BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN 2348-0513
Vol. 2, Issue 3, Mar 2014, 11-18

INCREASING COMPARISON PERFORMANCE USING K-HARMONIC MEAN

SUCHITRA REYYA[1], M. PUSHPA[2], B. PRIYANKA[3], K. DEVI[4] & G. RUTHROSY[5]

[1]Assistant Professor, Department of Computer Science & Engineering, LENDI Institute of Engineering & Technology, Andhra Pradesh, India

[2,3,4,5]Department of Computer Science & Engineering, LENDI Institute of Engineering & Technology, Andhra Pradesh, India

ABSTRACT

Data clustering is one of the common techniques used in data mining. A popular performance function for measuring goodness of data clustering is the total within-cluster variance. The K-Means (KM) algorithm is a popular algorithm which attempts to find a K-clustering. The K-Means [2] algorithm is a centre based clustering algorithm. The dependency of the K-Means performance on the initialization of the centres is a major problem; a similar issue exists for an alternative algorithm, Expectation Maximization (EM) [6]. In this paper, we propose a new clustering method called the K-Harmonic Means algorithm (KHM). KHM [3] is a centre-based clustering algorithm which uses the Harmonic Averages of the distances from each data point to the centres as components to its performance function. It is demonstrated that K-Harmonic Means is essentially insensitive to the initialization of the centres. In certain cases, K-Harmonic Means significantly improves the quality of clustering results comparing with both K-Means and EM, A unified view of the three performance functions, K-Means', K-Harmonic Means 'and EM's, are given for comparison. Experimental results of KHM comparing with KM on Iris [4] data.

KEYWORDS: Clustering, K-Means, K-Harmonic Means, EM, Iris

INTRODUCTION

We are considering two algorithms Expectation–Maximization and K-mean algorithms, due to the problems of initialization of centres in these two algorithms, we consider another algorithm KHM it improves accuracy of centres the definition of these algorithms are as follows

Definition Expectation Maximization

Expectation–maximization (EM)[6] algorithm is an iterative method for finding maximum likelihood or maximum a posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Definition K Means

K-means [2] clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as prototype of the cluster.

KM and EM, is a centre-based, iterative algorithm that refines the clusters defined by K centres.

Definition K Harmonic Mean

K-Harmonic Means takes the sum over all data points of the harmonic average of the squared distance from a data point to all the centres as its performance function. We apply an iris dataset on the three algorithms i.e., K-mean, EM algorithms and KHM [1]. We calculate the three algorithms on centres and iterations on this iris dataset it can form three clusters they are Iris setose, Iris virginica and Iris versicolor. Experimental results obtained for this algorithm are centres are accurately obtained in K-harmonic mean algorithm than k-mean and EM algorithms.

OBJECTIVES OF THE STUDY

Increasing cluster efficiency using KHM- Input data is applied on KHM,K-mean and EM algorithms, initially random centers are allocated. Accurate centroid is calculated by using Harmonic Mean in KHM algorithm.

METHODS

K-mean Equations performed in K-mean algorithm are based on arithmetic mean calculation. Taking the iris [5] data set consisting of 4 attributes i.e., sepal length, sepal width, petal length and petal width of 150 data elements.

Equation: Arithmetic Mean=sum of all data elements/number of data elements

Expectation Maximization: Equations performed in EM algorithm are based on variance calculation. Taking the iris [5] data set consisting of 4 attributes i.e., sepal length, sepal width, petal length and petal width of 150 data elements.

***K*-Harmonic Means:** Takes the sum over all data points of the harmonic average of the squared distance from a data point to all the centres as its performance function.

Algorithm: K-Harmonic Means Clustering

Input: Dataset *xi* of *n* objects numbers of clusters *k* .

Output: Partition of the input data into *k* clusters

Procedure

Step 1: Declare a matrix *U* of size $n \times k$

Step 2: Generate *k* cluster centroids randomly within the range of the data or select *k* objects randomly as initial cluster centroids. Let the centroids be *C1, C2, . . . , Ck*. Calculate objective function value using

KHM(X,C)=N/(1/X-C)^2

Step 3: Compute the *U* membership matrix using HM.

Step 4: Compute new cluster centroids with membership values of each data object.

Step 5: Repeat step 2 to step 4 until convergence

Step 6: Assign data object i to cluster j with biggest Uij value

RESULTS AND DISCUSSIONS

KM algorithm, centroid is calculated by using Arithmetic Mean and in EM algorithm centroid is calculated by using Variance.The results generated by these three algorithms can be compared and finally we observe that in KHM, it got more accurate clusters and less no of iterations. Increasing cluster efficiency using KHM- Input data is applied on KHM,K-mean and EM algorithms, initially random centers are allocated.accurated centroid is calculated by using Harmonic Mean in KHM algorithm.

Arithmetic mean is calculated by using K-Mean algorithm,the centers values are obtained as very high than KHM algorithm.In expectation Maximization algorithm is using variance,here initialization of clusters are big drawback in this algorithm the centers are very high than k-mean and KHM.by using thest two algorithms KHM increases the performance of cluster efficiency.

algorithm the centers are very high than k-mean and KHM.by using thest two algorithms KHM increases the performance of cluster efficiency.

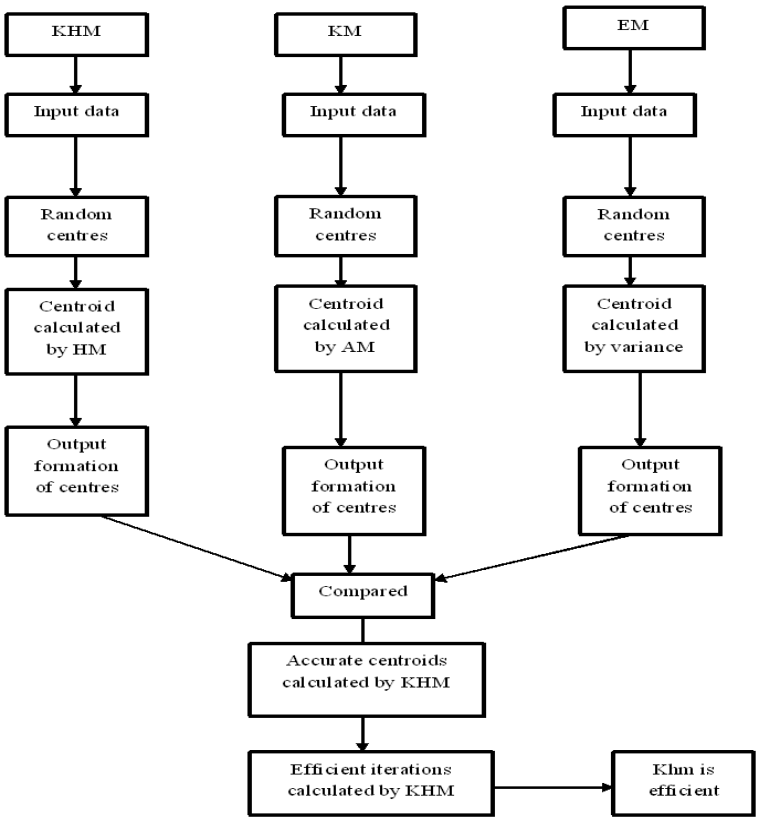

Figure 1: A Model for Increasing Accuracy

For the three algorithms can be run on the iris dataset we got the following iterations.

Iteration 1

The three algorithms are applied on Iris dataset. In the first iteration,the centres are as

Follows.if we observed that KHM centers are accurate.

Table 1

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KHM	6.21	3.21	2.86	0.57	5.49	2.63	3.05	0.52	4.54	3.18	1.32	0.18
EM	6.75	3.45	5.23	1.47	5.32	2.14	4.35	1.35	5.47	3.45	1.35	0.54
KM	6.24	3.25	5.21	1.25	5.21	2.14	4.25	1.25	5.21	3.25	1.25	0.25

sl-sepal length sw-sepal width pl-petal length pw-petal width

Iteration 2

The three algorithms are applied on Iris dataset. In the first iteration,the centres are as follows.if we observed that KHM centers are accurate.

Table 2

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KHM	6.79	3.10	5.33	1.83	5.49	2.65	4.41	1.41	4.98	3.37	1.44	0.20
EM	6.98	3.87	5.87	1.98	5.98	2.89	5.24	2.35	5.21	4.25	2.00	0.98
KM	6.85	3.45	5.21	1.52	5.84	2.78	4.57	1.25	5.21	3.54	1.47	0.54

Iteration 3

The three algorithms are applied on Iris dataset,we can get the third iteration,then the centres are obtained as follows, here KHM centers are accurate.

Table 3

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KHM	6.74	3.01	5.53	1.94	5.79	2.69	4.25	1.34	4.96	3.37	1.44	0.20
EM	6.25	3.87	8.25	2.36	6.35	3.45	4.23	2.54	5.87	3.87	1.25	1.25
KM	6.57	2.98	5.33	1.88	5.63	2.63	4.02	1.25	5.00	3.41	1.46	0.24

Iteration 4

The three algorithms are applied on Iris dataset we can get the fourth iteration, the centres are obtained as follows, KHM centers are accurate.at this iteration and KHM algorithm gets stabilized. Therefore it forms accurate centroids.

Table 4

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KHM	6.73	3.01	5.54	1.95	5.79	2.69	4.25	1.34	4.98	3.37	1.44	0.44
EM	6.87	3.11	6.12	3.21	2.14	3.12	4.98	2.36	5.23	3.25	1.09	0.45
KM	6.60	2.98	5.38	5.91	2.67	4.09	1.26	5.00	3.14	1.46	0.24	0.23

Iteration 5

EM and K-Mean algorithms are repeating the iterations and centres are obtained as follows

Table 5

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
EM	6.78	4.21	3.25	2.32	5.78	3.00	4.89	1.23	4.23	3.25	1.87	1.23
KM	6.63	2.99	5.43	1.93	5.72	2.69	4.15	1.29	5.00	3.41	1.46	0.24

Iteration 6

EM and K-Mean algorithms are repeating the iterations and centres are obtained as follows

Table 6

Cluster	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
EM	6.87	4.25	3.21	1.85	5.68	2.98	4.98	1.25	4.56	3.25	2.36	1.00
KM	6.66	3.00	5.49	1.96	5.78	2.71	4.20	1.33	5.00	3.41	1.46	0.24

Iteration 7

At this iteration the EM algorithm gets stabilized but it is less accurate than K-Mean algorithm. The centres obtained for the two algorithms are considered in the following table

Table 7

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
EM	7.25	4.25	3.14	1.23	6.21	2.98	4.98	1.25	5.98	3.25	1.87	0.99
KM	6.70	3.01	5.55	1.99	5.82	2.70	4.25	1.36	5.00	3.41	1.46	0.24

Iteration 8

At this iteration K-Mean algorithm gets centres are considered in the following table

Table 8

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KM	6.79	3.06	5.59	2.00	5.82	2.73	4.31	1.39	5.00	3.41	1.46	0.24

Iteration 9

At this iteration K-Mean algorithm gets centres are considered in the following table

Table 9

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KM	6.80	3.04	5.64	2.03	5.85	2.74	4.34	1.40	5.00	3.41	1.46	0.24

Iteration 10

At this iteration K-Mean algorithm gets centres are considered in the following table

Table 10

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KM	6.82	3.06	5.69	2.06	5.88	2.74	4.37	1.45	5.00	3.14	1.46	0.24

Iteration 11

At this iteration K-Mean algorithm gets centres are considered in the following table

Table 11

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KM	6.85	3.076	5.71	2.25	5.36	2.74	4.38	1.43	5.00	3.14	1.46	0.24

At the iteration 4 k-harmonic mean is stabilized and the remaining algorithms are forms centers until we get the repeated centers. At iteration 7 EM algorithm gets stabilized. By analyzing the above results it clearly shows that KHM algorithm can completed with in the 4 iterations with accurate centroids. EM algorithm can be stabilized at 7 iterations but it has very higher centroids than KM and KHM.K-MEAN algorithm can be stabilized at 11 iterations and has higher values than KHM.

Experiment Results

Centers obtained in the K-harmonic mean, K-mean, Expectation Maximization algorithms.

Table 12

Clusters	Cluster1				Cluster2				Cluster3			
Algorithm	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**	**SL**	**SW**	**PL**	**PW**
KHM	6.73	3.01	5.54	1.95	5.79	2.69	4.25	1.34	4.98	3.37	1.44	0.20
EM	6.88	7	4	3	7.11	5.24	3.01	3.86	2.09	1	6.03	5.86
KM	6.85	3.07	5.71	2.05	5.88	2.74	4.38	1.43	5.00	3.41	1.46	0.24

By observing above results for iris dataset it divides into three clusters. they consists of sepal length, sepal width, petal length, petal width. Centroids obtained for the 12 centers are listed above, similar for the Expectation Maximization algorithm-harmonic mean algorithms. Therefore from the above listed values we compare the three algorithms we got accurate values in k-harmonic mean for all the 12 centres.k-mean less better than KHM,EM got large values than KM and KHM.

CONCLUSIONS

This study showed that knowledge factor influence the use of nursing process more than other variables. One of the biggest problems currently facing the nursing profession is that of implementing the nursing process as lamented by Milne (1985) which the reporter believed that it can be influenced by the variables such as knowledge, profession, attitude, institution. Institutional factor ranks the second highest predictive value in the use of nursing process but currently, many institutions do not use nursing process for the care of their clients. for the negative attitude of nurses which is the least ranked in the use of nursing process.

ACKNOWLEDGEMENTS

We express our sincere and profound gratitude to LENDI institute management members of our Chairman Sri P. Madhusudhana Rao, Vice-Chairman Sri P. Srinivasa Rao and Secretary Sri K.Siva Rama Krishna for their valuable support and providing good infrastructure. Our special thanks to Principal Dr. V. V. Rama Reddy for his motivation and encouragement. We also thank our Head of the Department Professor A. Rama Rao for giving guidance and continuous support.

REFERENCES

1. Bradley, P., Fayyad, U. M., and Reina, C.A., “Scaling EM Clustering to Large Databases,” MS Technical Report, 1998.
2. [BF98] Bradley, P., Fayyad, U. M., C.A., “Refining Initial Points for KM Clustering”, MS Technical Report MSR-TR-98-36, May 1998
3. http://link.springer.com/chapter/10.1007%2F3-540-36175-8_7#page

4. http://archive.ics.uci.edu/ml/datasets/Iris

5. http://mlg.eng.cam.ac.uk/teaching/3f3/1011/iris.data

6. Bishop, C.M. (1995) Neural Networks for Pattern Recognition, Oxford University Press, New York

APPENDICES

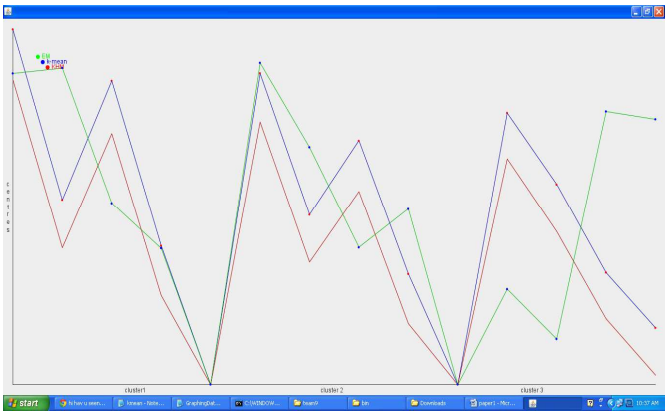

Figure 2: Graphs Generated for K-Harmonic Mean, K-Mean, EM Algorithms