Research issues on K-means Algorithm: An Experimental Trial Using Matlab

Joaquín Pérez Ortega[1] , Ma. Del Rocío Boone Rojas,[1,2], María J. Somodevilla García[2]

1 Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca Mor. Mex.
2 Benemérita Universidad Autónoma Puebla, Fac. Cs. de la Computación, México.
jperez@cenidet.edu.mx, {rboone,mariasg}@cs.buap.mx

Abstract. Clustering problems arise in many different applications: machine learning data mining and knowledge discovery, data compression and vector quantization, pattern recognition and pattern classification.
It is considered that the k-means algorithm is the best-known squared error-based clustering algorithm, is very simple and can be easily implemented in solving many practical problems.
This paper presents the results of an analysis of the representative works related to the reseach lines of k-means algorithm devoted to overcome its shortcomings. To establish a framework for a proposed improvement to the standard k-means algorithm, the results obtanained from experiments of the k-means in the Matlab package and databases of the UCI repository are presented.

Keywords: k-means, clustering, k-means-Matlab.

1 Introduction.

The problem of object clustering according to its attributes has been widely studied due to its application in areas such as machine learning [4], data mining and knowledge discovery [3, 11], pattern recognition and pattern classification [2]. The aim of clustering is to partition a set of objects which have associated multi-dimensional attribute vectors into homogeneous groups such that the patterns within each group are similar. Several unsupervised learning algoritms have been proposed which partition the set of objects into a given number of groups according to an optimization criterion. One of the most popular and widely studied clustering methods is K-means [10].

This work is part of a project devoted to proposals for improvement to k-means algorithm and its application to health care in México. In the works of [29] and [30] proposes an improvement to k-means algorithm using a new convergence condition and application in cancer incidence by municipalities in México is proposed. Currently, the project focuses on making a proposal for improvement of the algorithm based on the optimization of the classification step. In doing so, we have done an update and analysis of the state of the art in theoretical research of the algorithm. For purposes of establishing a framework to propose an improvement, a serie of experimental tests on the matlab package for kmeans in databases of the UCI repository has been made.

The works of [41] and [42] highlight the impact and relevance of the k-means algorithm, within the context of Data Mining and clustering techniques. There are over a thousand papers published to date, related to its different forms, applications and contributions of k-means algorithm. Research on the k-means algorithm, has been developed in two major directions, from theory and field of applications. From theory, a set of advantages and shortcomings has been identified, in an attempt of try in to overcome their shortcomings. Also, a serie of works have been developed, that have led different lines of research, on wich this work focuses. Study we have made no attempt to be exhaustive, but rather representative of the reseach. We include in this report, selected works that have been frequently cited and work that was deemed to provide a novel approach. This report is organized as follow: following this introduction, it is included in the Section 2, the approach to the clustering problem and the description of the k-means algorithm. The section 3 provides an analysis and synthesis of the works in the various research lines of k-means. Section 4, describes certain characteristics of k-means in Matlab and in the section 5, we present a reference database of UCI repository and results of test performed to illustrate some aspects of the performance of k-means algorithm on Matlab. Finally, in section 6, we will discuss our preliminary results and present the future work.

2 Clustering Problem and the k-means Algorithm.

According to [15], 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 R^d, find subjets of X wich are homogeneouns 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, wich lead to as many clustering problems and even more heuristic or exact algoritms, so clustering is a vast subject.

The work of [42] provides a good survey on the issue of clustering and provides a set of key references..

According to [15], a mathematical programming formulation of the minimum sum-of-squares clustering problema is a follows:

$$\min f(M,Z) = \sum_{i=1}^{M} \sum_{j=1}^{N} z_{ij} \left\| x_j \quad m_i \right\|^2$$

subject to

$$\sum_{i=1}^{M} z_{ij} = 1, \; j = 1,2,\ldots,N$$

z_{ij}€ $\{0,1\}$ i = 1,2, …, M; j =1,2, …,N

$$\text{where} \qquad m_i = \frac{\sum_{j=1}^{N} z_{ij} x_j}{\sum_{j=1}^{N} z_{ij}}, \qquad i = 1,2, \dots M.$$

The N entities to be clustered are at given points x_j = (x_{i1}, x_{j2}, …., x_{jd}) of R^d for j = 1, …., N; M cluster centroids must be located at unknown points m_i € R^d for i = 1, …, M. The decision variable z_{ij} is equal to 1 if point j is assigned to cluster i, at a squared Euclidean distance $\|x_j - m_i\|^2$ from its centroid. It is well-known that condition z_{ij}€ $\{0,1\}$ may be replaced by z_{ij}€ $[0,1]$, since in the optimal solution, each entity belongs to the cluster with the nearest centroid.

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 [23], [24] and [12].
And the best-known heuristic for minimum sum-of-squares clustering is MacQuee`s [24]. It proceeds by seleccting a first set of M points as candidate centroid set, then alternately (i) assigning points of X to their closest centroid and (ii) recomputing centroids of clusters so-obtained, until stability is attained.

In a more specific form in the work of [29] four algorithm steps can be identified:

Step 1. Initialization. A set of objects to be partitioned, the number of groups and a centroid for each group are defined.
Step 2. Classification. For each database object its distance to each of the centroids is calculated, the closest centroid is determined, and the object is incorporated to the group related to this centroid.
Step 3. Centroid calculation. For each group generated in the previous step, its centroid is recalculated.
Step 4. Convergence condition. Several convergence conditions have been used from which the most utilized are the following: stopping when reaching a given number of iterations, stopping when there is no exchange of objects among groups, or stopping when the difference among centroids at two consecutive iterations is smaller than a given threshold. If the convergence condition is not satisfied, steps two, three and four of the algorithm are repeated.

3 Analysis and Classification of Papers Reviewed.

This section, identifies the advantages of k-means. The following is the result of the analysis of the work that has been developed in different research lines of k-means, related to the proposals made to try to overcome their shortcomings. The results are organized by category it, has been included in each case, the authors' names, the title and a concise comment on the thrust of the work.

3.1 Algortihm K-means Advantajes.

MacQeen J. [24], the author of one of the initial k-means algorithm and the most frequently cited, states.

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 extend by mathematical analysis and practical experience. 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.

Likewise [39], summarizes the benefits of k-means, in the introduction to his work:

K-means algortihm is one of first which a data analyst will use to investigate a new data set because it is algorthmically simple, relatively robust and gives "good enough" answers over a wide variety of data sets.

3.2 Algorithm K-means Shortcomings.

Taking as a framework and as an extension and update, the k-means shortcomings that are identified in [42] the following is the result of the analysis previously cited in a series of tables grouped by category of work that arose as extensions k-means or as possible solutions to one or more of the limitations that have been identified above.

3.2.1 The algorithm's sensitivity to initial conditions: The number of partitions, the initial centroids.

According to [42] there is a universal and efficient method to identify initial patterns and the number k of clusters. In [40] 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:

Authors	Title and Commentary
[45] Zhang, Chen; Xia Shixiong.	"K-means Clustering Algorithm with improved initial Center." It avoids the initial random assignment of centers. Use strategy called "sub-merger"
[2] B. Bahmani Firouzi, T. Niknam, and M. Nayeripour.	"A New Evolutionary Algorithm for Cluster Analysis". It not depend on the initial centers. Algorithm PSO-SA-K combines the algorithms "Particle Swarm Optimization (PSO)," Simulated Annealing "(SA) and K-means.
[39] Barbakh Wesam And Colin Fyfe.	"Local vs global interactions in clustering algorithms: Advances over K-means." It focuses on the algorithm's sensitivity to initial conditions. Incorporate information on the role of overall performance. Define three new algorithms: Weighted k-means (WK), Inverse Weighted K-means (IWK) and Inverse Exponential k-means (IEK).

[19] Kao, Yi-Tung, Zahara, Erwie, Kao. I-Wei.	“A hibridized approach to data clustering”. Draft bioinformatics. Hybrid techniques called K-NM-PSO-based K-means, Nelder-Mead Simplex search and optimization of exchange of particles.
[6] 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.” Title explicit.
[34] Redmond, Stephen J., Heneghan, Conor.	A method for initialising the K-means clustering algorithm using kd-trees” A kd-tree used to calculate an estimate of the density of data and to select the number of clusters.
[15] P. Hansen & E. Nagai.	“Analysis of Global k-means, an Incremental Heuristic for Minimum Sum of Squares Clustering”. Commentary on work [22].
[31] Pham, D. Dimov, S.S. Nguyen, C.D.	“Selection of K in K-means clustering”. It proposes a measure to select the reference number of clusters.
[22] Likas, A., Vlassis, N., Verbeek, J.J.	“The Global K-means Clustering Algorithm.” Algorithm that consists of a series of k-means clusterings with varying number of clusters from 1 to k. It argues that it is independent of initial partitions and accelerates the calculations of k-means.
[28] J. Peña, J. Lozano and P. Larrañaga,	“An empirical comparision of four initialization methods for the k-means algorithm.” Compare initialization methods for k-means: Random, [12], [20] and [24].
[5] P. Bradley, U.Fayyad.	“Refining initial points for k-means clustering”. Use k-means M times for M random subsets of the original data.
[20] L. Kaufman and P. Rouseeuw.	L. Kaufman and P. Rouseeuw. Finding Groups in Data: An Introduction to Cluster analysis: Text. Def. K-means.
[3] G. Ball and D. Hall	“A clustering technique for summarizing multivariate data”, (ISODATA). Perform dynamic estimation of K.

3.2.2 The convergence of algorithm to a local optimum rather than a global optimum.

According to [24], the iterative procedure of k-means can not guarantee convergence to a global optimum, but in his work, some research is cited, which are special cases. Currently, there are several developments that analyze and / or proposed solutions to this constraint:

Authors	Títle and Commentary
[39] Barbakh Wesam And Colin Fyfe.	“Local vs global interactions in clustering algorithms: Advances over K-means.” Addresses the algorithm's sensitivity to initial conditions. Incorporating global information on the performance function. Define three new algorithms: Weighted k-means (WK), Inverse Weighted K-means (IWK) and Inverse Exponential k-means (IEK).
[29] Joaquín .Pérez O, Rodolfo Pazos R, Laura Cruz R.,Gerardo Reyes S. Rosy Basave T. Héctor Fraire H.	“Improvement the Efficiency and Efficacy of the K-means Clustering Algorithm through a New Convergence Condition”. Improvement to the k-means algorithm by new convergence conditions. Experimentally analyze the local convergence of k-means.
[44] Z. Zhang, B. Tian D. And Tung A.K.H.	“On the Lower Bound of Local Optimums in K-means Algorithm.” Estimate lower limit for local optimum.

[21] K. Krishna and M. Murty	“Genetic K-means algorithm”. Hybrid scheme based on Genetic Algorithm - Simulated annealing with new operators to perform global search and rapid convergence.
[24] MacQUEEN J.	“Some Methods for Classification and Analysis of Multivariate Observations.” Definition, Analysis and Applications of k-means.

3.2.3 The efficiency of the algorithm.

According to the work of [42] the complexity of the k-means algorithm is O (n, d, k) which involves the sample size, the number of dimensions and the number of partitions. There are several works that have focused on different aspects of the algorithm, in order to reduce computational load.

Authors	Títle and Commentary
[4] Moh`d Belal Al-Zoubi, Amjad Hudaib, Ammar Huneiti and Bassam Hammo	“New Efficient Strategy to Accelerate k-Means Clustering Algorithm”. Strategy to accelerate k-means algorithm, which avoids many calculations of distance, through a strategy based on an improvement to the partial distance algorithm (PD).
[43] Zalik, Krista Rizman	“An Efficient k`-means Clustering Algorithm.” Based on the algorithm Rival, it penalizes competitive Learninig (RPCL). It does not require pre-allocation of the number of clusters. Two-step process. Pre processes and uses the prior information to minimize the cost function.
[26] Cao. D. Nguyen & Cios, Krzysztof J.	“GAKREM: A novel hybrid clustering algorithm.” Eliminates the need to specify a priori the number of clusters. Combines genetic algorithms and logarithmic regression Màxima expectation.
[13] G. Frahling & Ch. Sohler.	“A Fast k-means implementation using coresets.” Implemented version of Lloyd's k-means [23], using a weighted set of points that approximate the original set.
[35] Taoying Li & Yan Chen	“An improved k-means algorithm for clustering using entropy weighting measures”. Improvement of the algorithm by introducing a variable to the function of cost.
[18] Kashima, H. Hu, J.; Ray,B; Singh, M.	“K-means clustering of proportional data using L1 distance”. K-means based on distance L1. Proportionate restrictions incorporated in the calculation of centroids.
[36] Tsai, Chieh-Yuan, Chiu, Chuang-Cheng.	“Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm.” Improvement of the quality of k-means clustering via FWSA mechanism called "Self-Adjusment Feature Weight." Is modeled as an optimization problem.
[29] Joaquín .Pérez O, Rodolfo Pazos R, Laura Cruz R.,Gerardo Reyes S. Rosy Basave T. Héctor Fraire H.	“Improvement the Efficiency and Efficacy of the K-means Clustering Algorithm through a New Convergence Condition”. Improvement to the k-means algorithm by new convergence conditions. Experimentally analyze the local convergence of k-means.
[30] J.Pérez, M.F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas, A. Mexicano.	“Improvement of the K-means algorithm using a new approach of convergence and its application to databases cancer population." Title explicit.

[33] Pun, W.K.D., Ali, A.S.	"Unique distance measure approach for K-means (UDMA-Km) clustering algorithm." Sets distance measure based on statistical data.
[7] Zejin Ding, Jian Yu, Self-Yang-Qing Zhang.	"A New improved K-Means Algorithm with Penalizaed Term". Define new objective function and minimize it with genetic algorithm.
[17] 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," presents an implementation of the version of Lloyd's k-means [23] called "filtering algorithm" based on a kd-tree.

3.2.4 K-means is sensitive to *outliers* and noise.

According to [42] even if an object is quite far away from the cluster centroid, it is still forced into a cluster and, thus, it distorts the cluster shapes. Here are works that focus on shortcoming:

Authors	Títle and Commentary
[1] Asgharbeygi, N. Maleki, A.	"Geodesic K-means clustering".Extends k-means by using a geodesic distance metric. Algorithm ensures resistance to outliers.
[9] V. Estivill-Castro and J. Yang	"A fast and robust general purpose clustering algorithm." It eliminates the effect of outliers through a process that considers real points as centroids.
[3] G. Ball and D. Hall	"A clustering technique for summarizing multivariate data".(ISODATA). It performs dynamic estimation of K. Considers the effect of outliers in the process of clustering.

3.2.5 The definition of "means" limits the application only to numerical variables.

Several works have been developed that extend the application of k-means for categorical variables or others:

Authors	Títle and Commentary
[38] Song, Wei, Li Cheng Hua, Park, Soon Cheo.	"Genetic Algorithm for text clustering using ontology and evaluating the vality of various semantic simility measures." Improving the k-means algorithm by using a genetic algorithm that finds similarities conceptual. Based on ontology, thesaurus corpus for clustering of text fields.
[14] S. Gupata, K. Rao, &Bhatnagar	"K-means clustering algorithm for categorical attributes". Title explicit.
[16] Z. Huang.	"Extensions to the k-means algorithm for clustering large data sets with categorical values." Title explicit.

4 The Algorithm k-means on Matlab.

Experimental tests were conducted for K-means in the Matlab [25]. The Matlab (Matrix Laboratory) is both, an environment and programming language for numerical calculations with vectors and matrices. It is a product of the company The

Math Works Inc. (Natick, MA). [1]. The K-means algorithm for clustering is in the following MATLAB function:

[IDX, C, SUMD, D] = KMEANS(X, K)

This function partitions the points in the N-by-P data matrix X into K clusters. This partition minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Rows of X correspond to points, columns correspond to variables. KMEANS returns an N-by-1 vector IDX containing the cluster indices of each point. By default, KMEANS uses squared Euclidean distances. The K cluster centroid is located in the K-by-P matrix C. The within-cluster sums point-to-centroid distances in the 1-by-K vector sumD. Distances from each point to every centroid in the N-by-K matrix D. It may include optional parameters to specify distance measure, the method used to choose the initial cluster centroid positions, display information.

5 Test Results of K-means in Matlab.

Tests for k-means in Matlab, used the well known UCI Machine Learning Repository [37]. The UCI Machine Learning Repository [37] is among other things, a collection of databases, which is widely used by the research community of Machine Learning, especially for the empirical algorithms analysis of this discipline.

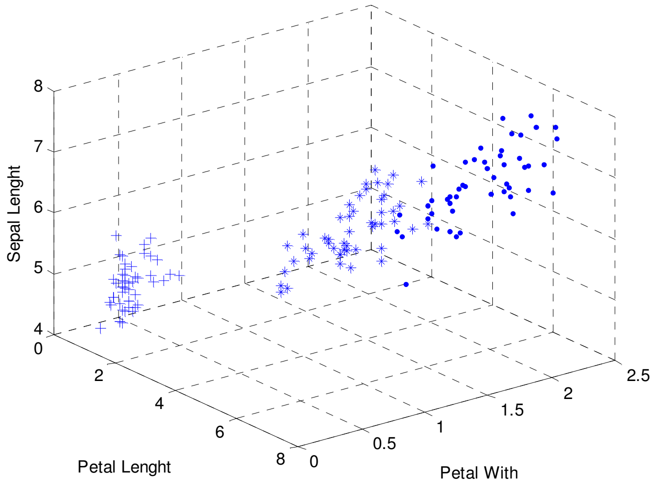

Fig.1 Representation of the Iris Data Set

For the experimental tests carried, the following data sets have been used: Iris, Glass and Wine. This report presents the results for the Iris data set.

The Iris Data Set is a database of types Iris plant, which has No. of instances: 150 (50 in each class), No. of attributes: 4 (Sepal length, Sepal width, Petal Length, Petal width) No. of classes: 3 (Hedges Iris, Iris versicolor, Iris virginica). One class is linearly separable from the other two; the latter are NOT linearly separable from each other. Based on data and classes defined in [37] and [42], fig.1 includes the Iris data, for illustrative purposes only are considered the attributes sepal length, petal length and petal width.

Test I4: >> [u,v,sumd,D]= kmeans(z,3,'display','iter');

```
iter    phase   num     sum       %inter
1       1       150     426.888   100
2       1       34      134.187   22.6
3       1       13      105.771   8.6
4       1       12      88.8948   8
5       1       6       85.2326   4
6       1       4       84.064    2.6
7       1       3       83.3704   2
8       1       5       82.073    3.3
9       1       3       81.3672   2
10      1       4       80.3157   2.6
11      1       3       79.6817   2
12      1       3       79.1156   2
13      1       1       78.9451   0.6
14      2       1       78.9408   0.6

14 iterations, total sum of distances = 78.9408
```

The "Test I4" is an example of the test results on Matlab and kmeans for the Iris data set. ***Iter*** column represents the number of iteration, the ***phase*** indicates the algorithm phase, ***num*** provides the number of exchanged points, ***sum***, provides the total sum of distances, ***inter%*** are the percentage of exchanged points in each iteration.

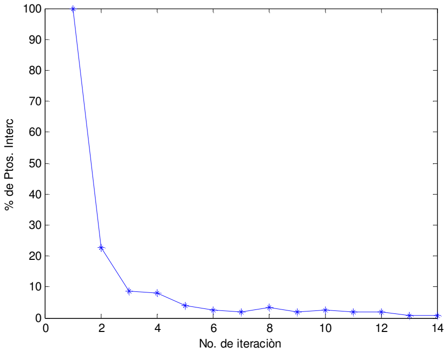

Fig. 2 % of Exchanged Points.

Fig. 2 corresponds to the test I4 and the graphical representation of the exchange behavior at each iteration. Likewise Fig. 3 represents the behavior of the sum of distances for the same test.

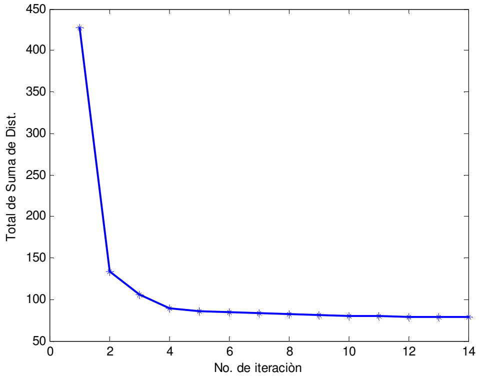

Fig. 3 Total Sum of Distance

No. Prue.	No. Iter.	No. Ptos.Interc.2	% Dif. Ptos. Interc. 1 a 2
1	5	14	90.7
2	10	7	95.4
3	3	5	96.7
4	6	8	94.7
5	14	34	77.4
6	4	5	96.7
7	8	8	94.7
8	8	36	97.6
9	7	73	95.4
10	3	4	97.4
11	7	37	97.6
12	11	9	94.0
13	6	53	64.7
14	5	12	92.0
15	3	2	98.7
16	6	15	90.0
17	4	3	98.0
18	10	9	96.0
19	10	4	97.4
20	6	32	79.0
21	12	16	89.4
22	8	33	78.0
23	7	31	80.0
24	6	14	90.7
25	11	7	95.4

Table 1 Summary of Results for Iris Data Set

5.1 Summary of Results for the Iris data set.

Table 1 provides a summary of 25 experimental tests performed on the IRIS data set, the first column identifies the test number and the second column includes the number of iterations performed by the algorithm.

With regard to exchanges between groups that the algorithm makes in the tests conducted, it was observed that the most significant changes occurred from first to second iteration. For all cases, in the first step all points are located (100%), the third column includes the number of points to be exchanged in the second iteration and the fourth column is the percentage difference in the number of items exchanged between the first and second iteration.

According to the results in Table 1:

It can see that for 150 points in the Iris database, and a set of 25 tests, the k-means algorithm in Matlab:

- ✓ Converge in an average of 7.2 iterations.
- ✓ The average number of points exchanged during the second iteration was 18.84 points
- ✓ The percentage of points located on the second iteration in the corresponding group was 91.0%

6 Conclusions.

The results of the analysis for our sample work, allow us to establish a framework and analyze the theoretical study of the k-means algorithm Also we reseach and distinguish the different lines on which there is still a fertile field for investigation.
As we can see several attempts at overcoming the shortcomings of the k-means algorithm have been done and different approaches in different disciplines have been proposed: Optimization, Probability and Statistics, Neural Networks, Evolutionary Algorithms, among others. The vast majority of contributions have focused on the first three lines of research identified in this study: The sensibility of the algorithm to initial conditions, convergence of the algorithm to a local optimum rather than a global optimum and the efficiency of the algorithm. Notes that challenges still to be resolved in such research and has been relatively little work done on the lines related to the implementation of the algorithm to other variables as well as treatment to outliers and noise.
Acording the tests conducted in Matlab, this laboratory showed that it is actually very conducive to experimental testing, the implementation of k-means, allows to monitor the performance of the algorithm through the information that can be deployed at runtime, such as result of the objective function and the number of points exchanged in each iteration.The results allow us to establish a framework to compare the proposal improvement algorithm with the previous work. As part of this project and to give continuity to previous work [29] [30], also ventures into different applications to k-means, such as in the areas of health care in Mexico and in Web Usage Mining for Log files from the server of the Faculty of Compute Science BUAP, México.

References

1. Asgharbeygi, N. Maleki, A. “Geodesic K-means clustering”. Pattern Recognition, 2008, ICPR 2008, 19th International Conference on. Dec. 2008.

2. 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.

3. Ball, G. and D. Hall, “A clustering technique for summarizing multivariate data”, (ISODATA), Behav Sci., vol. 12, pp. 153-155, 1967.

4. 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.

5. Bradley P., U.Fayyad. “Refining initial points for k-means clustering”, in Proc. 15 th Int. Conf. Machine Learning, 1998 pp.91-99.

6. 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.

7. 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.

8. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis, John Wiley &Sons, New York, NY, 1973.

9. Estivill-Castro V. and J. Yang, “A fast and robust general purpose clustering algorithm.” In Proc. 6 th Pacific Rim Int. Conf. Artificial Intelligence (PRICAI`00), R. Mizoguchi and J. Slaney, Eds., Melbourne, Australia, 2000, pp, 208 – 218.

10. Fayyad, U.M., Piatetsky-Shanpiro, G., Smyth P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

11. Fissher, D.: Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning, Vol.2, No. 2 (1987) 139-172.

12. Forgy E. “Cluster analysis of multivariate data: Efficiency vs. Interpretability of classification”, Biometrics, vol. 21, pp.768-780.1965

13. Frahling, G. & Ch. Sohler. “A Fast k-means implementation using coresets.”. International Journal of Computational Geometry & Applications. Dec. 2008. Vol. 18 Issue 6. P605-625.

14. Gupata S., K. Rao, and V. Bhatnagar, “K-means clustering algorithm for categorical attributes”, in Proc. 1st Int. Conf. Data Werehousing and Knowledge Discovery (DaWak`99). Florence, Italy, 1999, pp. 203 – 208.

15. Hansen, P. & E. Nagai. “Analysis of Global k-means, an Incremental Heuristic for Minimum Sum of Squares Clustering”. Journal Classification 22:287-310.

16. Huang, Z., “Extensions to the k-means algorithm for clustering large data sets with categorical values.”. Data Mining Knowl. Discov., vol. 2, pp. 283 – 304, 1998.

17. 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.

18. Kashima, H. Hu, J.; Ray,B; Singh, M. “K-means clustering of proportional data using L1 distance”. Pattern Recognition, 2008, ICPR 2008. International Conference On. Volume Issue, Dec. 2008.

19. 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.

20. Kaufman L. and P. Rouseeuw. Finding Groups in Data: An Introduction to Cluster analysis: Wiley, 1990.

21. 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.

22. 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

23. 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.

24. 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.

25. Matworks. http: //www.matworks.com

26. Mehmed, K.: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. 2003.

27. Nguyen, Cao. D. & Cios, Krzysztof J. “GAKREM: A novel hybrid clustering algorithm. Information Sciences. Vol. 178 Issue 22, p4205-4227- Nov. 2008.

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. Pérez J., Rodolfo Pazos R, Laura Cruz R.,Gerardo Reyes S. Rosy Basave T. Héctor Fraire H. “Improvement the Efficiency and Efficacy of the K-means Clustering Algorithm through a New Convergence Condition”. Computational Science and Its Applications – ICCSA 2007 – International Conference Proceedings. Springer Verlag.

30. 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”.

31. 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.

32. Proietti, Guido and Christos Faloutsos. “Analysis of Range Queries on Real Region Datasets Stored Using an R-Tree.” IEEE Transactions on Knowledge and Data Engieneering., Vol. 12, No. 5, Sep./Oct. 2000.

33. 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.

34. 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.

35. 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.

36. Tsai, Chieh-Yuan, Chiu, Chuang-Cheng. “Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm.” Computational Statistics & Data Analysis. Vol. 52 Issue 10. Jun. 2008.

37. UCI. Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

38. Wei, Song, Li Cheng Hua, Park, Soon Cheo. “Genetic Algorithm for text clustering using ontology and evaluating the vality of various semantic simility measures.” Expert Systems with Applications. Vol. 36, Issue 5, Jul. 2009.

39. Wesan, Barbakh And Colin Fyfe. “Local vs global interactions in clustering algorithms: Advances over K-means.” International Journal of knowledge-based and Intelilligent Engineering Systems 12 (2008).83 – 99.

40. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers. San Diego, CA (1999)

41. Xindong Wu,, V.Kumar, J. Ross Q., J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng,, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand and D. Steinberg. “Top 10 algorthms in data mining”. Knowl Inf Syst (2008). Springer.

42. Xu,_Rui and Donald Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, Vol., 16, No. 3, May 2005.

43. Zalik, Krista Rizman . “An Efficient k`-means Clustering Algorithm.” Pattern Reconition Letters, Vol. 29, I.9. Pag. 1385-1391. Elsevier 07/2008.

44. 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.

45. 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.