A New Improved Hybridized K-MEANS Clustering Algorithm with Improved PCA Optimized with PSO for High Dimensional Data Set

H. S. Behera, Abhishek Ghosh, Sipak Ku. Mishra

Abstract*— The day to day computation has made the data sets and data objects to grow large so it has become important to cluster the data in order to reduce complexity to some extent. K-means clustering algorithm is an efficient clustering algorithm to cluster the data, but the problem with the k-means is that when the dimension of the data set becomes larger the effectiveness of k-means is lost. PCA algorithm is used with k-means to counter the dimensionality problem. However K-means with PCA does not give much optimisation. It can be experimentally seen that the optimisation of k-means gives more accurate results. So in this paper we have proposed a PSO optimised k-means algorithm with improved PCA for clustering high dimensional data set.

Index Terms*— Data mining, Clustering, Particle Component Analysis, Centred vector, Squared Sum Error, Lower bound, Bound Error, Particle Swarm Optimisation.

I. INTRODUCTION

Data mining is a convenient way of extracting patterns, which represents knowledge. It can be viewed as an essential step in the process of knowledge discovery. Data are normally pre-processed through data cleaning, data integration, data selection, and data transformation prepared for the mining task. Data mining can be performed on various types of databases and information repositories, but the kind of patterns to be found are specified by various data mining functionality like class description, association, correlation, analysis, classification, prediction, cluster analysis etc. High dimensional data are transformed into lower dimensional data using the principal component analysis (PCA). Such dimension reduction technique has wide range of application such as meteorology, image processing, genomic analysis, and information retrieval. It is also common that PCA is used to project data to a lower dimensional subspace and K-means is then applied in the subspace (Zha et al., 2002)[4].The main basis of PCA-based dimension reduction is that PCA picks up the dimensions with the largest variances. Mathematically, this is equivalent to finding the best low rank approximation (in L2 norm) of the data via the singular value decomposition (SVD) (Eckart& Young, 1936)[5].Dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. The data transformation is linear in case of PCA. In order to improve the efficiency of the clustering algorithm the redundant data may be removed so to minimize space required and the execution time. To do so, we can choose dimensionality reduction methods such as principal component analysis (PCA), Singular value decomposition (SVD), and factor analysis (FA). Among this, PCA is preferred to our analysis and the results of PCA are applied to a popular model based clustering technique. Using simple PCA for k-means clustering algorithm provides a less tight lower bound to the optimal k-means. To overcome this drawback an improved PCA is being developed which uses dataset with centred data as the input which has been proved to provide a tighter lower bound to the optimal k-means algorithm. The complexity of k-means can be reduced using optimisation algorithm. Some of the optimisation algorithms are Evolutionary Programming (EP), Genetic Programming (GP), Differential Evolutionary (DE), Genetic Algorithms (GA) etc. Here we have used Particle Swarm Optimisation (PSO) for optimisation purposes. Particle swarm optimization is a class of evolutionary algorithms which aims to find a solution to a given optimization problem [9].

The PSO has particles driven from natural swarms with communications based on evolutionary computations. PSO combines self-experiences with social experiences. In this Algorithm, a candidate solution is presented as a particle. It uses a collection of flying particles (changing solutions) in a search area (current and possible solutions) as well as the movement towards a promising area in order to get to a global optimum. One of the PSO problems is its tendency to a fast and premature convergence in mid optimum points. A lot of effort has been made so far to solve this problem [10].

II. RELATED WORK

For improving the performance and efficiency of k-means clustering, various and numerous methods have been proposed.

Manuscript received on April 14, 2012.

H.S. Behera, Faculty in Dept. of Computer Science and Engineering is Veer Surendra Sai University of Technology(VSSUT), Burla, Odisha, India, (e-mail: hsbehera_india@yahoo.com).

Abhishek Ghosh, B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India. (e-mail: abhishekgh3@gmail.com).

Sipak ku. Mishra, B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India. (e-mail: sipakmishra208@gmail.com).

A hybridized K-Means clustering approach for high dimensional data set was proposed byDash, et al.[1] and in his paper he used PCA for dimensional reduction and for finding the initial centroids anew 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 improvising time complexity and reducing squared sum error an improvised hybridized k-means clustering algorithm (IHKMCA) was proposed H.S. Behera et.al.[2] but it used simple PCA without centred data vector where again it had a less tight lower bound and the squared sum error increased in compared to PCA using centred data vector.

Improvising the performance of k-means clustering was proposed by P.Prabhuet. al[3].

Dimensionality reduction and the problems of the high dimensional data sets were proposed in Maaten[4] and Davy Micheal[5] respectively.

Performance analysis of k-means with different initialization method for high dimensional data was proposed by- Tajunisha and Saravan[6].

A new method of dimensionally reduction using k-means clustering algorithm for high dimensional data set was proposed by D.Napolean and S.Pavaloki.[7]

A improved k-means using PCA was discussed by Tanjunisha and Saravan[8].

A new auto clustering algorithm was developed in by Jakob R. Olesen, Jorge Cordero H., and Yifeng Zeng[9].

Different types of PSO techniques with their advantages and disadvantage was discussed by Davoud Sedighizadeh and EllipsMasehian[10].

III. METHODOLOGIES

A. K-Means Clustering Algorithm

K-means is a prototype-based, simple partition clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice.

This algorithm consist of two separate phases: the first phase is to select k centroids randomly, where the value of k is assigned earlier. The next phase is to assign each data object to the nearest centre. Euclidean distance method is generally used to determine the distance between each data object and the cluster centres. When all the data objects are assigned to each of the clusters, the cluster centres recalculation is done. This iterative process continues repeatedly until the criterion function of finding new cluster centres becomes minimum or reduced.

B. Principal Component Analysis (PCA).

Principal component analysis (PCA) involves a mathematical procedure that transforming a number of correlated variables of higher dimension into a smaller dimension of unrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. PCA involves the calculation of the eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centring the data for each attribute. The results of PCA are usually discussed in terms of component scores and loadings. PCA is the simplest of the true eigenvector-based multivariate analyses.

The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables. It is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set.

> The k-means clustering algorithm works as follows:
> a) Randomly select k data object from dataset D as initial cluster centres.
> b) Repeat
> a. Calculate the distance between each data object di (1<=i<=n) and all k cluster centres' c_j(1<=j<=n) and assign data object d_i to the nearest cluster.
> b. For each cluster j(1<=j<=k), recalculate the cluster centres.
> c. Until no change in the cluster centre.[8]

> The steps involved in PCA algorithm are:-
> Step1: Obtain the input matrix Table
> Step2: Calculate the covariance matrix
> Step3: Calculate the eigenvectors and eigen values of the covariance matrix
> Step4: Choosing components and forming a feature vector
> Step5: deriving the new data set

The eigenvectors with the highest eigen value is the principal component of the data set. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by Eigen value, highest to lowest. To reduce the dimensions, the first d (no. of principal components) eigenvectors are selected. The final data has only d dimensions.

C. Improved Principal Component Analysis.

Improved principal component analysis consists of finding out a particular vector matrix called as centred vectored matrix. In this centred vectored matrix consist of data points that are more centrally oriented towards the principal axis than PCA. In the improved PCA the centred data vectored matrix is calculated by applying the mean and standard deviation to the data sets. The rest of the procedures follow the PCA algorithm.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

By taking out the standard deviation and applying to the centred data, data points are more centrally aligned to the principal axis. There is more dimension reduction than primitive PCA because of the data points being aligned to the principal axis more.

In the experiments to be followed it is proved that using centred vectored matrix with PCA gives more optimal values than simple PCA. The factor of dimensionality has been reduced to a great extent with the use of improved PCA.

The algorithm to find out the centred vectored matrix is as follows:-

Step1:The original n data points to be clustered in m-dimensional space is contained in the data matrix $X=(x_1,\ldots,x_n)$

Step2: In general data is not centred around the origin. We denote the centred data matrix $Y = (y_1,\ldots\ldots,y_n)$, where$[y_i = {(x_i - \mu_x)}/{\sigma_x}]$, $\mu_x = \Sigma x/n$, $\sigma_x = \sqrt{\frac{(\sum_{i=1}^{n}(x_i-\mu_x)^2)}{n-1}}$

Step 3: The covariance matrix is given by $S =(xi -\mu_x)(xi -\mu_x)^T=YY^T$

Step 4: The principal eigenvectors u_k of YY^T are the principal directions of the data Y . The principal eigenvectors v_k of the Gram matrix Y^TY are the principal components; entries of each v_k are the projected values of data points on the principal direction u_k. v_k and u_k are related via: $v_k = Y^Tu_k/\lambda_k^{1/2}$: where λ_k is the eigenvalue of the covariance matrix YY^T.

D. Particle Swarm Optimization Algorithm (PSO).

Particle swarm optimization is a form of stochastic based optimization algorithm which was developed by observing the behavior of birds swarm.PSO was originally developed by Eberhartand Kennedy in 1995.In PSO the basic idea is that each particle represents a possible solution and is updates its self with respect to its neighbor. The basic principle of this algorithm is that it performs a search of an m- dimensional search space say ζ by using the particles and updates its position and velocity. Each particle is believed to occupy a position $x_i(t) = \{x_{i,j}(t)|j = (1,2,3 \ldots m)\}$ and velocity$v_i(t) = \{v_{i,j}(t)|j = (1,2,3 \ldots m)\}$.

The PSO algorithm works as follows:-

At first the particles are assigned an initial position$x(0), x_{min} < x(0) < x_{max}$, and the initial velocity $v(0)$ is set to zero. Then there is a updating of position and the velocity of the particles which are classified in local updating and global updating.

The local updating of the position is given by equation (1)

$$y(t+1) = \begin{cases} y_i(t) \; if \; f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) \; otherwise, \end{cases} \quad (1)$$

Where $y(t)$ are the local optima of the particles and $f(x(t))$ is the fitness function of the local optimization.

The global optimization of the position of the particles is given by the equation (2).

$$Y(t+1) = \begin{cases} y_i(t) \; iff(y_i(t+1)) \geq f(Y_i(t) \; \forall y_i(t)) \\ y_i(t) \; if \; \exists y_i(t) | \; f(y_i(t)) < f(Y_i(t)) \end{cases} \quad (2)$$

Next there is the velocity and position updating step which selects new values for the position $x_i(t+1)$ and velocity $v_i(t+1)$ using equations 3 and 4 respectively:

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (3)$$

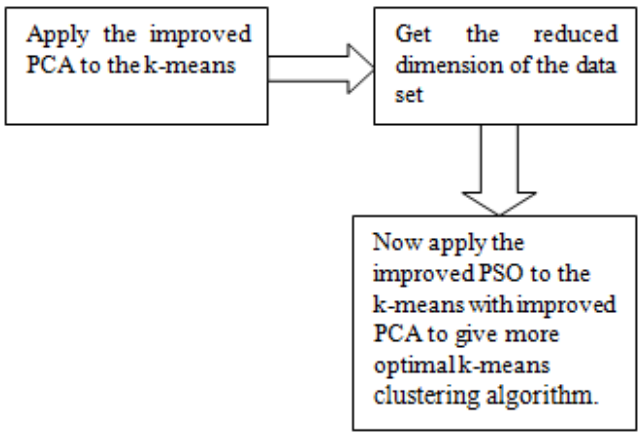
$$v_i(t+1) = v_i(t) + c_1r_1(y_i(t) - x_i(t)) + c_2r_2(Yi(t) - x_i(t)) \quad (4)$$

Where c1 and c2 are the parameters for local and global and r1 and r2 ϵ |0, 1|.

The updating of position and velocity of the particles are stopped till a convergence criterion is met [9].

IV. PROPOSED ALGORITHM

Improved PCA is applied to the k-means for the dimension reduction. PSO is applied subsequently to the improved PCA based k-means algorithm. The PSO optimised improved PCA based k-means algorithm gives a more optimal algorithm.



V. ALGORITHM

Input:Data set D={d_1,d_2.....d_n},where d_i=data points, n= no of data points

Cluster centre C={c_1,c_2,.....c_k), where c_i=cluster centres' ,k = no of cluster centres'.

Step 1: Given the data set points we find out the centered vectored matrix Y = (y_1,……,y_n) by the following method.

Step 2: We find out the mean and standard deviation of each data points by the given formula, $\mu_x = \Sigma x/n$ and $\sigma_x = \sqrt{\frac{(\sum_{i=1}^{n}(x_i-\mu_x)^2)}{n-1}}$ respectively.

Step 3: Then we find out the elements of the centred vectored matrix by the formula $y_i = (x_i - \mu_x)/\sigma_x$. The centered vectored matrix is the new data set D.

Step 4: The covariance matrix is given by S =YY^T. : The principal eigenvectors u_k of YY^T are the principal directions of the data Y. The principal eigenvectors v_k of the Gram matrix Y^TY are the principal components; entries of each v_k are the projected values of data points on the principal direction u_k. v_k and u_k are related via: $v_k = Y^Tu_k/\lambda_k^{1/2}$: where λ_k is the eigenvalue of the covariance matrix YY^T.

Step 5: Calculate the distance of each data points d_n and the k cluster centres c_k mostly preferred is the Euclidean distance.

Step 6: Repeat the following Steps 7-13 till a convergence criteria is met or we can say no new centroids are found.

Step 7: For each data object d_i, find the closest centroid c_j and assign d_i to the cluster with nearest centroid cj.

Step 8: Update the local and global position of each data points using the equation (1) and equation (2).

Step 9: Update the velocities and positions of each of the data points using equation (3) and equation (4).

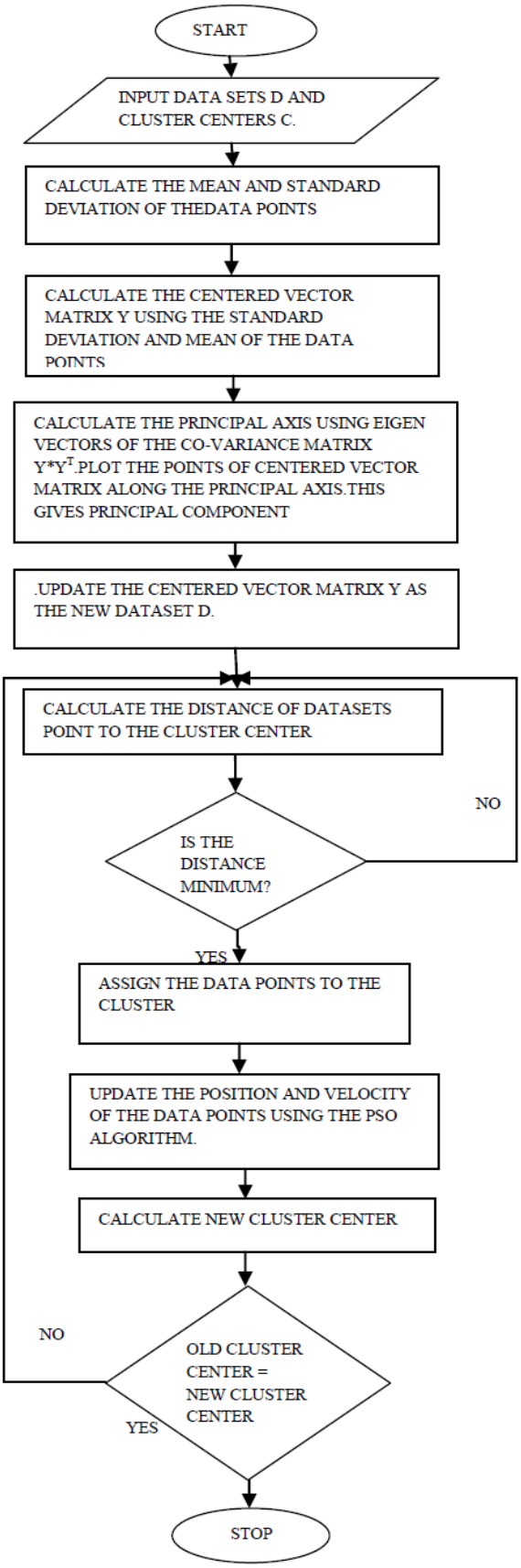
Step 10: For each cluster cj (1<=j<=k), recalculate the centroids.

Step 11: For each data points d_i compute its distance from the centroid c_j of the present nearest cluster.

Step 12: If the calculated distance is less than or equal to the previous calculated distance then the data points stay in the previous cluster.

Step 13: Else, calculate the distance of the data point to each of the new cluster centers and assign the data point to the nearest cluster based on the distances from the cluster centers.

VI. FLOW CHART OF PROPOSED ALGORITHM



Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

VII. EXPERIMENTAL ANALYSIS

In this experiment, an analysis is done on the behaviour of the clusters of different algorithms. In this experiment comparison study of different kinds of algorithm is done on the basis of inter cluster and intra cluster distances.

Here first inter cluster and intra cluster analysis is done on k-means algorithm. Then PCA is applied on the k-means algorithm, the algorithm derived as in [1] and [2] are used for the calculation of inter cluster and intra cluster distances. Finally inter cluster and inter cluster distances is done on the proposed algorithm.

The data sets used for above experimental analysis are given as:-Iris, Wine, Breast-cancer and Artificial.

The bar graph of the comparison of the parameters is plotted for the different algorithm (k-means, improved hybridised k-means [1] and [2] and proposed algorithm "A New Improved Hybridised K-MEANS Clustering Algorithm with Improved PCA Optimised with PSO for High Dimensional data set").

The clustering analysis of k-means using improved PCA and proposed algorithm was plotted.

Table 1:Intra Cluster distance Comparison

Data Sets	Intra Cluster Distance		
	K-means	Improved Hybridized K-means	Proposed Algorithm
Iris	3.489	3.374	3.304
Wine	4.911	4.202	4.199
Breast-cancer	7.285	6.599	6.551
Artificial	0.911	0.873	0.869

Table 2: Inter Cluster distance Comparison

Data Sets	Inter Cluster Distance		
	K-means	Improved Hybridized K-means	Proposed Algorithm
Iris	0.852	0.881	0.887
Wine	1.010	2.799	2.977
Breast-cancer	1.824	3.335	3.545
Artificial	0.796	0.814	0.8 15

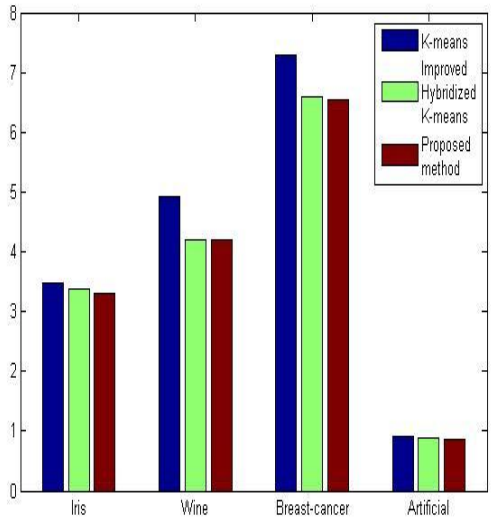


Fig 1::Intra cluster distance using Various Algorithms

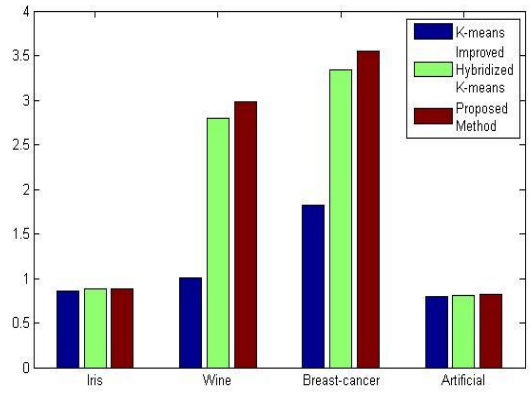


Fig 2: Inter Cluster distance using Various Algorithms

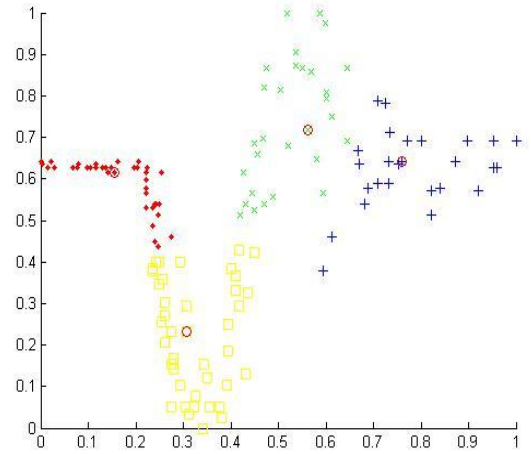


Fig 3: Improved hybridised k-means with improved PCA

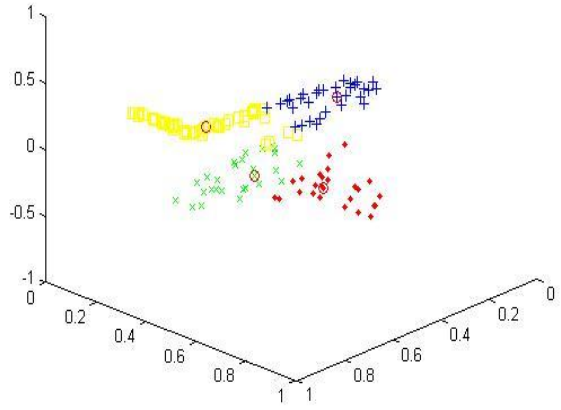


Fig 4: A New Improved Hybridised K-MEANS Clustering Algorithm with Improved PCA Optimised with PSO for High Dimensional data set

VIII. RESULT AND PERFORMANCE ANALYSIS

From the experimental analysis above we can see that the inter cluster and intra cluster distance for our proposed algorithm "A New Improved Hybridised K-MEANS Clustering Algorithm with Improved PCA Optimised with PSO for High Dimensional data set" is optimal than simple k-means and algorithm proposed by Dash et. al. [1] and Behera et. al [2] . Thus we can say that proposed algorithm "A New Improved Hybridised K-MEANS Clustering Algorithm with Improved PCA Optimised with PSO for High Dimensional data set" is more optimised than k-means and algorithm proposed by Behera et. al [2].

IX. CONCLUSION

Thus it can be concluded that combining PSO with improved PCA and k-means give more optimal clusters than k-means clustering algorithm. Experimental studies can be conducted with various optimising algorithm with various form of k-means and can be analysed to give more optimal results. Thus we can see that implementing improved PCA with PSO gives more optimal k-means with data reduction or dimensional reduction of the data sets.

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. H.S. Beheraet.al. "An Improved Hybridized K-Means Clustering Algorithm (IHKMCA) For High dimensional Dataset &its Performance Analysis International" Journal on Computer Science and Engineering (IJCSE) vol 3 no 3 march 2011
3. P.Prabhuet et al. "Improvising the performance of K-means clustering for high dimensional data set" International journal on computer science and engineering vol 3, Jun 2011
4. Dimensionality reduction: A comparative review", by Maaten L.J.P., Postma E.O. and Herik H.J. van den, Tech. rep.University of Maastricht ,2007.
5. Davy Michael and Luz Saturnine, 2007. "Dimensionality reduction for active learning with nearest neighbour classifier" in text categorization problems, Sixth International Conference on Machine Learning and Applications, pp. 292-297
6. Performance analysis of K-means with different initialization for high dimensional data" by Tanjunisha and Saravan International journal of Artificial Intelligence and application vol1 no.4, October 2010.
7. New method of dimensionality reduction using K-means clustering algorithm for high dimensional data set" by D Napoleon and S.Paralakodi international journal of computer science application vol13 no.7, January 2011.
8. An efficient method to improve clustering performance for high dimensional data by principal component analysis and modified K-means" by Tanjunisha and SaravanInternationaljournalofdatabase management system vol3 no.1, February 2011.
9. Auto-Clustering Using Particle Swarm Optimization and Bacterial Foraging".byJakob R. Olesen, Jorge Cordero H., and YifengZeng.Cao et al. (Eds.): ADMI 2009, LNCS 5680, pp. 69–83, 2009 Springer-Verlag Berlin Heidelberg 2009.
10. Particle Swarm Optimization Methods, Taxonomy and Applications" by DavoudSedighizadeh and EllipsMasehian, International Journal of Computer Theory and Engineering, Vol. 1, No. 5, December 20091793-8201.

AUTHORS PROFILE

Dr. H.S Behera is currently working as a Faculty in Dept. of Computer Science andEngineering is Veer Surendra Sai University of Technology(VSSUT), Burla, Odisha, India. Hisresearch areas of interest include Operating Systems, Data Mining , Soft Computing and Distributed Systems.

Mr.Abhishek Ghosh is a B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India.

Mr.Sipak ku. Mishra is a B. Tech. student in Dept. of Computer Science and Engineering, Veer Surendra Sai University of Technology (VSSUT), Burla, Odisha, India.