



A Genetic Algorithm for K-Mean Clustering

Varsha Singh

Asst. Prof. JSSATE, Noida, Uttar Pradesh, India

Prof A K Misra

Professor, Deptt of CSE,

MNNIT Allahabad, Uttar Pradesh, India

Abstract: Clustering techniques have obtained adequate results when are applied to data mining problems. Clustering is the process of subdividing an input data set into a desired number of subgroups so that members of the same subgroup are similar and members of different subgroups have diverse properties. Many heuristic algorithms have been applied to the clustering problem, which is known to be NP Hard. Genetic algorithms have been used in a wide variety of fields to perform clustering, however, the technique normally has a long running time in terms of input set size. In this paper we investigate the use of Genetic Algorithms to determine the best initialization of clusters, as well as the optimization of the initial parameters. The genetic algorithm uses the most time efficient techniques along with preprocessing of the input data set. The experimental results show the great potential of the Genetic Algorithms for the improvement of the clusters. The techniques of clustering are most used in the analysis of information or Data Mining, this method was applied to Data Set at mining

I. INTRODUCTION

In the area of clustering, there are many tasks involved in the process of identifying a cluster, which need appropriate parameter selection and definite criteria in complex spaces in order to obtain optimum solutions. Therefore, the application of Genetic Algorithms for solving certain problems of clustering (which need optimization of computational requirements, and robust, fast and close approximate solution) appears to be appropriate and natural. Recently, an application of Genetic Algorithms has been reported in the area of clustering [14][24][19][2]. When the only data available are unlabeled, the classification problems are sometimes referred to as unsupervised classification. Clustering is an important unsupervised classification technique where a set of patterns, usually vectors in a multidimensional space, are grouped into clusters in such a way that patterns in the same cluster are similar in some sense and patterns in different clusters are dissimilar in the same sense [52]. Since our aim is to propose a clustering technique based on Genetic Algorithms, a criterion is required whose optimization would provide the final clusters. However, unlike the K-means algorithm, which may get stuck at values, which are not optimal [47], the proposed technique should be able to provide good results irrespective of the starting configuration.

II. Genetic Algorithm

The genetic algorithm is a search algorithm based on the mechanics of natural selection and natural genetics [41]. As summarized by Tomassini [25], the main idea is that in order for a population of individuals to adapt to some environment, it should behave like a natural system. This means that survival and reproduction of an individual is promoted by the elimination of useless or harmful traits and by rewarding useful behavior. The genetic algorithm belongs to the family of evolutionary algorithms, along with genetic programming, evolution strategies, and evolutionary programming. Evolutionary algorithms can be considered as a broad class of stochastic optimization technique [58]. An evolutionary algorithm maintains a population of candidate solutions for the problem at hand. The population is then evolved by the iterative application of a set of stochastic operators. The set of operators usually consists of mutation, recombination, and selection or something very similar. Globally satisfactory, if sub-optimal, solutions to the problem are found in much the same way as populations in nature adapt to their surrounding environment.

The evolutionary cycle can be summarized as follows [25]:

```
generation = 0
seed population
while not (termination condition) do
generation = generation + 1
```

calculate fitness
selection
crossover
mutation
end while

III. CLUSTERING

Fast retrieval of the relevant information from the databases has always been a significant issue. Different techniques have been developed for this purpose; one of them is Data Clustering. Some applications of Data Clustering are Data Mining using Data Clustering, pattern recognition and marketing research. Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined.

A. K-Means Clustering Algorithm [59]

k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-means clustering is an effective algorithm to extract a given number of clusters of patterns from a training set. Once done, the cluster locations can be used to classify patterns into distinct classes.

B. Clustering Problem and the k-means Algorithm

According to [53], clusters analysis aims at solving the following very general problem: given a set X of N entities, often described by measurements as points of the real d -dimensional space, find subjects of X which are homogeneous and/or well-separated. Homogeneity means that entities in the same cluster must be similar and separation between entities in the same cluster must be similar and separation between entities in different clusters must differ one from the other. These concepts can be made precise in a variety of ways, which lead to as many clustering problems and even more heuristic or exact algorithms, so clustering is a vast subject. The work of [16] provides a good survey on the issue of clustering and provides a set of key references. Among the clustering algorithms based on minimizing an objective function or the squared error, perhaps the most widely used and studied is called the k-means algorithm. This algorithm has been discovered by several research across different disciplines, most notably [49], [54] and [56].

C. Algorithm K-means Advantages

The process, which is called “k-means”, appears to give partitions which are reasonably efficient in the sense of within-class variance, corroborated to some extent by mathematical analysis and practical experience [54]. Also, the k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer. K-means algorithm is one of first which a data analyst will use to investigate a new data set because it is algorithmically simple, relatively robust and gives “good enough” answers over a wide variety of data sets.

D. Algorithm K-means Shortcomings

The algorithm's sensitivity to initial conditions: The number of partitions, the initial centroids. According to [16] there is a universal and efficient method to identify initial patterns and the number k of clusters. In [26] briefly is discussed the sensitivity of the algorithm for the allocation of initial centroids, that in practice the usual method is to test iteratively with a random allocation to find the best allocation in terms of minimizing the total squared distance. However, there have been various investigations aimed at making various proposals related to these limitations [1, 3, 4, 5, 10, 11, 17, 21, 28, 30, 40, 53, and 55].

Clustering using Genetic algorithm is the proposed algorithm for:

1. Finding best k cluster centers
2. Applying k-mean for storing the k cluster center i.e. superimposing GA over k-Means.

In this algorithm, the first step is to randomly generate an initial population of size 20 in which each individual (cluster center) is randomly initialized. After generating population the k -cluster center is found by using Genetic Algorithm. Then simple Euclidean distance is used to assign the remaining instances to their closest cluster center and then mean value for each cluster is calculated. Each individual is evaluated according to its fitness function. In this, the concept of Elitism for individual selection is used.

Elitism is the process of selecting a better individual or more to the point-selecting individual with a biased towards the better ones. Elitism is important since it allows the solution to get better overtime and allows the population to converge quicker this means all the individuals will more or less all be the same. Contrarily to general belief, the solution will not necessary be optimal. This approach is fine for small problem and small

population size. For bigger problems more diversity is needed. This allows the genetic algorithm to search through a wider variety of solutions, and there by not get stuck in local maxima. This may be achieved by adding a random parameter to the elitist selection scheme. For example, use a tournament selection scheme, which involves comparing random numbers scaled by the magnitude of the fitness.

D.1 String representation for GA

Each string is a sequence of real numbers representing the K cluster centers. For an N-dimensional space, the length of a chromosome is N*K words, where the first N positions (or, genes) represent the N dimensions of the first cluster center, the next N positions represent those of the second cluster center, and so on.

D.2 Population initialization

The k cluster centers encoded in each chromosome are initialized to k randomly chosen points from the data set. This process is repeated for each of the P chromosomes in the population, where P is the size of the population.

D.3 Fitness computation

In the fitness computation process, the constant number k is generated randomly which is considered as the number of clusters to be formed and the clusters are then formed according to the centers encoded in the chromosome under consideration.

This is done by calculating the sum of the distance of each data from each cluster center in a cluster and this should be done for each individual. After this, by comparing the sum of each individual the individual having minimum sum is identified which means that the sum of all clusters in that individual is minimum and this will give us the fitness function. For example, suppose there are three clusters i.e. C_1, C_2, C_3 , which are formed according to the centers, encoded in the chromosome i.e. individuals. Data having minimum distance from the cluster center is assigned to the corresponding cluster C_i . Then the sum of distances between the data and the cluster center for each cluster is computed as follows where i vary from 1 to L/2 where L represents the length of chromosomes :

$$S_i = \sum_{j=1}^m (\text{sum of distances between } D_j \text{ and cluster center } C_i) \quad [3.1]$$

Where D_j represents the data in cluster C_i and “m” is the number of data in cluster C_i .

And then we compute

$$S = \sum_{i=1}^{L/2} S_i \quad [3.2]$$

for each individual and as the minimum value of S represents the fitness value therefore the individual having minimum value of S selected as the fittest individual in the population.

D.4. Selection

The selection process selects chromosomes from the mating pool directed by the survival of the fittest concept of natural genetic systems. A chromosome is assigned a number of copies according to the selection process in which the chromosomes having minimum fitness value (which is proportional to its fitness in the population) goes into the mating pool for further genetic operations. For this, Roulette wheel selection process is used.

D5. Crossover

Crossover is a probabilistic process that exchanges information between two parent chromosomes for generating two child chromosomes. In this Paper single point crossover with a fixed crossover probability that is k_c is used. For chromosomes of length L, a random integer, called the crossover point, is generated in the range [1, L-1]. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offspring.

D.6. Mutation

Each chromosome undergoes mutation with a fixed probability k_m that is called as mutation probability. For binary representation of chromosomes, a bit position (or gene) is mutated by simply flipping its value. Since we are considering floating point representation in this Paper, we use the following mutation.

A number d in the range [0, 1] is generated with uniform distribution. The value at a gene position is X, then if the value of $d < 0.5$ then after mutation it becomes X-1 otherwise it becomes X+1.

D.7. Termination criterion

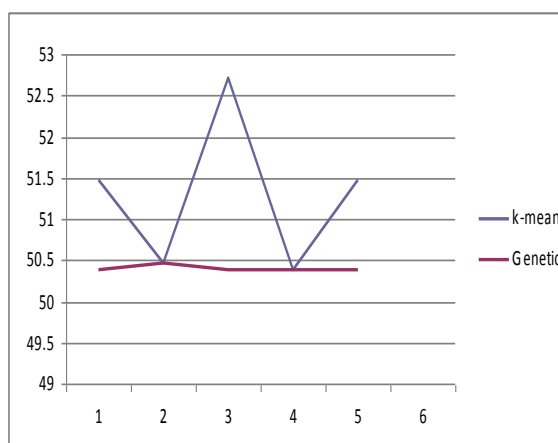
In this Paper the processes of fitness computation, selection, crossover, and mutation are executed for a maximum number of iterations. As twenty five generations are assumed here therefore the process terminates after twenty five generations. The best string seen up to the last generation provides the solution to the clustering problem. The elitism has been implemented at each generation by preserving the best string seen up to that generation in a location outside the population. Thus on termination, this location contains the centers of the final clusters. The next section provides the results of implementation of the GA-clustering algorithm, along with its comparison with the performance of the K-means algorithm for several artificial and real-life data sets.

IV. RESULT AND DISCUSSION

In this Paper, clustering using genetic, we are removing drawback of k-mean, which is to random selection of chromosomes of variables. Here we are using genetic with three operators which are selection, crossover, mutation. In the taken data of genetic we assume the population size 20 and number of generation 25. We are implementing both algorithms i.e. K-Mean and K-Mean using Genetic Algorithm in MAT lab and the corresponding results are shown below.

A. Comparison between the k-mean and Genetic Algorithm:

In the below graph we can see the result of k-mean alter after each iteration in large variation. And the result of Genetic is not vary as K-mean gives straight line in graph. Its show that the genetic clustering algorithm gives minimum and global result.



Comparison between K-Mean and K-mean using Genetic Algorithm(x-axis represents the number of O/P taken and Y-axis represents the minsum that is fitness value)

V. RESULTS COMPARISON BETWEEN THE DIFFERENT APPROACHES

Different approaches	Data 1(Avg. error)
K-means	1.34036
GA-clustering with 25 iterations	0.02
GA-clustering with 100 iterations	0

In the above table, the result comparison between the different approaches is shown in which Avg Error (the average of the difference between the outputs taken) is compared for the K-mean, GA-clustering with 25 iterations and GA-clustering with 100 iterations. We found that the Avg Error for K-Mean comes 1.34036 and if we use GA clustering with 25 iterations then Avg error will come .01945 and if we increase the iterations from 25 to 100 then it becomes zero.

VI. CONCLUSION & FUTURE WORK

Genetic algorithm-based clustering algorithm has been developed in the present work. Genetic algorithm has been used to search for the cluster centers which minimize the fitness value (minsum) for the

individual. In order to demonstrate the effectiveness of the GA-clustering algorithm in providing optimal clusters, two artificial data sets with the number of clusters ranging from 2 to 2000 have been considered. The results show that the GA-clustering algorithm provides a performance that is significantly superior to that of the K-means algorithm, a very widely used clustering technique..

This Paper can be used for designing genetic algorithm tools for various problems of Optimization. Floating-point representation of chromosomes has been adopted in this, since it is very closer to the problem space and provides easy way of mapping from the encoded cluster centers to the actual ones. Instead of floating-point representation of chromosomes, a binary representation may be implemented for the same, and the results may be compared with the present representation form.

REFERENCES

- [1] Zhang, Chen; Xia Shixiong. "K-means Clustering Algorithm with improved initial Center." Knowledge Discovery and Data Mining, 2009. Second International Workshop on. Vol. Issue, 23-25 Jan., 2009.
- [2] Dong-Xia, Xian-Da and Chang-Wen "A genetic algorithm with gene rearrangement for K-means clustering," Pattern Recognition Volume 42, Issue 7, July 2009, Pages 1210-1222
- [3] Bahmani, B., Firouzi, T. Niknam, and M. Nayeripour. "A New Evolutionary Algorithm for Cluster Analysis". Proceedings of world Academy of Science, Engineering and Technology Vol. 36, Dec.2008.
- [4] Wesan, Barbakh And Colin Fyfe. "Local vs global interactions in clustering algorithms: Advances over K-means." International Journal of knowledge-based and Intelligent Engineering Systems 12 (2008).83 – 99.
- [5] Kao, Yi-Tung, Zahara, Erwie, Kao, I-Wei. "A hibridized approach to data clustering". Expert Systems with Applications. Vol. 34 Issue 3. P 1754-1762. Apr.2008.
- [6] Belal Al-Zoubi, Al-Zoubi, Amjad Hudaib, Ammar Huneiti and Bassam Hammo. "New Efficient Strategy to Accelerate k-Means Clustering Algorithm". American Journal of Applied Sciences 5(9) 1247-1250, Science Publications. 2008
- [7] Taoying Li & Yan Chen "An improved k-means algorithm for clustering using entropy weighting measures". Intelligent Control and Automation, 2008, WCICA 2008, 7th World Congress on. June 2008.
- [8] Tsai, Chieh-Yuan, Chiu, Chuang-Cheng. "Developing a feature weight selfadjustment mechanism for a K-means clustering algorithm." Computational Statistics & Data Analysis. Vol. 52 Issue 10. Jun. 2008.
- [9] Zalik, Krista Rizman . "An Efficient k'-means Clustering Algorithm." Pattern Reconition Letters, Vol. 29, I.9. Pag. 1385-1391. Elsevier 07/2008
- [10] Deelers S. And S. Auwatanamongkol. "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance." Proceedings of world Academy of Science, Engineering and Technology. Vol. 26, Dec. 2007.
- [11] Redmond, Stephen J., Heneghan, Conor. "A method for initialising the K-means clustering algorithm using kd-trees". Pattern Recognition Letters; Vol. 28 Issue 8, Jun. 2007.
- [12] Pun, W.K.D., Ali, A.S. "Unique distance measure approach for K-means (UDMA-Km) clustering algorithm". TENCON 2007 – 2007 IEEE Region 10 Conference. Oct. 30 2007
- [13] Ding Zejin, Jian Yu, Yang-Qing Zhang. "A New improved K-Means Algorithm with Penalizaed Term". Granular Computing, 2007, GRC 2007, IEEE International Conference on. Nov. 2007.
- [14] Michael Laszlo and Sumitra Mukherjee "A genetic algorithm that exchanges neighboring enters for k-means clustering," Pattern Recognition Letters Volume 28, Issue 16 (December 2007) Pages 2359-2366 ISSN: 0167-8655
- [15] Zhang, Z., B. Tian D. And Tung A.K.H. "On the Lower Bound of Local Optimums in K-means Algorithm." Data Mining 2006, ICDM'06 Sixth International Conference on Data Mining. Dec. 2006
- [16] Xu, Rui and Donald Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, Vol., 16, No. 3, May 2005
- [17] Pham, D.T. Dimov, S.S. Nguyen, C.D. "Selection of K in K-means clustering". "Proceedings of the Institution of Mechanical Engineers – Part C – Journal of Mechanical Engineering Science; Vol. 219 Issue 1, p103-109. Jan 2005.
- [18] Xu, Rui and Donald Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, Vol., 16, No. 3, May 2005.
- [19] C. C. Lai, "A novel clustering approach using hierarchical genetic algorithms," Intelligent Automation and Soft Computing, vol. 11, no. 3, 143-153, 2005.
- [20] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown, "FGKA: A Fast Genetic K-Means Clustering Algorithm," ACM Symposium on Applied Computing, 2004
- [21] Likas, A., Vlassis, N., Verbeek, J.J.: "The Global K-means Clustering Algorithm. Pattern Recognition." The Journal of the Pattern Recognition Society. Vol. 36, No. 2 (2003) 451-461
- [22] Mehmed, K.: "Data Mining: Concepts, Models, Methods, and Algorithms." John Wiley & Sons. 2003.
- [23] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: "An Efficient K-means Clustering Algorithm: Analysis and Implementation". Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 24, No. 7 (2002) 881-892.
- [24] L. Y. Tseng and S. B. Yang, "A genetic approach to the automatic clustering algorithm," Pattern Recognition, vol. 34, no. 2, pp. 415-424, 2001
- [25] Tomassini, M. (1999). "Parallel and Distributed Evolutionary Algorithms:A Review."In K. Miettinen, M. M'akel'a, P. Neittaanmäki and J. Periaux (Eds.), *Evolutionary Algorithms in Engineering and Computer Science* (pp. 113 - 133). Chichester: J. Wiley and Sons.
- [26] Witten, I.H., Frank, E.: "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations." Morgan Kaufmann Publishers. San Diego, CA (1999)
- [27] Krishna, K. and M. Murty, "Genetic K-means algorithm". IEEE Trans. Syst., Man, Cybern. B., Cybern., vol. 29, no. 3, pp. 433 – 439, Jun. 1999.
- [28] Peña, J. Lozano and P. Larrañaga, "An empirical comparision of four initialization methods for the k-means algorithm." Pattern Recognit Lett., vol. 20 pp. 1027 – 1040, 1999
- [29] Tomassini, M. (1999). "Parallel and Distributed Evolutionary Algorithms: A Review." In K. Miettinen, M. M'akel'a, P. Neittaanmäki and J. Periaux (Eds.), *Evolutionary Algorithms in Engineering and Computer Science* (pp. 113 - 133). Chichester: J. Wiley and Sons.
- [30] Bradley P., U.Fayyad. "Refining initial points for k-means clustering", in Proc. 15 th Int. Conf. Machine Learning, 1998 pp.91-99.

- [31] P. Bradley, and U. Fayyad, "Refining Initial Points for K-Means Clustering," In Proceeding of 15th International Conference on Machine Learning, 1998, pp. 91-99.
- [32] D. Bhandari, C.A. Murthy, S.K. Pal, "Genetic Algorithm with elitist model and its convergence", Int. J. Pattern Recognition Artif. Intell. 10 (1996) 731-747.
- [33] L.J. Eshelman (Ed.), Proceedings of the Sixth International conference Genetic Algorithms, Morgan Kaufmann, San Mateo, 1995.
- [34] J.L.R.Filho, P.C.Treleaven, C. Alippi, Genetic algorithm programming environments, IEEE Comput. 27 (1994) 28-43.
- [35] S.K. Pal, D. Bhandari, Selection of optimal set of weights in a layered network using genetic algorithms, Inform. Sci.80 (1994) 213-234.
- [36] A. Homaifar, S.H.Y. Lai, X. Qi, "Constrained Optimization via genetic algorithms," Simulation 62 (1994) 242-254.
- [37] A. Korol, I. Preygel, and S. Preygel, "Algorithms Of Estimation And Population-Genetic Models," Central Library of Imperial College, Chapman and Hall, 1994.
- [38] C.Z. Janikow, Z. Michalewicz, "An experimental comparison of binary and floating point representations in genetic algorithms," in: R.K. Belew, J.B. Booker (Eds.), Proceedings of the Fourth International Conference Genetic Algorithms, Morgan Kaufmann, San Mateo, 1991, pp. 31-36.
- [39] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.
- [40] Kaufman L. and P. Rousseeuw. "Finding Groups in Data: An Introduction to Cluster analysis": Wiley, 1990.
- [41] Goldberg, D. E. (1989). "*Genetic Algorithms in Search, Optimization, and Machine Learning*." Boston: Addison-Wesley.
- [42] D. Goldberg, "Genetic Algorithms in Search, Optimization, and Machine Learning," Addison-Wesley, New York, 1989
- [43] P. Chi, "Genetic Search with Proportion Estimation", In proceedings of the Third Int. Con. on Genetic Algorithms (ICGA), San Mateo, California, 1989, pp. 92-97.
- [44] Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Boston: Addison-Wesley.
- [45] Oliver, I.M., Smith, D.J., & Holland, J.R.C. (1987, July). A Study of Permutation Crossover Operators on the Traveling Salesman Problem. In John J. Grefenstette (Ed.), *Proceedings of the 2nd International Conference on Genetic Algorithms, Cambridge, MA, USA, July 1987*, (pp. 224-230). Lawrence Erlbaum Associates.
- [46] Oliver, I.M., Smith, D.J., & Holland, J.R.C. (1987, July). A Study of Permutation Crossover Operators on the Traveling Salesman Problem. In John J. Grefenstette (Ed.), *Proceedings of the 2nd International Conference on Genetic Algorithms, Cambridge, MA, USA, July 1987*, (pp. 224-230). Lawrence Erlbaum Associates.
- [47] S.Z. Selim, M.A. Ismail, K-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell.6 (1984) 81-87.
- [48] D. Goldberg, "Computer-Aided Gas Pipeline Operation Using Genetic Algorithms And Rule Learning," Ph.D. Paper, University of Michigan, Ann Arbor, 1983.
- [49] Lloyd SP "Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical statistics Meeting Atlantic City, NJ, sep. 1957. IEEE Trans. inform, Theory (Special Issue on Quantization), vol. IT-28, pp 129 – 137 Mach 1982.
- [50] R.C. Dubes, A.K. Jain, Clustering techniques: the user's dilemma, Pattern Recognition 8 (1976) 247-260.
- [51] J. Holland, "Adaptation In Natural And Artificial Systems," University of Michigan Press, 1975.
- [52] M.R. Anderberg, Cluster Analysis for Application, Academic Press, New York, 1973.
- [53] Hansen, P. & E. Nagai. "Analysis of Global k-means, an Incremental Heuristic for Minimum Sum of Squares Clustering". Journal Classification 22:287-310.
- [54] MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings Fifth Berkeley Symposium Mathematics Statistics and Probability. Vol. 1. Berkeley, CA (1967) 281-297.
- [55] Ball, G. and D. Hall, "A clustering technique for summarizing multivariate data", (ISODATA), Behav Sci., vol. 12, pp. 153-155, 1967
- [56] Forgy E. "Cluster analysis of multivariate data: Efficiency vs. Interpretability of classification", Biometrics, vol. 21, pp.768-780.1965
- [57] Pérez, J., M.F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas, A. Mexicano. Mejora al Algoritmo de K-means mediante un Nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer. 2do Taller Latino Iberoamericano de Investigación de Operaciones, "La IO aplicada a la solución de problemas regionales". México. "In Spanish".
- [58] IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.9, September 2008
- [59] A.K. Jain, M.N.Murty and P.J.FLYNN "Data Clustering: A Review", Advances in Image understanding, IEEE Computer Society Press, January 1999.