Perimeter Clustering Algorithm to Reduce the Number of Iterations

G.V.S.N.R.V.Prasad
Department of Computer Science & Engineering, Gudlavalleru Engg., College, Gudlavalleru, A.P,India

Dr Ch. Satyanarayana
Department of Computer Science and Engineering, J.N.T.U.K, Kakinada, A.P, India.

Dr V.Vijaya Kumar
Dept of Computer Science Engineering & Information Technology, G.I.E.T Rajamundry, A.P, India

ABSTRACT
Clustering is a division of data into groups of similar objects. Clustering is an unsupervised learning, due to its unknown label class in the search domain. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. It has capability to cluster large data. The main idea of K-Means is to define k centroids for each cluster. The K-means algorithm clusters the data with more complexity and the complexity further increases based on the dimensionality and data size. To overcome this we present a novel approach called perimeter K-means (PKM) clustering algorithms, which considers two data points and evaluates the perimeters. From this the two data pints are assigned to the nearest cluster center. By this the PKM reduces the overall complexity issues of K-means algorithms. The experimental result on various datasets, with various instances clearly indicates the efficacy of the proposed method. Further cluster quality and stability issues are tested by the proposed PKM.

Keywords
DataMining,clustering,similarity,stability.

1. INTRODUCTION
Partitioning a large dataset of objects into homogeneous clusters is a fundamental operation in data mining. Clustering problems arise in many different applications, such as data-mining and knowledge- discovery [6], data-compression /reduction and vector quantization [7], and pattern recognition and pattern classification [4]. A good cluster depends on a proper application and there are many methods for finding clusters subjected to various criteria, both adhoc and systematic. These include approaches based on splitting and merging such as ISODATA [1,10], randomized approaches such as CLARA [11], CLARANS [16], methods based on neural nets [12], and methods designed to scale to large databases, including DBSCAN [5], BIRCH [19] and ScaleKM [3].

One of the most widely used and studied clustering forms which are based on minimizing a formal objective function is K-means- clustering. Given a set of n data points in real d-dimensional space, Rd, and an integer k, the problem is to determine a set of k points in Rd called centers, so as to minimize the mean squared distance from each data point to its nearest center. This measure is often called the squared-error distortion [7, 10] and this type of clustering falls into the general category of variance-based clustering [8,9]. A direct implementation of the k-means method is computationally very intensive. This is especially true for typical data mining applications with a large number of pattern vectors. Most of the previous studies concentrated on various similarity or dissimilarity measures, dimensional reduction and various data structures like KD-trees are used to find the nearest data-point from the center. The above studies did not concentrate much on the reducing number of iterations over dataset which is also the main component in reducing the complexity of any clustering algorithm.

The K-means clustering algorithm was implemented in Introduction Detection system [20.21].This reduces time for detection and also reduces false positives

After a careful and thorough study on the existing literature the present study has outlined a novel approach to reduce the number of iterations over dataset which is one of the main factors that influence the complexity of partitional clustering algorithms. In this scheme the number of iterations is reduced by selecting two data-points, instead of one in the K-means clustering algorithm and a perimeter is evaluated by taking the centre point as the third point. This scheme is known as Perimeter K-means (PKM) clustering algorithm

The paper is structured as follows. In Section 2, we describe our Perimeter K-Means (PKM) clustering algorithm in detail and illustrate the algorithm with an example. In Section 3, we describe the details of our experiment and present the results graphically. The section four deals with conclusions and some possible future directions of the investigation.

2. PERIMETER K-MEANS (PKM) CLUSTERING ALGORITHM
The simple K-means clustering algorithm randomly picks one data point from a dataset and calculates the distance between cluster centers. This data-point is assigned to the cluster that is having least distance from its center.

The proposed PKM methodology considers two data points from input dataset and calculates the perimeter by calculating the distance between points and center of clusters. These two points are assigned to the respective

cluster having least perimeter between the center and two data points. The proposed PKM clustering algorithm picks two data points at a time and hence it reduces the number of scans over dataset, i.e., the number of iterations over dataset. This reduces the overall complexity of the clustering process in terms of number of iterations. The proposed PKM algorithm contains seven steps, as given below.

Begin
Step 1: Choose K number of clusters.
Step 2: Assume K number of initial seed points.
Step 3: Randomly assign the data into K (non-empty) clusters. Determine the center of each cluster by averaging the observations in the cluster and update the initial centroids ith new centroids.
Step 4: Considers two data points from the dataset and calcu- -late the perimeter between data-points and cluster-centroid(s) of all clusters and assign the two points to that cluster for which the perimeter is comparatively small.
Step 5: Compute new centroids after assigning all data points to k clusters.
Step 6: Repeat steps 4 and 5 until the difference between the previous and current centroids is less than the specified threshold value.
Step 7: Repeat steps 2 to 6 with different initial seed points until the algorithm reaches the minimum objective function
End

Let D1, D2 be the data points from the set D. Let C1, C2 be two initial cluster centroids assumed initially of clusters C1 and C2. Based on this the proposed PKM calculates the perimeters of the ΔC1D1D2 and ΔC2D1D2. If the perimeter of ΔC1D1D2 is less than the perimeter of ΔC2D1D2, then the two points D1, D2 are to be assigned to the cluster C1 or else to cluster C2. With this procedure the proposed PKM approach assigns two points D1, D2 to the nearest cluster center instead of assigning only one data point as in the basic K-means clustering algorithm.

The proposed PKM clustering algorithm evaluates similarity measures based on the Euclidean distance that satisfies the tetrahedral inequality in two-metric space. The two-metric is a function d(x, y, z), which is symmetric under permutations, satisfying the tetrahedral inequality and positive definiteness as given in equation (1)

$$d(x, y, z) \leq d(x, y, a) + d(x, a, z) + d(a, y, z) \quad (1)$$

for all the items $x, y, z, a \in X$.

Where d(x, y, z) is also called the G-metric space, which is the area of the triangle spanned by x, y, z. In G-metric space [14,15], the tetrahedral inequality is replaced by an inequality involving repetition of indices by which function d(x, y, z) is thought of as representing the perimeter of a triangle.

3. EXPERIMENT ANALYSIS OF PKM CLUSTERING ALGORITHM

To establish the practical efficiency of the proposed algorithms, the algorithm is tested using synthetic dataset of 1000 instances, 10000 instances, generated using a Gaussian distribution around the centers. The Gaussian type construction function treats data points neighbouring each other which is in a spherical shape and for the purpose of clustering intuitively. The major advantage of adopting the Gaussian type function is that all partitional clustering algorithms are able to cluster in spherical shapes qualitatively rather than in arbitrary shapes.

The algorithms are also tested using the most popular multi-dimensional real datasets like Breast Cancer dataset, Iris dataset, Intrusion Detection (ID) dataset and 8000 instances dataset. The proposed PKM is measured with respect to complexity, quality and stability of clusters. The efficiency of PKM is also compared with K-means by using the above measures.

3.1 Measure of complexity using PKM

The proposed PKM algorithm runs considerably less number of times over the dataset for both uni-dimensional and multi-dimensional datasets of moderate as also for large dataset when compared to the original K-means algorithm; thus it reduces the overall complexity. The graphs are plotted for the number of clusters versus the number of iterations for the above dataset. The graphs (Figures 1 to 6) clearly indicate that the proposed PKM algorithm takes a less number of iterations when compared to K-means algorithm for the synthetic dataset of 1000 instances, 10000 instances, real datasets like Breast Cancer dataset, Iris dataset, ID dataset and 8000 instances of dataset respectively. From this it is evident that the proposed PKM clustering algorithm is clustering large and multidimensional data with less complexity by reducing the number of iterations over dataset.

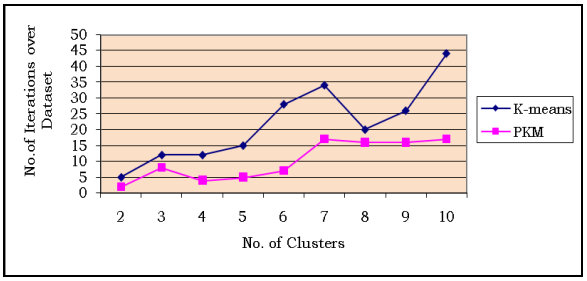

Figure 1: Performance Comparison for Synthetic One Dimensional 1000 Instances Dataset.

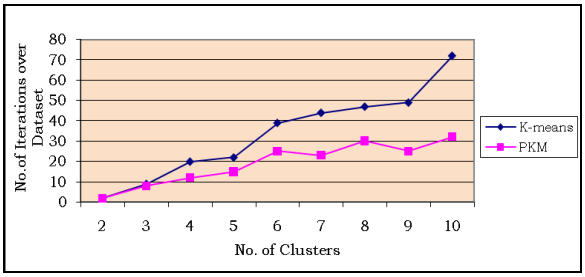

Figure 2: Performance Comparison for Synthetic One Dimensional 10000 Instances Dataset.

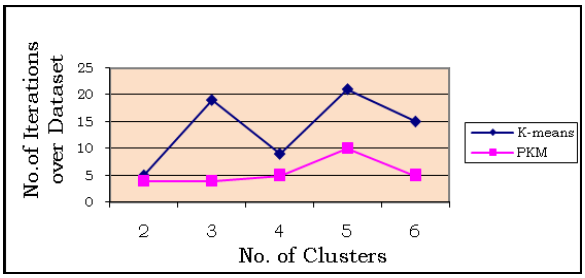

Figure 3: Performance Comparison for Breast Cancer Dataset.

Figure 4: Performance Comparison for Iris Dataset.

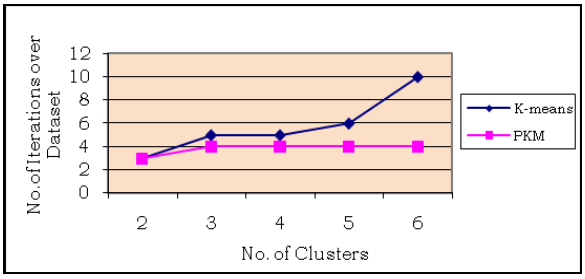

Figure 5: Performance Comparison for Intrusion Detection Dataset.

Figure 6: Performance Comparison for 8000 Instances Dataset.

3.2 Measure of quality of clusters using PKM

The popular K-means algorithm uses the sum of squared error function as an objective function to indicate the cluster quality. The minimum value of the sum of squared error function is an indication for good quality of clustering procedure. The sum of squared error function j is given by the equation. 2

$$J = \sum_{j=1}^{k} \sum_{i=1}^{x} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (2)$$

Where $\left\| x_i^{(j)} - c_j \right\|$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center. This distance is an indicator of the distance of the n data points from their respective cluster centers. Cluster quality graphs that indicate the number of clusters versus error-function calculated from equation 2, are plotted in Figures 7 to 12 for the synthetic dataset of (a)1000 instances, (b)10000 instances, real dataset (c)Breast Cancer dataset, (d)Iris dataset, (e)ID dataset and (f)8000 instances of dataset respectively. The cluster quality of the proposed PKM scheme is significantly either different or similar when compared to K-means algorithm for Breast Cancer dataset, Iris dataset, ID dataset and 8000 instances of dataset as shown in Figures 9 to 12 respectively.

The graphs 7 to 12 on different data sets with different instances indicates that the proposal PKM maintains the cluster quality as K-means algorithm even by picking up to points randomly.

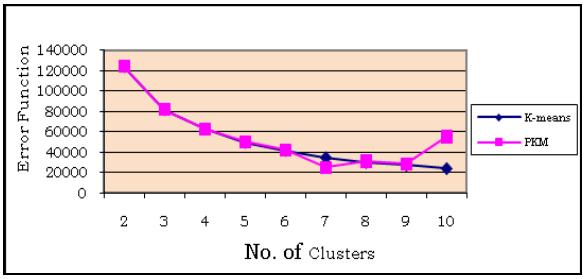

Figure 7: Comparison of Cluster Quality in One Dimensional Synthetic 1000 Instances Dataset.

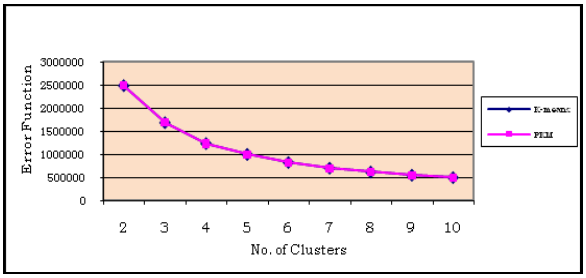

Figure 8: Comparison of Cluster Quality in One Dimensional Synthetic 10000 Instances Dataset.

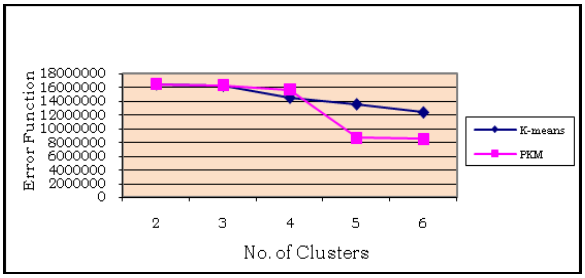

Figure 9: Comparison of Cluster Quality in Breast Cancer Dataset.

Figure 10: Comparison of Cluster Quality in Iris Dataset.

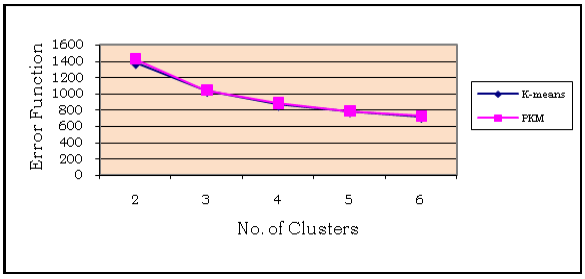

Figure 11: Comparison of Cluster Quality in Intrusion Detection Dataset.

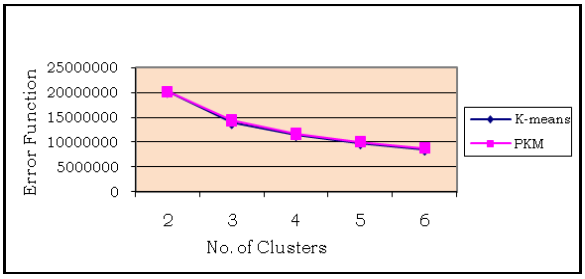

Figure 12: Comparison of Cluster Quality in 8000 Instances Dataset.

3.3 Measure of stability of clusters using PKM

Validation is very important in cluster analysis, because clustering methods tend to generate clusters even for fairly homogeneous datasets. Most clustering methods assume a certain model or prototype for clusters, and this may be adequate for some parts of a data, but not for others. Cluster analysis is often carried out in an exploratory manner, and the patterns found by cluster analysis are not necessarily meaningful. An important aspect of cluster validity is stability. Stability means that "a meaningful valid cluster should not disappear easily if the dataset is changed in a non-essential way". Stability in cluster analysis is strongly dependent on the dataset, especially on how well-separated and how homogeneous the clusters are. In the same clustering, some clusters may be very stable and others may be extremely unstable.

Many authors have differed in taking data sub-sampling. For instance, Ben-Hur [2] randomly chose overlapping portions of the data and evaluated the distance between the resulting clustering solutions on the common samples. Lange [13], on the other hand, divided the sample into disjoint subsets. Similarly, Ben-David [17, 18] studied stability with respect to complete change of the data. These different approaches of choosing K have prompted the present study to give a precise characterization of clustering stability with respect to partial changes of the data.

The simplest idea is to draw a subsample of X_n without replacement. This avoids multiple points and shortens computation times, which can be a big issue with large datasets. Sub-setting requires choice of the size m<n of the subsample. If number of data points m is too large, sub-setting will not generate enough variation to be informative. If m is too small, the clustering results can be expected to be much worse than that obtained from the original dataset. To avoid these problems with m the present paper considered intelligently m as(n/2). This is shown in the form of subsets from 1 to 10 in the following figures from 13 to 20.

The present paper investigates stability analysis on the proposed PKM and K-Means algorithms for K=2,3,4 and 5 for Iris dataset, represented in Figures 13 to 20 which consider the number of iterations and cluster quality.

3.3.1 Stability in performance using number of iterations

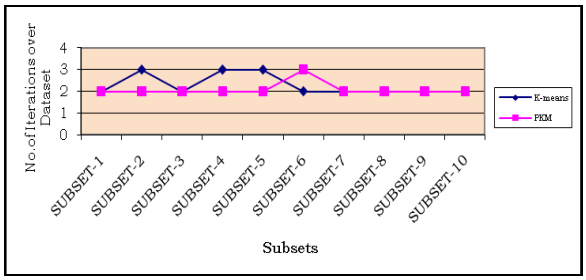

Figure 13: Comparison for Stability in Performance of Iris Dataset at K=2.

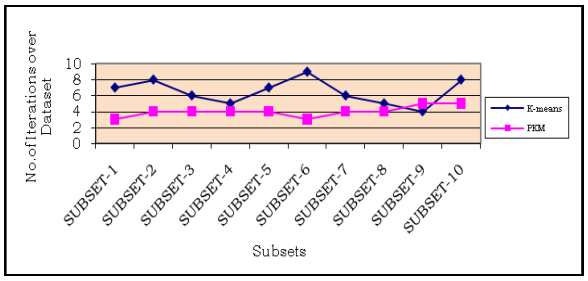

Figure 14: Comparison for Stability in Performance of Iris Dataset at K=3.

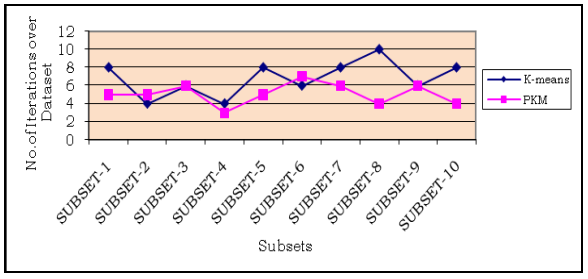

Figure 15: Comparison for Stability in Performance of Iris Dataset at K=4.

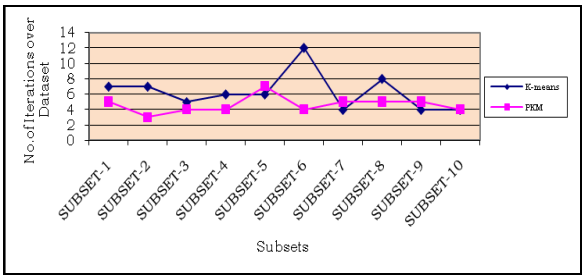

Figure 16: Comparison for Stability in Performance of Iris Dataset at K=5.

3.3.2 Stability in cluster quality

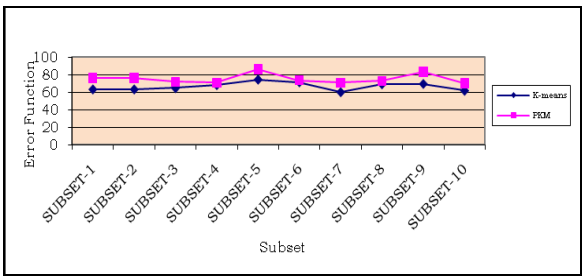

Figure 17: Comparison for Stability in Quality of Iris Dataset at K=2.

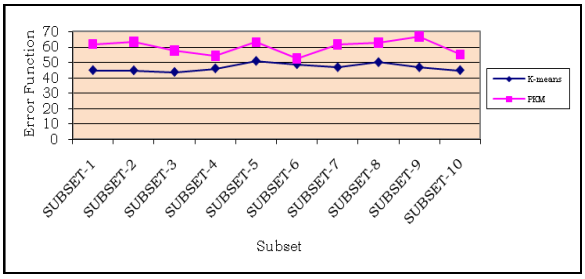

Figure 18: Comparison for Stability in Quality of Iris Dataset at K=3.

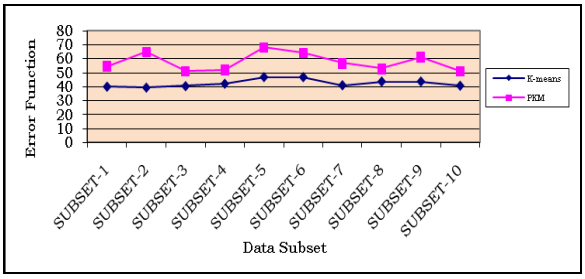

Figure 19: Comparison for Stability in Quality of Iris Dataset at K=4.

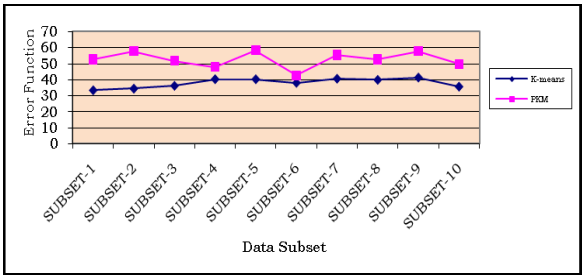

Figure 20: Comparison for Stability in Quality of Iris Dataset at K=5.

The uniformity in the graphs (Figures 13 to 20) clearly indicates higher or the same stability level for the proposed PKM scheme when compared to K-means scheme.

4. CONCLUSION

Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed and are available in the literature. The proposed PKM clustering algorithm clusters the data with less complexity when compared to K-means clustering algorithm by reducing the number of iterations over dataset. The proposed PKM clustering algorithm selects two data points instead of one point as in the case of K-means. The novelty of this algorithm is it reduces the number of iterations over large multi dimensional datasets with an increase in K (number of clusters) value when compared to K-means which has a linear relation between K-value and the number of iterations. This is rigorously tested with various one-dimensional, multi-dimensional synthetic datasets and also using popular real datasets of multi-dimensional, moderate and large dataset sizes. The qualities of clustering algorithms are verified using K-means objective function which is an error function. The quality of the proposed PKM algorithm is similar to K-means for some datasets and an improvement for some datasets. The stability of clustering algorithms plays a vital role in assessing the performance of the clustering algorithms. This stability is verified using the most popular sub-setting method and results show that the proposed PKM clustering produces more stable clusters when compared to K-means algorithm. This proposed PKM clustering algorithm can be implemented in segmentation of images.

5. REFERENCES

[1] Ball.G.H. and Hall.D.J., "Some Fundamental Concepts and Synthesis Procedures for Pattern Recognition Preprocessors," Proc. Int'l Conf. Microwaves, Circuit Theory, and Information Theory, Sept. 1964.

[2] Ben-Hur.A., Elisseeff.A. and Guyon.I. "A stability based method for discovering structure in clustered data," In Pasific Symposium on Bio-computing, Vol. 7, pp 6–17, 2002.

[3] Bradley.P., Fayyad.U. "Scaling clustering algorithms to large databases," KDD-98, 1998.

[4] Duda.R.O. and Hart.P.E. "Pattern Classification and Scene Analysis," New York: John Wiley & Sons, 1973.

[5] Ester.M., Kriegel.H. and Xu.X. "A Database Interface for Clustering in Large Spatial Databases," Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD-95), pp. 94-99, 1995.

[6] Fayyad.U.M., et.al. "Advances in Knowledge Discovery and Data Mining," AAAI/MIT Press, 1996.

[7] Gersho.A. and Gray.R.M. "Vector Quantization and Signal Compression," Boston: Kluwer Academic, 1992.

[8] Inaba.M., Imai.H. and Katoh.N. "Experimental Results of a Randomized Clustering Algorithm," Proc.12th Ann. ACM Symp. Computational Geometry, pp. C1-C2, May 1996.

[9] Inaba.M., Katoh.N. and Imai.H. "Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based clustering," Proc. 10th Ann. ACM Symp. Computational Geometry, pp. 332-339, June 1994.

[10] Jain.A.K. and Dubes.R.C. "Algorithms for clustering Data", Prentice Hall, 1988.

[11] Kaufman.L. and Rousseeuw.P.J. "Finding Groups in Data: An Introduction to Cluster Analysis," New York: John Wiley & Sons, 1990.

[12] Kohonen.T., "Self-Organization and Associative Memory," third ed. New York: Springer-Verlag, 1989.

[13] Lange.T., Braun.V., Roth.V. et al. "Stability based model selection," In NIPS, 2003.

[14] Mustafa.Z. and Sims.B. "A new approach to generalized metric spaces," J. of Nonlinear and Convex Analysis 7, pp. 289–297, 2006.

[15] Mustafa.Z. and Sims.B. "Fixed point theorems for contractive mappings in complete G-metric spaces," Fixed Point Theory Appl, Art. ID 917175, pp. 10, 2009.

[16] Ng.R. and Han.J. "Efficient and effective clustering methods for spatial data mining," VLDB-94, 1994.

[17] Shai Ben-David., Ulrike von et al. "A sober look at clustering stability," In COLT, 2006.

[18] Ulrike von Luxburg and Shai Ben-David. "Towards a statistical theory of clustering," PASCAL Workshop on Statistics and Optimization of Clustering, 2005.

[19] Zhang.T., Ramakrishnan.R. and Livny.M. BIRCH: "A New Data Clustering Algorithm and Its Applications," Data Mining and Knowledge Discovery, Vol. 1, no. 2, pp. 141-182, 1997.

[20] Prof. G.V.S.N.R.V.Prasad, Prof.Y. Dhanalakshmi, Prof. V.Vijaya Kumar and Prof. I.Ramesh Babu **"**Modeling An Intrusion Detection System Using Data Mining And Genetic Algorithms Based On Fuzzy Logic", International Journal of Computer Science and Network Security, IJCSNS, VOL.8 No.7, July 2008.

[21] Prof. G.V.S.N.R.V.Prasad, Prof.Y. Dhanalakshmi, Prof.V.Vijaya Kumar and Prof. I.Ramesh Babu "Mining for optimized data using clustering along with fuzzy association rules and genetic algorithms",International Journal Of Artificial Intelligence & Applications (IJAIA),Vol. No. 2,April 2010.

6. AUTHORS PROFILE

G.V.S.N.R.V.Prasad did his MS Software Systems, BITS Pilani and M.Tech in Computer Science and Technology in Andhra University .He has 15 years of teaching experience. Published 7 Research Papers in various National and International Conferences and 3 Research papers in National and International Journals. He is a member in various Professional Bodies . Presently working as Professor in CSE at Gudlavalleru Engineering College , Gudlavalleru ,A.P. His area of interest is Data Mining, Network Security and Image Processing.

Ch. Satyanarayana is working as an Associate Professor in the Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Kakinada, India. He received B. Tech (CSE) in 1996 and M.Tech (CST) in 1998, from Andhra University. Ph.D in Computer Science. He has been working in Jawaharlal Nehru Technological University Kakinada from the past 10 years. His research areas of interest are Pattern Recognition, Image Processing, Speech Processing, Computer Graphics and Compiler writing. He had supervised 92 M.Tech and 65 MCA projects. He has published 35 papers in international conferences and Journals. He is a member of different professional bodies like ISTE, IETE and CSI.

Prof.Vijaya Kumar did his MS Engineering in Computer Science [USSR –TASHKENT STATE UNIVERSITY] and Ph.D in Computer Science . Worked as Associate Professor in Department of CSE and School of Information Technology (SIT) at Jawaharlal Nehru Technological university (JNTU) Hyderabad. Having a total of 13 years of experience. He Published 60 Research Papers in various National and International Conferences /Journals. Guiding 10 Research scholars . He is a Member for various National and Inter National Professional Bodies.Presently working as Dean for CSE & IT at GODAVARI INSTITUTE OF ENGINEERING AND TECHNOLOGY Rajamundry .