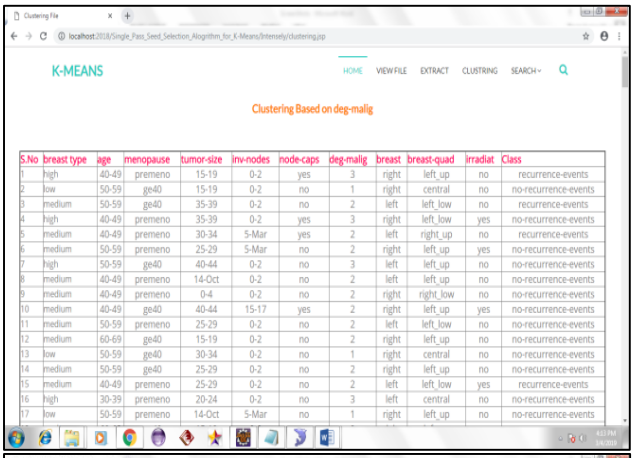
Choose a set C of k initial centers from a point-set (x_1, x_2 ,.., x_n). Where k is number of clusters and n is number of data points:

1) Calculate distance matrix Dist in which Dist (i,j) represents distance from i to j
2) Find Sumv in which Sumv (i) is the sum of the distances from i^{th} point to all other points.
3) Find the point i which is min (Sumv) and set Index = i
4) Add First to C as the first centroid
5) For each point x_i, set D (x_i) to be the distance between xi and the nearest point in C
6) Find y as the sum of distances of first n/k nearest points from the Index
7) Find the unique integer i so that
8) $(x_1)^2$+ $D(x_2)^2$+….+ $D(x_i)^2$>=y> $D(x_1)^2$+ $D(x_2)^2$+….+ $D(x_{i-1})^2$
9) Add x_i to C
10) Repeat steps 5-8 until k centers

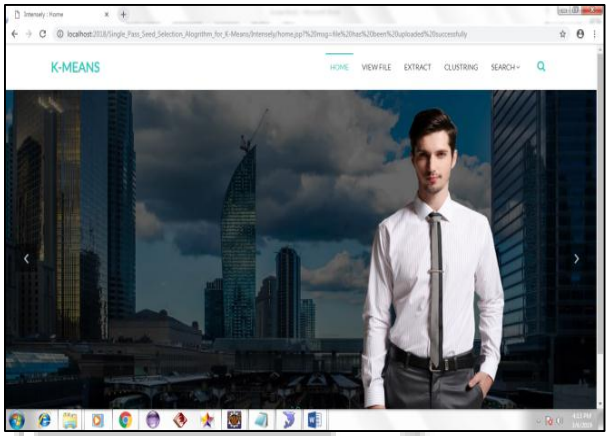
IV. RESULT

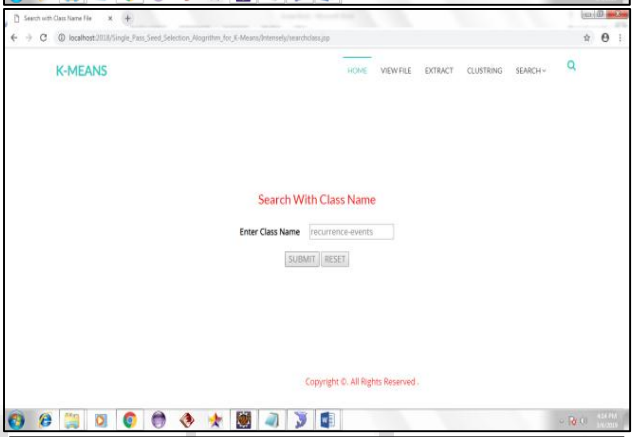


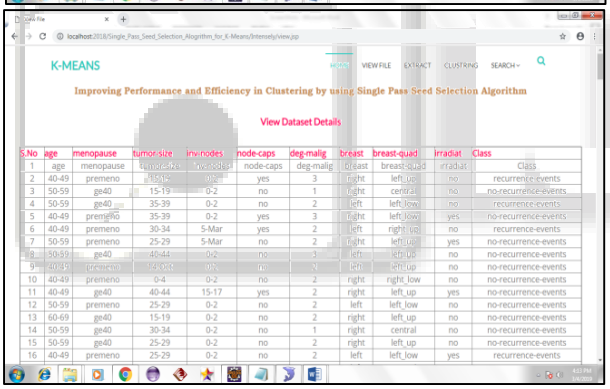




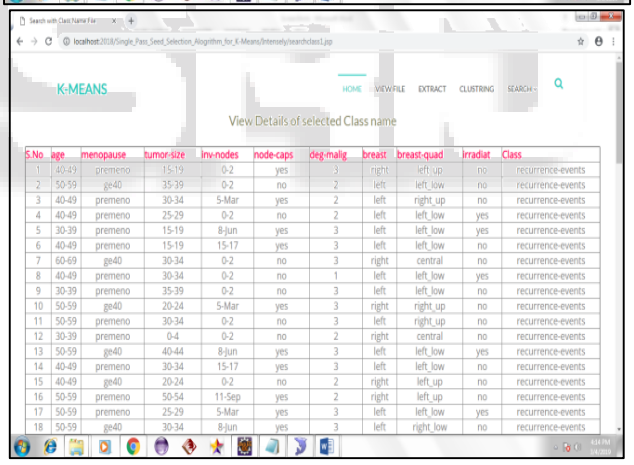
S.No	breast type	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
1	high	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurrence-events
2	low	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recurrence-events
3	medium	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurrence-events
4	high	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recurrence-events
5	medium	40-49	premeno	30-34	5-Mar	yes	2	left	right_up	no	recurrence-events
6	medium	50-59	premeno	25-29	5-Mar	no	2	right	left_up	yes	no-recurrence-events
7	high	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-recurrence-events
8	medium	40-49	premeno	14-Oct	0-2	no	2	left	left_up	no	no-recurrence-events
9	medium	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recurrence-events
10	medium	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recurrence-events
11	medium	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recurrence-events
12	medium	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recurrence-events
13	low	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recurrence-events
14	medium	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recurrence-events
15	medium	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurrence-events
16	high	30-39	premeno	20-24	0-2	no	3	left	central	no	no-recurrence-events
17	low	50-59	premeno	14-Oct	5-Mar	no	1	right	left_up	no	no-recurrence-events



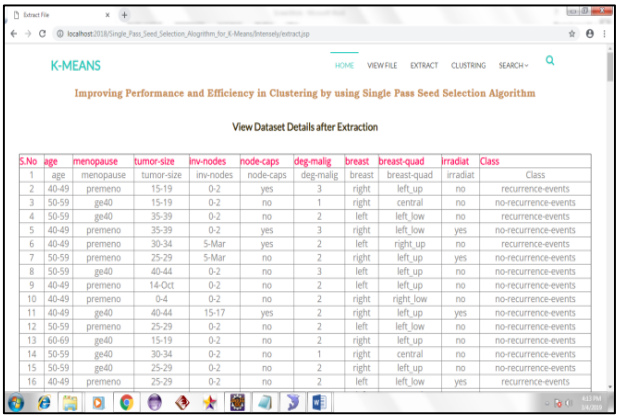




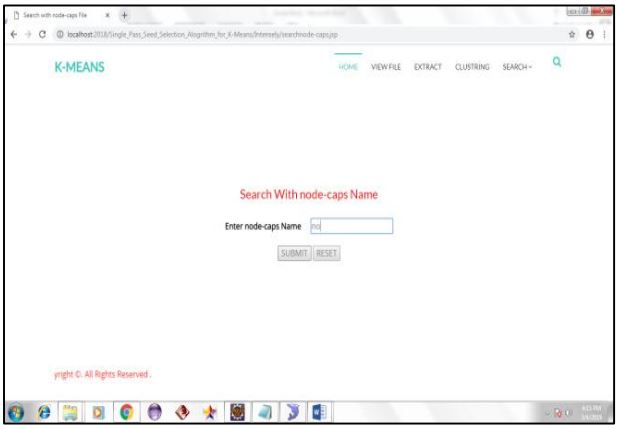
S.No	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
1	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
2	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurrence-events
3	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recurrence-events
4	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurrence-events
5	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recurrence-events
6	40-49	premeno	30-34	5-Mar	yes	2	left	right_up	no	recurrence-events
7	50-59	premeno	25-29	5-Mar	no	2	right	left_up	yes	no-recurrence-events
8	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-recurrence-events
9	40-49	premeno	14-Oct	0-2	no	2	left	left_up	no	no-recurrence-events
10	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recurrence-events
11	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recurrence-events
12	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recurrence-events
13	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recurrence-events
14	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recurrence-events
15	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recurrence-events
16	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurrence-events

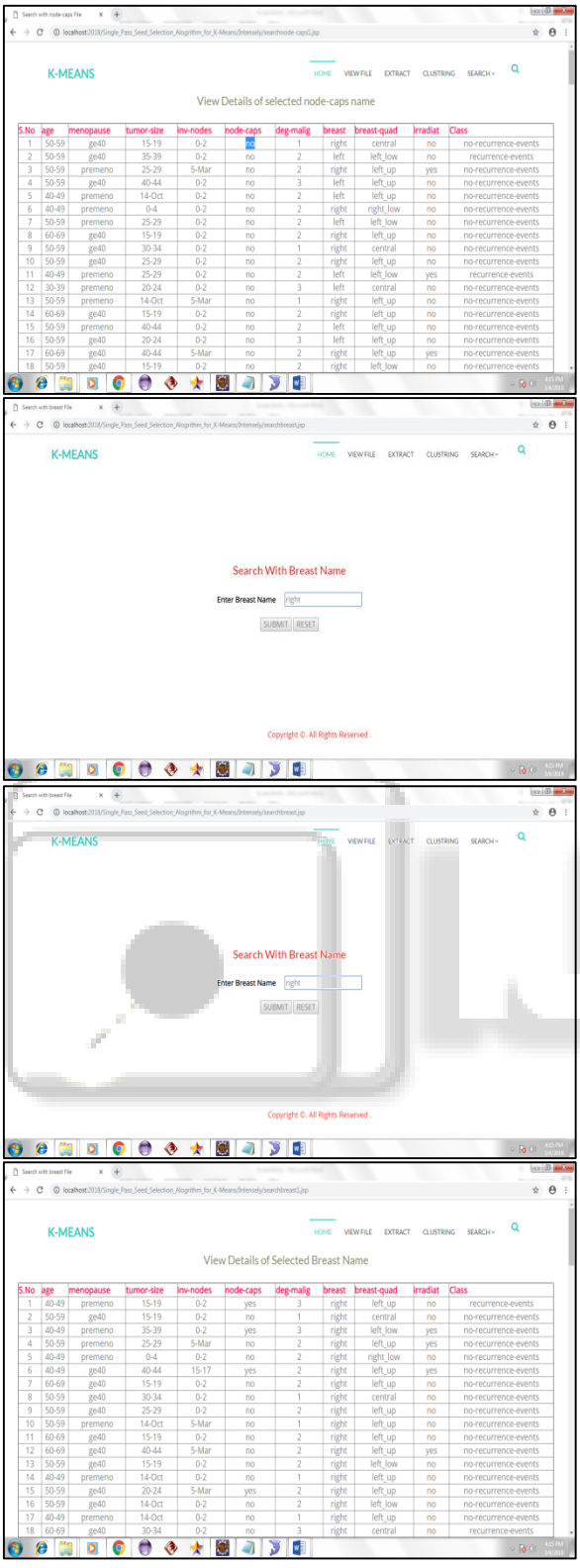


S.No	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurrence-events
2	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurrence-events
3	40-49	premeno	30-34	5-Mar	yes	2	left	right_up	no	recurrence-events
4	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurrence-events
5	30-39	premeno	15-19	8-Jun	yes	3	left	left_low	yes	recurrence-events
6	40-49	premeno	15-19	15-17	yes	3	left	left_low	no	recurrence-events
7	60-69	ge40	30-34	0-2	no	3	right	central	no	recurrence-events
8	40-49	premeno	30-34	0-2	no	1	left	left_low	yes	recurrence-events
9	30-39	premeno	35-39	0-2	no	3	left	left_low	no	recurrence-events
10	50-59	ge40	20-24	5-Mar	yes	3	right	right_up	no	recurrence-events
11	50-59	premeno	30-34	0-2	no	3	left	right_up	no	recurrence-events
12	30-39	premeno	0-4	0-2	no	2	right	central	no	recurrence-events
13	50-59	ge40	40-44	8-Jun	yes	3	left	left_low	yes	recurrence-events
14	40-49	premeno	30-34	15-17	yes	3	left	left_low	no	recurrence-events
15	40-49	ge40	20-24	0-2	no	2	right	left_up	no	recurrence-events
16	50-59	premeno	50-54	11-Sep	yes	2	right	left_up	no	recurrence-events
17	50-59	premeno	25-29	5-Mar	yes	3	left	left_low	yes	recurrence-events
18	50-59	ge40	30-34	8-Jun	yes	3	left	right_low	no	recurrence-events



S.No	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
1	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
2	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurrence-events
3	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recurrence-events
4	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurrence-events
5	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recurrence-events
6	40-49	premeno	30-34	5-Mar	yes	2	left	right_up	no	recurrence-events
7	50-59	premeno	25-29	5-Mar	no	2	right	left_up	yes	no-recurrence-events
8	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-recurrence-events
9	40-49	premeno	14-Oct	0-2	no	2	left	left_up	no	no-recurrence-events
10	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recurrence-events
11	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recurrence-events
12	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recurrence-events
13	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recurrence-events
14	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recurrence-events
15	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recurrence-events
16	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurrence-events





V. Conclusion and Future Scope

This paper talks k-means++ is a cautious seeding for k-means. However, for proper clustering outcomes it has to repeat number of times. The proposed SPSS set of rules is a SPSS algorithm is a single pass algorithm yielding unique solution with consistent clustering outcomes in comparison to k-means++. The SPSS algorithm offers precise consequences whilst the attributes of the records set are greater in variety. The computational project required by means of the SPSS set of rules is less comparative to k-means++ set of rules as the first seed and the minimum in all likelihood distance is selected randomly, this will boom the wide variety of iterations and accordingly it takes more time to attain final solution. Improving the performance of the proposed SPSS set of rules for low dimensional information sets and proposing an algorithm to generate variety of clusters with ultimate centroids is our destiny endeavor.

References

[1] Dembele, D. and P. Kastner, 2003. Fuzzy C-means method for clustering microarray data. Bioinformatics, 19: 973-980.http://bioinformatics.oxfordjournals.org/cgi/reprint/19/8/973

[2] Duda, R.O. and P.E. Hart, 1973..Pattern Classification and Scene Analysis. John Wiley Sons, New York, ISBN: 0471223611, pp: 482.

[3] Duda, R.O., P.E. Hart and G. David, 2001. Stork Pattern Classification. 2nd Edn., John Wiley and Sons, ISBN: 0471056693, pp: 654.

[4] Eisen, M.B., P.T. Spellman, P.O. Brown and D. Botstein, 1995. Cluster analysis and display of genome- wide expression patterns. Proc. Natl. Acad. Sci. USA., 95: 14863-14868. http://www.ncbi.nlm.nih.gov/pubmed/9843981

[5] Ester, M., H. Kriegel, J. Sander and X. Xu, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining, (KDD'96), Germany, pp: 1-6.

[6] Fahim, A.M., A.M. Salem, F.A. Torkey and M. Ramadan, 2006. An efficient enhanced k-means clustering algorithm. J. Zhejiang Univ. Sci. A., 7: 1626-1633. http://www.zju.edu.cn/jzus/2006/A0610/A061002. pdf

[7] Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, 1996. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, ISBN: 0262560976, pp: 611.

[8] Gersho, A. and R.M. Gray, 1992. Vector Quantization and Signal Compression, Kluwer Academic, Boston, ISBN: 0792391810, pp: 761.

