AN IMPROVED HYBRIDIZED K-MEANS CLUSTERING ALGORITHM (IHKMCA) FOR HIGHDIMENSIONAL DATASET & IT'S PERFORMANCE ANALYSIS

H.S Behera
Department of Computer Science and Engineering,
Veer Surendra Sai University of Technology (VSSUT), Burla
Sambalpur, Orissa, India

Rosly Boy Lingdoh
Department of Computer Science and Engineering,
Veer Surendra Sai University of Technology (VSSUT), Burla
Sambalpur, Orissa, India

Diptendra Kodamasingh
Department of Computer Science and Engineering,
Veer Surendra Sai University of Technology (VSSUT), Burla
Sambalpur, Orissa, India

***Abstract*— In practical life we can see the rapid growth in the various data objects around us, which thereby demands the increase of features and attributes of the data set. This phenomenon, in turn leads to the increase of dimensions of the various data sets. When increase of dimension occurred, the ultimate problem referred to as the 'the curse of dimensionality' comes in to picture. For this reason, in order to mine a high dimensional data set an improved and an efficient dimension reduction technique is very crucial and apparently can be considered as the need of the hour. Numerous methods have been proposed and many experimental analyses have been done to find out an efficient reduction technique so as to reduce the dimension of a high dimensional data set without affecting the original data's. In this paper we proposed the use of Canonical Variate analysis, which serves the purpose of reducing the dimensions of a high dimensional dataset in a more efficient and effective manner. Then to the reduced low dimensional data set, a clustering technique is applied using a modified k-means clustering. In our paper for the purpose of initializing the initial centroids of the Improved Hybridized K Means clustering algorithm (IHKMCA) we make use of genetic algorithm, so as to get a more accurate result. The results thus found from the proposed work have better accuracy, more efficient and less time complexity as compared to other approaches.**

***Index Terms*—Data mining, Clustering, Dimensionality Reduction, Genetic Algorithm, curse of dimensionality, K-means clustering Algorithm, Canonical variate analysis.**

1. INTRODUCTION

Clustering is an unsupervised classification technique, which means that it does not have any prior knowledge of its data and results before classifying the data. Clustering can be defined as the process of organizing all similar data objects into a single group called Clusters and it's important that each cluster differs from every other cluster in one or many forms. The similarities are based entirely on the features of the data object. Clustering technique in itself has a very large and a wide range of applications under its belt, may it be in the field of pattern recognition, image processing, or may it be in the field of bioinformatics or in data mining. Basically Data mining deals with extracting meaningful information from the large junk of data sets. So the demand of an effective data mining methods is essential towards extracting the implicit information from large data sets stored in the data base. On the other hand dimensional reduction technique deals with transforming the high dimensional data in to a low dimensional which corresponds and project only the important features of the high dimensional data set. They are basically categorized in to two types:. Feature selection and . Feature reduction

In Feature reduction method features are reduced by transforming the original high dimensional data set in to a lower dimensional one through Eigen value decomposition. Canonical variate analysis is one of the

effective feature reduction technique employed for the task of reducing the dimensions of high dimensional data sets . In Feature selection method the best or the important features of the dataset are selected, so that it is converted to low dimension one. Canonical variate analysis is the dimension reduction technique that goes naturally with linear discriminant analysis. It finds the linear combinations that show the maximum ratio of between groups to within-groups variation. It explicitly looks for linear combinations that reveal differences among groups. It does so by solving a generalized eigen decomposition, looking for the eigen values and vectors of the between-groups covariance matrix "in the metric of" the within-groups covariance matrix. For high dimensional data set K-means clustering algorithm does not work effectively and accuracy is less due to noise and complexity in data sets. Here we proposed to use the canonical variate analysis to reduce the dimensionality, thus thereby less important attributes are eliminated. Then we apply the K means algorithm to the reduced data set. We use the genetic algorithm for finding the initial centroids which gave more effective and efficient results. It is better in terms of better accuracy and more efficiency than any other methods.

II. RELATED WORK

For improving the performance and efficiency of k-means clustering, various and numerous methods have been proposed. A hybridized K-Means clustering approach for high dimensional data set was proposed by Dash, et al.[1] and in his paper he used PCA for dimensional reduction and for finding the initial centroids a new method is employed that is by finding the mean of all the data sets divided in to k different sets in ascending order. This approach stumble, when time complexity is taken into account and it may eliminates some of the features which are also important for explicit extraction of information.

For improving the performance of K-Means clustering M Yedla et al[6] proposed an enhanced K-Means algorithm with improved initial center by the distance from the origin. The approach seems to solve the initialization problem but does not give any guarantee regarding the performance of the algorithm in terms of Time complexity and other matters.

Fahim A M et al[5]. Proposed an efficient method for assigning data points to clusters. Here the required computational time is reduced by using a new approach of assigning the data elements to the appropriate and required clusters. But in this method the initial centroids are selected randomly so this method is very sensitive to initial starting point and it doesn't produce unique clustering result. For obtaining the best results this approach is applied repeatedly to the same datasets and lastly the best one is selected.

Zhang Chen et al[4] proposed the initial centroids algorithm based on K-Means that have avoided alternate randomness of initial centroids. Bashar Al Shboul et.al [3] proposed an efficient way of initializing K-Means clustering by using Genetic algorithm thus there by solve the problem of randomly initializing the centroids.

A. K-MEANS CLUSTERING ALGORITHM:

The K-Means algorithm is one of the partitioning based, nonhierarchical clustering technique. For any given set of numeric objects X and an integer number K, the K-Means algorithm searches for a partition of X into k clusters that minimizes the within groups sum of squared errors. The K-means algorithm starts by initializing the k cluster centers. The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster. The processes of re-assigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers. The K Means clustering method can be considered, as the cunning method because here, to obtain a better result the centroids, are kept as far as possible from each other. The steps for the K-means algorithm are given below:

1. Initialization: choose randomly K input vectors (data points) to initialize the clusters.
2. Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.
3. Mean update: update the cluster centers in each cluster using the mean (Centroid) of the input vectors assigned to that cluster.
4. Stopping rule: repeat steps 2 and 3 until no more change in the value of the means.

B. PRINCIPAL COMPONENT ANALYSIS:

The high dimensional datasets or the multivariate datasets usually are stretched out along certain directions due to correlations among the variable in the variable space. They generally do not fill all of the variable space uniformly. These correlations arise due to several different measured attributes that respond in similar ways to some common underlying factor, so that there is some degree of redundancy among the variables. The overall orientation of the data cloud in multivariate space is given by the rough estimation of the covariance (or correlation) matrix of the data. Principal component analysis uses an eigen decomposition of the covariance matrix to construct a rotated set of orthogonal axes in the variable space such that the first axis– or principal component – is oriented in the direction of greatest variability in the data, the second principal component (P.C.) shows the second highest variability along orthogonal to the first P.C., and so forth.

C. CANONICAL VARIATE ANALYSIS:-

Canonical variate analysis is the dimension reduction technique that blends and goes naturally with linear discriminant analysis. We know principal component analysis tries to find the linear combinations of variables that show maximum overall variation, however canonical variate analysis on the other hands, finds the linear combinations that shows the maximum ratio of between groups to within-group variation. PCA uses an Eigen decomposition of the global covariance matrix, treating all the data as one group. Though PCA may help to reveal clusters in the data, but it does so by happen instance – if the groups are reasonably well separated, then the differences between them will be a significant component of the overall variation and PCA will pick up on this. However CVA, it explicitly looks for linear combinations that reveal differences among groups. It does so by solving a generalized eigen decomposition, looking for the eigen values and vectors of the between-groups covariance matrix "in the metric of" the within-groups covariance matrix, thereby we can say CVA pick up the data not by happen instance but by looking in the light of the problem.

III. PROPOSED ALGORITHM

It is well known and obvious that K-Means clustering algorithm does not work well when applied on high dimensional data. However the problem is solved up to a considerable extent, using hybridized K-means clustering algorithm using PCA but it has its own merit and demerit. One of its demerits is that it does not ensure and gave any guarantee about its effectiveness and accuracy. Here we are going to used another but improved hybridized clustering technique whereby we employed the Canonical Variate analysis which has an better upper hand with regards to performance than any other methods. The high dimensional data sets is reduced to low dimensional using the canonical variate analysis. To the reduced dimensional data set we then applied the K-Means algorithm merged with genetic algorithm for initializing purpose. The results is found out to be more effective and more accurate as compared to the hybridized K mean algorithm using PCA approach for dimension reduction.

The improved hybridized K-Means clustering algorithm (IHKMCA) is applied on any given high dimensional datasets is as follows:

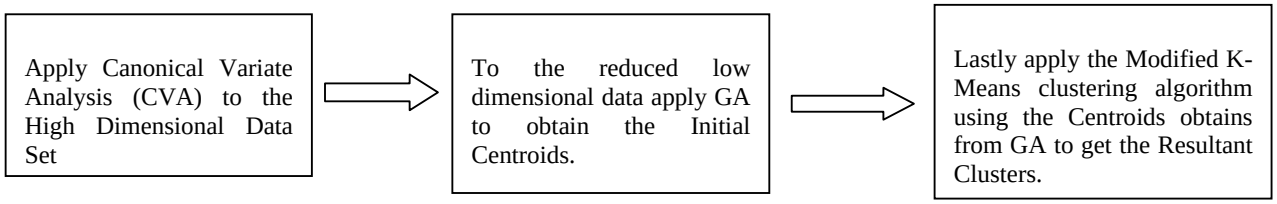

Fig 1. Working Model of IHKMCA

PROPOSED ALGORITHM:-

Input: X = {d1, d2,……..,dn} // set of n data items.

K // Number of desired clusters.

A. IMPROVED HYBRIDIZED K MEANS CLUSTERING ALGORITHM (IHKMCA):-

1. Organize the dataset in a matrix X.
2. Normalize the data set using Z-score.
3. Calculate the singular value decomposition of the data matrix.
4. Calculate the variance using the diagonal elements of D.
5. Find the co variance matrix S within the group by finding the variation of each group respective of its group mean.
6. .Then we find the between group covariance matrix B

$$B = \frac{K}{n(K-1)} \sum_{k=1}^{K} n_k (\overline{x_k} - \overline{x})(\overline{x_k} - \overline{x})^t$$

// Here k is the no of groups.nk is the no of data within th group.kx is the group mean.X is the global mean vector.

7. Find the eigen vector U of S-1B.
8. If u1 is the 1st eigen vector then u1 is first canonical variate is v1=u1'x then find the 2nd canonical variate accordingly.
9. Choose the "α" no of canonical variates which represents the data.
10. Find the reduced dataset Y1.

B. Genetic Algorithm to find the Initial Centroids:-pseudo code

```
Input: A resulting reduced data set W from CVA
Output: a set of K variables for initial centroids
    t=0;
    Initialize P(t);
    Evaluate P(t);
    While not (termination condition)
     Begin
     t=t+1;
    Select P(t) from P(t-1);
    Recombine pairs in P(t);
    Mutate P(t);
    Evaluate P(t);
    End
```

C. K-Means algorithm:-

```
Input:
 W //set of n data points.
 K // number of desired clusters
Output: a set of K clusters.
Steps:
 K initial centroids from genetic algorithm.
Iterative process:
Assign each point ai, to the cluster which has the closest centroid.
Calculate the new mean for each cluster UNTIL the convergence
criteria is met.
```

IV. EXIPERIMENTAL ANALYSIS AND PERFORMANCE EVALUATION

Initially, we used the proposed algorithm on a synthetic data objects of 15 having 10 attributes which is given by Dash et.al[1]. We have given an algorithm which is more efficient and accurate than the algorithm proposed by Dash et.al. The original data matrix X with 15 data objects and 10 attribute values:

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Data1	1	5	1	1	1	2	1	3	1	1
Data2	2	5	4	4	5	7	10	3	2	1
Data3	3	3	1	1	1	2	2	3	1	1
Data4	4	6	8	8	1	3	4	3	7	1
Data5	5	4	1	1	3	2	1	3	1	1
Data6	6	8	10	10	8	7	10	9	7	1
Data7	7	1	1	1	1	2	10	3	1	1
Data8	8	2	1	2	1	2	1	3	1	1
Data9	9	2	1	1	1	2	1	1	1	5
Data10	10	4	2	1	1	2	1	2	1	1
Data11	11	1	1	1	1	1	1	3	1	1
Data12	12	2	1	1	1	2	1	2	1	1
Data13	13	2	1	1	1	2	1	2	1	1
Data14	14	5	3	3	3	2	3	4	4	1
Data15	15	1	1	1	1	2	3	3	1	1

Table-1 Experimental synthetic data objects

Step1:-
Finding the reduced data set by applying conanical variate analysis which performs the task of reducing the dimension of a given high dimensional dataset

Step2:-
Now on the reduced dataset we apply the genetic algorithm of n variables where n signifies the number intial centroid required .Genetic algorithm in one way give the most accurate way of initializing as it selects variable from a data set which are best fitted .

Step3:-
After getting the number of initial centroids an efficient K-Mean clustering algorithm is applied and the clusters are determined

Step4:-Comparison of efficiency and accuracy of the proposed algorithm with the hybridized k means algorithm for high dimensional data set.

	V1	V2
Data1	0.536985	1.045863
Data2	-2.76212	1.914667
Data3	0.8597	0.73036
Data4	-2.78911	-0.43292
Data5	0.502727	0.44459
Data6	-7.19228	-0.525
Data7	0.691562	0.513194
Data8	1.149898	-0.19297
Data9	2.060598	1.744708
Data10	1.08491	-.31268
Data11	1.749453	-0.78052
Data 12	1.620462	-0.6419
Data 13	1.656486	-0.7997
Data14	-0.70786	-1.51675
Data15	1.540447	-1.16586

Table-2 Reduced dataset containing two attributes using PCA

	V1	V2
Data1	0.0099	-00076
Data2	0.0792	-0.5472
Data3	-0.0571	-0.0611
Data4	0.0298	-0.2075
Data5	-0.1152	0.0145
Data6	0.0572	-0.5571
Data7	-0.1695	-0.5196
Data8	-0.2237	0.33306
Data9	-0.2460	0.2391
Data10	-0.2879	0.0057
Data11	-0.3345	0.0168
Data12	-0.3634	0.0181
Data13	-0.3969	0.0101
Data14	-0.3721	-0.0781
Data15	-0.4591	-0..0941

Table-**3** Reduced dataset containing two attributes using IHKMCA

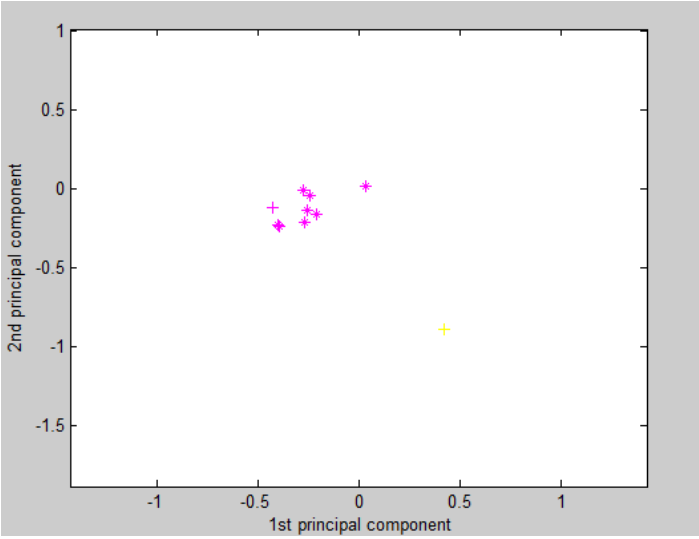

Figure-2 CLUSTERING USING HYBRIDIZED K-MEANS ALGORITHM

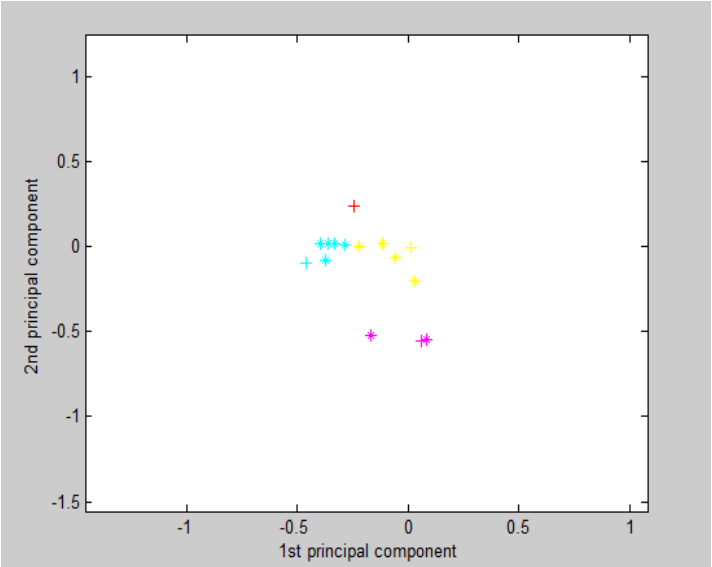

Figure-3 CLUSTERING USING IHKMCA

Data sets	K Means algorithm		Hydridized K Means algorithm using PCA		IHKMCA	
	Time complexity	Sum of Squared error(SSE)	Time complexity	Sum of Squared error(SSE)	Time complexity	Sum of Squared error(SSE)
Synthetic Data	6	23.285	5	10.12	3	2
Nutrients in Meat, Fish and Fowl	5	18.12	3	7.8	3	2..5
Breast Cancer	14	98.536	9	16.53	7	6.33

Table-4: Performance analysis

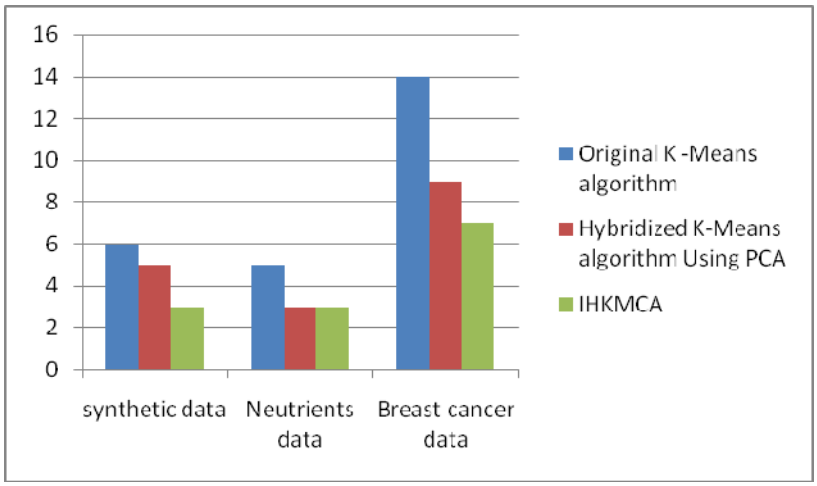

Figure-4: Comparison of Time complexity for different dataset

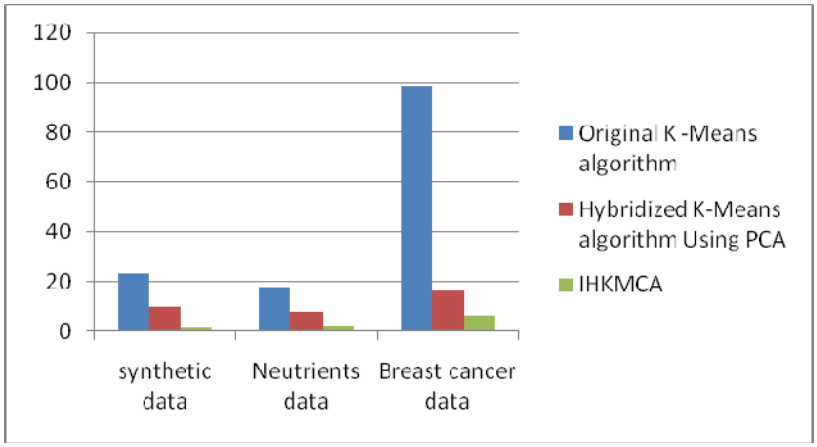

Figure-5: Comparison of Sum of Squared Error for different dataset

The above result shows that the proposed algorithm IHKMCA shows a better performance in terms of both speed and accuracy when compared it to the earlier proposed hybridized K Means clustering algorithm using PCA and original K Means Clustering algorithm taken on the experimental data. The performance is more efficient and effective even comparing the higher principal component using CVA as compares to using PCA which adds to the upper hand that CVA has over PCA and also over the conventional algorithm.

CONCLUSION AND FUTURE WORK:-

In this paper we have proposed an Improved Hybridized K-Means Clustering Algorithm (IHKMCA) using CVA as a dimension reduction technique and initialized the centroid of Improved Hybridized K-means algorithm(IHKMCA) by using genetic algorithm. We found out that IHKMCA has a better performance compare to the earlier Hybridized K-Means Clustering algorithm using PCA . Now the problem of determining the number of clusters before hand is still worth working for and also certain area of improvement for dimension reduction and outlier elimination is still to be explored and unveiled.

REFERENCES

[1] Dash et.al , “A Hybridized k-Means Clustering Algorithm for High Dimensional Dataset”, International Journal of Engineering, Science and Technology, vol. 2, No. 2, pp.59-66,2010.
[2] “Dimension Reduction And Cluster Analysis” By Geoff Bohling, EECS 833, 6 March 2006
[3] Bashar Al Shboul et.al “Initializing K-Means Clustering Algorithm by using Genetic Algorithm” , World Academy of Science, Engineering and Technology 54 2009
[4] Chen Zhang, Shixiong Xia et al, "K-means Clustering Algorithm with Improved Initial Center," Second International Workshop on Knowledge Discovery and Data Mining, wkdd, pp.790-792,2009

[5] Fahim A. M., Salem A. M., Torkey F. A., Saake G. and Ramadan et al, "An efficient k-means with good initial starting points", Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol. 2, No. 19, pp. 47-57,2009

[6] M Yedla et al[6]. "Enhancing K means algorithm with improved initial center", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , pp- 121-125,2010.

SHORT BIO DATA OF ALL THE AUTHORS

1. Prof H.S Behera is currently working as a Senior Lecturer in Dept. of Computer Science and Engineering is Veer Surendra Sai University of Technology(VSSUT), Burla, Orissa, India. His research areas of interest include Operating Systems, Data Mining and Distributed Systems.

2.Rosly Boy Lyngdoh is a Final year B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology(VSSUT), Burla, Orissa, India.

3.Diptendra Kodamasingh is a Final year B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology(VSSUT), Burla, Orissa, India.