International Journal of Soft Computing 7 (5): 249-255, 2012
ISSN: 1816-9503

Genetic Algorithm Based Dimensionality Reduction for Improving Performance of K-Means Clustering: A Case Study for Categorization of Medical Dataset

Asha Gowda Karegowda, Vidya T. Shama, M.A. Jayaram and A.S. Manjunath
Department of Master of Computer Applications, Siddaganga Institute of Technology,
573103 Tumkur, Karnataka, Bangalore, India

Abstract: Medical data mining is the process of extracting hidden patterns from medical data. Among the various clustering algorithms, k-means is the one of most widely used clustering technique. The performance of k-means clustering depends on the initial cluster centers and might converge to local optimum. k-means does not guarantee unique clustering because it generates different results with randomly chosen initial clusters for different runs of k-means. In addition the performance of any data mining depends on feature subset selection. This study attempts to improve performance of k-means clustering using two stages. As part of first stage, this study investigates the use of wrapper approach for feature selection for clustering where Genetic Algorithm (GA) is used as a random search technique for subset generation, wrapped with k-means clustering. In second stage, GA and Entropy based Fuzzy Clustering (EFC) are used to find the initial centroid for k-means clustering. Experiments have been conducted using standard medical dataset namely Pima Indians Diabetes Dataset (PIDD) and Heart statlog. Results show markable reduction of 8.42 and 18.89% in the classification error of k-means clustering for PIDD and Heart statlog dataset using features identified by proposed wrapper approach and initial centroids identified by GA when compared to k-means performance with all the features and centroids initialized by random method for PIDD and Heart statlog dataset.

Key words: k-means clustering, genetic algorithm, dimensionality reduction, wrapper approach, cluster center initialization, entropy based fuzzy clustering, medical dataset

INTRODUCTION

The data mining functionalities mainly include association rule mining, classification, prediction and clustering. Classification is supervised learning algorithms in contrasts with clustering which are unsupervised learning algorithm. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters (Han and Kamber, 2001). Huge data repositories, especially in medical domains, contain enormous amounts of data. These data includes also currently unknown and potentially interesting patterns and relations which can be uncovered using knowledge discovery and data mining methods. Medical data mining has enormous potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. Data pre-processing is a significant step in the knowledge discovery process, since quality decisions are based on quality data. In real-world situations, relevant features are often unknown a priori. Hence, feature selection is a must to identify and remove irrelevant/redundant features. It can be applied in both unsupervised and supervised learning. The goal of feature selection for unsupervised learning is to find the smallest feature subset that best uncovers clusters from data according to the preferred criterion (Dy and Bradley, 2004). Feature selection in unsupervised learning is much harder problem, due to the absence of class labels. Feature selection for clustering is the task of selecting important features for the underlying clusters (Dash *et al*., 2002). One of most common clustering method is a k-means Clustering. The performance of k-means clustering mainly depends on the initial cluster centers and might converge to a local optimum. The performance of k-means can be improved by identifying the significant feature subset and the initial cluster centroids. This study investigates the use of GA (Goldberg, 1989) to improve the performance of k-means clustering by identifying significant features and selection of k-means initial cluster centroids. Performance of proposed research is compared with initializing k-means initial centroids using random method and EFC centroids.

Corresponding Author: Asha Gowda Karegowda, Department of Master of Computer Applications,
Siddaganga Institute of Technology, 573103 Tumkur, Karnataka, Bangalore, India

LITERATURE REVIEW

Feature selection for unsupervised learning can be subdivided into Filter Methods and Wrapper Methods. Filter methods in unsupervised learning is defined as using some intrinsic property of the data to select feature without utilizing the clustering algorithm (Dy and Bradley, 2004). Entropy measure has been used as filter method for feature selection for clustering (Dash and Liu, 1997). Wrapper approaches in unsupervised learning apply unsupervised learning algorithm to each candidate feature subset and then evaluate the feature subset by criterion functions that utilize the clustering result (Dy and Bradley, 2004). Volker Roth and Tilman Lange propose a Wrapper Method where Gaussian Mixture Model combines a Clustering Method with a Bayesian inference mechanism for automatically selecting relevant features (Roth and Lange, 2003). Sun and Xiong (2009) has used GA for feature subspace selection of k-means clustering. They have used binary encoding to represent feature subspace and cluster centers. Several methods have been proposed to solve the cluster initialization for k-means algorithm. Bradley and Fayyad (1998) proposed the method which forms a set of small random sub-samples of the data and then apply k-means to each of sub-samples. The centroids of each sub-samples is given an initial center for k-means for all centroids of all subsamples. The centers of the final clusters that give minimum clustering error are to be used as the initial centers for clustering the original set of data using k-means algorithm. Al-Shboul and Myaeng (2009) have used Genetic algorithm to initialize the k-means cluster centers. Jimenez *et al.* (2007) have used GA to determine the best initial cluster centers and find the number of clusters using binary encoding. Eltibi and Ashour (2011) has used statistical information from the data set to initialize the k-means prototypes. Erisogly *et al.* (2011) has used two principal variables based on maximum coefficient of variation and minimum absolute value of the correlation. The reduced dataset is partitioned one at a time till the desired number of clusters is obtained. The cluster membership for each point is determined according to candidate initial cluster centers and selected two axis. Dey *et al.* (2011) as used EFC (Yao *et al.*, 2000) to initialize the fuzzy c-mean initial cluster centers. Maulik and Bandyopadhyay (2000) have applied GA to find the k-means cluster centers using floating point representation of clustering for three datasets: iris, crude oil and vowels dataset.

MEDICAL DATASET

Diabetes mellitus is a disease in which the body is unable to produce or unable to properly use and store glucose (a form of sugar). Glucose backs up in the bloodstream causing one's blood glucose or sugar to rise too high. Type 1 and 2 are main types of diabetes. In addition, the people who develop diabetes while pregnant (a condition called gestational diabetes) are more likely to develop full-blown diabetes later in life. Poorly managed diabetes can lead to a host of long-term complications among these are heart attacks, strokes, blindness (diabetic retinopathy), kidney failure (nephropathy), blood vessel disease (neuropathy) (ADA, 2008). The PIMA Indian diabetic dataset is availed from UCI Machine Learning Repository. All patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, Arizona, USA. The database consist of two categories in the data set (i.e., tested positive, tested negative) each having 8 features: number of times pregnant, plasma glucose concentration a 2 h in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2 h serum insulin (mu U mL^{-1}), body mass index (weight in kg/(height in m)^2), diabetes pedigree function and age (years). Out of 768 cases in PIDD, the missing value is found for 5 patients with glucose value of zero, 11 patients had a body mass index of zero, 28 patients had a diastolic blood pressure of zero, 192 others had skin fold thickness readings of zero and 140 others had serum insulin levels of 0 which is biologically impossible (Breault, 2001). After deleting these cases there were 392 cases with no missing values (130 tested positive cases and 262 tested negative).

In addition to diabetic dataset, Heart statlog dataset is used for experiments. Heart statolog dataset contains 13 attributes namely age, sex, chest pain type, resting blood pressure, serum cholestoral in mg/dL, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak, slope of peak exercise ST segment, number of major vessels colored by fluoroscopy, thal and two class labels: absence or presence of heart disease. Both the datasets are availed from the UCI Machine Learning Repository.

K-MEANS CLUSTERING

k-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms and follows partitioning method for clustering. k-means algorithm takes the input parameter, k as number of clusters and partitions a dataset of n objects into k clusters, so that the resulting objects of one cluster are dissimilar to that of other cluster and similar to objects of the same cluster. In k-means algorithms begins with randomly selected k objects,

representing the k initial cluster center or mean. Next each object is assigned to one the cluster based on the closeness of the object with cluster center. To assign the object to the closest center, a proximity measure namely Euclidean distance is used that quantifies the notion of closest. After all the objects are distributed to k clusters, the new k cluster centers are found by taking the mean of objects of k clusters, respectively. The process is repeated till there is no change in k cluster centers. k-means algorithm aims at minimizing an objective function namely Sum of Squared Error (SSE). SSE is defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \quad (1)$$

Where:

E = Sum of the square error of objects with cluster means for k cluster
p = The object belong to a cluster C_i
m_i = The mean of cluster C_i

The time complexity of k-means is O (t×k×n) where t is the number of iterations, k is number of clusters and n is the total number of records in dataset.

k-means partitioning algorithm: Input is k is the number of clusters, D is input data set. Output is k clusters:

- Randomly choose k objects from D as the initial cluster centers
- Repeat
- Assign each object from D to one of k clusters to which the object is most similar based on the mean value of the objects in the cluster
- Update the cluster means by taking the mean value of the objects for each of k cluster
- Until no change in cluster means/min error E is reached

GENETIC ALGORITHM BASED DIMENSIONALITY REDUCTION FOR K-MEANS CLUSTERING

Genetic algorithm: GA (Goldberg, 1989) is an optimization techniques inspired by natural selection and natural genetics. Unlike many search algorithms which perform a local, greedy search, GA is a stochastic general search method, capable of effectively exploring large search spaces. GA is mainly composed of three operators: reproduction, crossover and mutation. As a first step of GA, an initial population of individuals is generated at random or heuristically. The individuals in the genetic space are called chromosome. Gene is the basic building block of the chromosome. Locus is the position of particular gene in the chromosome. In each generation, the population is evaluated using fitness function. In the selection process, the high fitness chromosomes are used to eliminate low fitness chromosomes. But selection alone does not produce any new individuals into the population. Hence, selection is followed by crossover and mutation operations. Crossover is the process by which two-selected chromosome with high fitness values exchange part of the genes to generate new pair of chromosomes. The crossover tends to facilitate the evolutionary process to progress towards the potential regions of the solution space. Mutation is the random change of the value of a gene which is used to prevent premature convergence to local optima. The new population generated undergoes the further selection, crossover and mutation till the termination criterion is not satisfied. Convergence of the genetic algorithm depends on the various criterions like fitness value achieved or the maximum number of generations (Maulik and Bandyopadhyay, 2000; Eduardo *et al*., 2009).

GA has been used in this paper to identify the significant feature subset for k-means clustering and to find initial k-means cluster centers. The working of both binary encoding and integer encoding GA identifying significant features and initial centroids using binary encoding is explained.

Binary encoded GA for identifying significant attributes: The binary encoded chromosome used for experiments is briefed as follows. The length of the chromosome is equal to total number of features say F. The GA is experimented with different values of MaxF where MaxF is the maximum number of features selected as subset features in the range of [1 to F-1]. As example, for diabetic data set, F = 8 features. The chromosome length is 8. With MaxF = 4, for example 10011001 binary encoded chromosome represents 1st, 4th, 5th and 8th features are selected as input for k-means clustering and 2nd, 3rd, 6th and 7th attribute are not selected. With binary encoding of chromosomes, 1 represents the feature is selected and 0 represents the feature is not selected as part of feature subset selection. Each of the chromosomes has exactly MaxF number of ones where MaxF represents the size of feature subset to be selected. The working of GA for finding the feature subset for k-means using binary encoded chromosomes:

Step 1: Initialize the chromosome population randomly using binary encoding (each chromosome length is equal to total number of features F where number of ones is equal to MaxF).

Setp 2: Repeat the steps following till terminating condition is reached:

- Apply k-means clustering to individual chromosome, representing the significant features and find the Sum of Square Error (SSE)
- Select the chromosome with least SSE as the fittest chromosome. Replace the low fit chromosome by highest fit chromosome
- Select any two chromosomes randomly and apply crossover operation
- Apply mutation operation by randomly selecting any one chromosome and randomly change the bit 1 to 0 and bit 0 to 1

Setp 3: The position of bit 1 in the best-fit chromosome is considered as significant attributes for k-means clustering.

Integer encoded GA for identifying significant attributes: The integer encoded chromosome used for experiments is briefed as follows. With integer encoding, let F represent the total number of features. The length of the chromosome is equal to MaxF where MaxF is the maximum number of features selected as subset features in the range of 1 to F-1. As example with diabetic data set F = 8 features. Each gene in integer encoding may take a value in the range of 1 to F. With integer encoding of chromosomes, each gene represents the feature selected as part of feature subset selection. For example 1457 with integer encoded chromosome represents a chromosome with MaxF = 4 with 2nd, 3rd, 6th and 8th features not selected and 1st, 4th, 5th and 7th features are selected as input for k-means clustering. Experiments are conducted using different values of MaxF in the range of 1 to F-1. The working of GA for finding the feature subset for k-means using Integer encoded chromosomes:

Step 1: Initialize the chromosome population randomly using integer encoding. Each chromosome length is equal to MaxF where each gene is an integer number in the range of 1to F.

Setp 2: Repeat the steps following till terminating condition is reached:

- Apply k-means clustering to individual chromosome and find the Sum of Square Error (SSE)
- Select the chromosome with least SSE as the fittest chromosome. Replace the low fit chromosome by highest fit chromosome
- Select any two chromosomes randomly and apply crossover operation
- Apply mutation operation by randomly selecting any one chromosome and randomly change the randomly selected gene by an integer value in the range of 1-F

Setp 3: The integer numbers in the best-fit chromosome are considered as significant features to be given as input to k-means clustering.

METHODS USED TO INITIALIZE K-MEANS INITIAL CENTROIDS

GA based k-means initial cluster centroids: As a part of second stage of the proposed research, the GA is further used to identify the initial centroids for k-means clustering. Chromosomes are encoded using binary encoding where 1 represents the sample selected as initial cluster center and 0 represents the sample is not selected as initial cluster center (Karegowda *et al*., 2012). The length of the chromosome is equal to total number of samples. The GA was experimented with population's size of 50-110 chromosomes, number of generations with 15-30 and with both one point and two-point crossover. After mutation and crossover operation the number of ones in the chromosomes must be checked for not exceeding the number of required clusters. The one point and two-point crossover, resulted in almost the same results. The terminating condition is that 80% of the chromosomes represent the same initial cluster centers. Once the terminating condition is reached, the highest fittest chromosomes decide the samples which will be k-means initial cluster centers.

Entropy based Fuzzy Clustering (EFC): Yao introduced EFC (Yao *et al*., 2000) which identifies the number of clusters and initial cluster prototypes by itself. The entropy is calculated for each sample using Eq. 2:

$$E_i = \sum_{k \in x}^{j \neq i} (S_{ij} \log_2 S_{ij} + (1 - S_{ij}) \log_2 (1 - S_{ij})) \quad (2)$$

Where:

$s_{ij} = e^{-\alpha d_{ij}}$ = The similarity between two data points (i, j)

d_{ij} = The Euclidean distance between points (i, j)

The algorithm for entropy based fuzzy clustering is as follows. The inputs for the algorithm are dataset D with N samples, β the threshold value can be viewed as a threshold of similarity among the data points in the same cluster, an constant α is which is computed as (ln 0.5/ ($\bar{D}$)) where $\bar{D}$ is the mean distance among the pairs of data points in a hyper-space and is usually set to 0.5.

Step 1: Compute entropy Ei for each sample xi from dataset D for i = 1 to N.

Step 2: Identify xi that has the minimum Ei value as the cluster centre.

Step 3: Remove xi and data points having similarity xi greater than some threshold β from D.

Step 4: If D is not empty then go to Step 2.

The k centroids identified by EFC are selected as k-means initial cluster centres.

EXPERIMENTAL RESULTS

In the first state of proposed research, GA is experimented using binary encoding and integer encoding for feature selection. Both one point and two-point crossover were experimented. The terminating condition is 80% of the chromosomes represent the same feature subset. Once the terminating condition is reached, the highest fittest chromosome (with the least Sum of Square Error (SSE)) decides the significant feature subset for k-mean clustering. For both binary and integer encoding, the total number of features F is 8 for diabetic dataset and MaxF is experimented with 4, 5 and 6 values. With binary encoding the population's size is varied with 50-100 chromosomes, number of generations with 15-30 with different values of MaxF in the range of 4-6. With integer encoding the population size is experimented with 10-20 chromosomes, 10-70 with Maxf values as 4, 5 and 6. Table 1 shows that for diabetic dataset the SSE is least with MaxF = 6. The following significant attributes were identified: plasma glucose, pedigree, BMI, age, insulin and diastolic blood pressure with MaxF = 6. The significant attributes sets identified by both binary and integer encoded chromosomes are same with MaxF value as 5 and 6. For MaxF = 4, both integer and binary encoded chromosomes identified different sets of significant attributes.

For Heart statlog dataset the total number of features is 13 and experiments were conducted with different values of MaxF. The performance of k-means with MaxF with the value 10, 8 and 7 is showed in Table 1, among which k-means showed the least SSE with MaxF = 8 with the following features: age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate achieved, exercise induced angina, number of major vessels.

As a part of second stage the proposed research after identifying the significant features for diabetic with MaxF = 6 and Heart statlog dataset with MaxF = 8, further experiments were conducted by initializing the k-means with GA and EFC identified centroids using GA identified attributes. Table 2 shows comparative performance of

Table 1: Performance of k-means clustering with random centroids for different values of MaxF

Dataset	Attributes selection using binary/integer encoding	No. of features	TP	FP	TN	FN	No. of iterations	Sum of square error	Error
Diabetic	-	8	74	56	197	65	7	0.0087	30.86
	Binary/Integer	6	92	38	198	64	7	0.0027	26.02
	Binary/Integer	5	89	41	196	66	7	0.0099	27.29
	Binary	4	91	39	196	66	5	0.0076	26.78
	Integer	4	69	61	205	57	5	0.0058	30.14
Heart Statlog	-	13	74	46	86	64	6	0.0064	40.74
	Binary/Integer	10	74	46	86	64	5	0.0054	40.74
	Binary/Integer	8	70	50	128	22	5	0.0013	26.66
	Binary/Integer	7	87	33	100	50	3	0.0021	30.74

Table 2: Comparing k-means clustering using random, GA and EFC centroids with all and significant attributes identified by GA

Dataset	Method used to initialize k-means centroids	No. of features	TP	FP	TN	FN	No. of iterations	Sensitivity	Specificity	Recall	F-measure	Precision	Error
Diabetic	Random	8	74	56	197	65	7	0.53	0.78	0.53	0.55	0.57	30.86
	Random	6	92	38	198	64	7	0.59	0.84	0.59	0.64	0.71	26.02
	GA	8	86	44	197	65	5	0.57	0.82	0.57	0.61	0.66	27.80
	GA	6	92	38	206	56	5	0.62	0.84	0.62	0.66	0.70	22.44
	EFC	8	89	41	192	70	5	0.56	0.82	0.56	0.62	0.68	28.32
	EFC	6	92	38	198	64	4	0.59	0.84	0.59	0.64	0.71	26.02
Heart statlog	Random	13	74	46	86	64	6	0.53	0.34	0.53	0.57	0.61	40.74
	Random	8	70	50	128	22	5	0.76	0.28	0.76	0.66	0.58	26.66
	GA	13	96	24	79	71	5	0.57	0.23	0.57	0.66	0.80	35.18
	GA	8	95	25	116	34	3	0.73	0.17	0.73	0.76	0.79	21.85
	EFC	13	100	20	67	83	5	0.54	0.22	0.54	0.66	0.83	38.14
	EFC	8	80	40	121	29	3	0.73	0.24	0.73	0.69	0.66	25.55

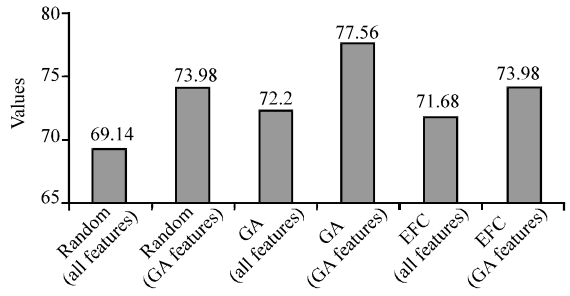

Fig. 1: k-means accuracy by initializing cluster centroids using random, GA and EFC Methods with all attributes and with significant attributes identified by GA for diabetic dataset

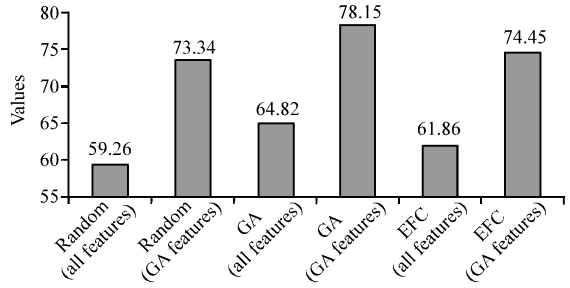

Fig. 2: k-means accuracy by initializing cluster centroids using random, GA and EFC Method with all attributes and with significant attributes identified by GA for Heart statlog dataset

k-means with all attributes and GA identified attributes using random initial centroids, k-means with all attributes and GA identified attributes using initial centroids identified by GA and k-means with all attributes and GA identified attributes using initial centroids identified by EFC in terms of sensitivity, specificity, recall and Precision, Fmeasure and classification error. Further more Table 2, depicts that the k-means requires less processing times (in terms of iterations) with significant attributes identified by GA.

Figure 1 illustrates the improved performance of k-means with a classification accuracy of 77.56% and 73.98% using GA identified attributes and centroids initialized by GA and EFC, respectively for diabetic dataset. Figure 2 illustrates the improved performance of k-means with a classification accuracy of 78.15 and 74.45% using GA identified attributes and centroids initialized by GA and EFC, respectively for Heart statlog dataset. The GA identified attributes have improved the performance of k-means using all the three methods of centroid initialization: random, GA and EFC. The hike in the accuracy of k-means is the result of not only GA identified attributes but also because of the initial centroids identified by GA and EFC. For both the medical dataset experimented, the GA centroids proved to be better compared to EFC centroids.

CONCLUSION

The performance of k-means clustering depends not only on the initial cluster centers but also on the significant dimensions. This study illustrates the improved performance of k-means clustering using GA identified attributes and with centroids identified by both GA and EFC. The GA identified centroids outperformed the performance of k-means when compared to EFC identified centroids. In addition, the k-means requires lower processing time in term of number of iterations using the proposed GA identified attributes. The performance of proposed research is experimented successfully for two medical dataset: PIMA Indian diabetic and Heart statlog dataset.

REFERENCES

ADA, 2008. Diagnosis and classification of diabetes mellitus. Diabetes Care, 31: S55-S60.

Al-Shboul, B. and S.H. Myaeng, 2009. Initializing K-means using genetic algorithms. World Acad. Sci. Eng. Technol., 54: 114-118.

Bradley, P.S. and U.M., Fayyad, 1998. Refining initial points for k-means algorithm. Proceedings of the 15th International Conference on Machine Learning, January 1998, Morgan Kaufmann, San Francisco, pp: 91-99.

Breault, J.L., 2001. Data mining diabetics databases: Are rough sets a useful addition? Proceedings of the 33rd Symposium on Interface, Computing Science and Statistics, June 13-16, 2001, Costa Mesa, CA., USA.

Dash, M. and H. Liu, 1997. Feature selection for classification. Intell. Data Anal., 1: 131-156.

Dash, M., K. Choi, P. Scheuermann and H. Liu, 2002. Feature selection for clustering: A filter solution. Proceedings of the IEEE International Conference on Data Mining, December 9-12, 2002, Maebashi City, Japan, pp: 115-122.

Dey, V., D.K., Pratihar and G. Lal Datta, 2011. Genetic algorithm-tuned entropy-based fuzzy C-means algorithm for obtaining distinct and compact clusters. Fuzzy Optim. Decis. Making, 10: 153-166.

Dy, J.G. and C.E. Bradley, 2004. Feature selection for unsupervised learning. J. Mach. Learning Res., 5: 845-889.

Eduardo, R., R.J.G.B. Campello, A.A. Freitas and A.C.P.L.F. de Carvalho, 2009. A survey of evolutionary algorithm for clustering. IEEE Trans. Syst. Man Cyber., 39: 133-155.

Eltibi, M.F. and W.M. Ashour, 2011. Initializing K-means clustering algorithm using statistical information. Int. J. Comput. Appli., 29: 51-55.

Erisogly, M., N. Calis and S. Sakalliglu, 2011. A new algorithms for initial cluster centers in k-means algorithm. Pattern Recognit. Lett., 32: 1701-1705.

Goldberg, D.E., 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, New York, USA.

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, CA.

Jimenez, J.F., F.J. Cuevas and J.M. Carpio, 2007. Genetic algorithms applied to clustering problem and data mining. Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization, September 15-17, 2007, Beijing, China, pp: 219-224.

Karegowda, A.G., S. Shama, T.R. Vidya, M.A. Jayaram and A.S. Manjunath, 2012. Improving performance of K-means clustering by initializing cluster centes using Gentic algorithm and Entropy based Fuzzy clustering for categorization of diabetic patients. Advances in Computing, M S Ramaiah Institute of Technology, Bangalore, India.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, December 27, 1965-January 7, 1966, Berkeley, USA., pp: 281-297.

Maulik, U. and S. Bandyopadhyay, 2000. Genetic algorithm-based clustering technique. Pattern Recogn., 33: 1455-1465.

Roth, V. and T. Lange, 2003. Feature selection in clustering problems. Proceedings of the Conference on Advances in Neural Information Processing Systems, December 9-14, 2002, Vancouver, British Columbia, Canada, pp: 1-8.

Sun, H.J. and L.H. Xiong, 2009. Genetic algorithm-based high-dimensional data clustering technique. Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery, August 14-16, 2009, Tianjin, China, pp: 485-489.

Yao, J., M. Dash, S.T. Tan and H. Liu, 2000. Entropy based fuzzy clustering and fuzzy modeling. Fuzzy Sets Syst., 113: 381-388.