A hybrid Genetic K-Means Algorithm forFeatures Selection to Classify Medical Datasets

Mohammed Abdullah Naser,
Zainab Falah Hasan and Esraa Abdalluh Hussein

mohamed_1276@yahoo.com
zainab_ga@yahoo.com, esraa_zd@yahoo.comand

University of Babylon/Collage of Science for Women / Computer Science Dep.

Abstract

Relevant features selection is become primary preprocessing step for building almost intelligence machine learning systems. Feature Selection (FS) is more and more important in many applications such as patterns recognition, medical technologies, data mining environments and others. The main objective of FS is to choice the important features among multi set in order to building effective machine learning models such as pattern analysis model by cancelling irrelevant or redundant attributes. An addition to that, there is a fact that the efficiency of the desired system is very sensitive to choose of the features that effect on classification or any analysis procedure of small or high dimensional datasets. Furthermore, the analysis of medical datasets has become growing claiming problem, due to huge datasets that cause time consuming and uses additional computational effort, which may not be suitable for many applications.

This work attempts to introduce a hybrid genetic k-means of feature selection algorithm for multi medical diseases datasets. The proposed algorithm uses a genetic algorithm combine with k-means algorithm as a powerful tool to select the relevant features from different large medical datasets of Mirjan hospital diabetes, heart and breast cancer diseases which play the important role in maximum the classification accuracy and efficiency of the system. Experimental results show the efficiency of the proposed system for the used datasets and satisfy maximum classification accuracy performance compared with others states.

Keywords: Genetic Algorithm, Features Selection, K-means and Classification.

الخلاصة

اصبحت عملية اختيار اهم خصائص البيانات من الخطوات الرئيسية في بناء اغلب انظمة التعلم الالي الذكي. اذ ان استخلاص تلك الصفات مهم جدا في عدة تطبيقات مثل تمييز الانماط، التقنيات الطبية، مجالات تنقيب البيانات وغير ذلك. الهدف الرئيسي من العملية المذكورة هو لاختيار افضل الخصائص من بين عدد منها لبناء نماذج الية كفوءة كنموذج تحليل الانماط من خلال اهمال او استبعاد الخصائص الغير مهمة او المكررة. ايضا فان هناك حقيقة مفادها ان كفاءة تلك الانظمة تتأثر بشكل كبير بنوعية الصفات المختارة والتي بدورها تنعكس على عملية تصنيف او عنقدة او اية عملية تحليل للبيانات الصغيرة او الكبيرة. من جانب اخر فان تحليل البيانات الطبية بات امرا ملحا وضروريا نتيجة النمو الهائل لهذا النوع من البيانات ، الامر الذي يسَبب هدرا كبيرا للوقت ويفضي الى جهد حسابي مضاعف ،وهذا مما لايناسب عدة تطبيقات.

العمل الحالي هو محاولة لتقديم خوارزمية هجينة تعتمد الخوارزمية الجينية وخوارزمية الكيمينز (k-means) لعدد من البيانات الطبية.الخوارزمية المقترحة تستخدم الخوارزمية الجينية مع خوارزمية الكيمينز كأحد الوسائل

الكفوءة في اختيار الخصائص المهمة و لعدد من البيانات الطبية مثل بيانات مرض السكري لمستشفى مرجان الطبية، وبيانات القلب وسرطان الثدي ، لما تلعبه من دور اساسي في زيادة دقة وكفاءة النظام المقترح .النتائج التجريبية اثبتت كفاءة النظام المقترح على البيانات المستخدمة ،كما وحققت دقة عالية بالمقارنة مع غيرها من الاجراءات.

1. Introduction

Feature selection plays a main role in the data analysis process since irrelevant features often degrade the performance of algorithms designed to data characterization, rule extraction and construction of predictive models, both in speed and in predictive accuracy. The goal of the feature selection process is, given a dataset described by *n* attributes (features), to find the minimum number of relevant attributes which describe the data such as the original set of attributes do [1].

Usually the classification field problems require selection of attributes or features to prepare the patterns to be classified; therefore, the feature or attribute selection procedure is very important which selects the informative features for used classification process or other machine learning techniques. The methods of features selection aim for selecting small set of features leading to the best performance of the classifier. In many applications, there is a practical need to reduce the number of measurements without significantly degrading the performance of the system [2].

Different methods have been used for feature selection in the literature such as breadth first search and branch and bound algorithms. These algorithms gave good results with conventional statistical classifiers; so, they could not achieve expected results with non-linear classifiers. In addition to these approaches, heuristic search and randomized population based search techniques were also used. In recent years, a feature selection algorithm using genetic algorithm was presented. Moreover, a hybrid genetic algorithm for feature selection was developed which performed better than simple genetic algorithms [3].A way to enhance the performance of a model that combines genetic algorithm and k-means algorithm for relevant feature selection and clustering analysis respectively is proposed in this work. Early diagnosis of any disease with less cost is preferable.

2-Related Works

There are many algorithms of features selection which identify the features that are relevant but not redundant to the solution. In paper [4] has proposed a medical diagnosis model that combines genetic algorithm and fuzzy logic technique for feature selection and classification operations. This system use diabetes datasets to identify and select useful subsets of pattern features from its larger set of features. While the fuzzy logic uses in classification process. The proposed system shows to improve the accuracy of classification. The paper [5]is presented a method that using a genetic algorithm(GA) to select a subset of relevant features combinations from small or high dimension medical datasets for improve the classification accuracy using. The combinations of these features are used for classification and then the accuracy of classification from support vector machine classifier to define the fitness

in GA. The results show the effectiveness of the proposed method for small and high dimension datasets. The paper [6] is introduced a new algorithm which is based on mimetic evolutionary idea that uses accurate set of fuzzy if then rules from cancer disease datasets that can classify gene expression data. This work is the first proposal of memetic methods with the multi view fitness function approach. The proposed algorithm classifies the tumors as (cancerous or benign) in efficiently and has good classification accuracy. The paper [7] presents a genetic programming (GP) based method has been used for classification of diabetes disease datasets. The GP has been used to generate new features from each subset of original datasets by combine of the current features, without prior information of the probability distribution. The performance of classification is computed using k-nearest neighbor and support vector machine classifiers. The experimental results of this method show a good performance over other methods. In paper [8] has provided the combining a feature selection based on genetic algorithm (GA) and support vector machines (SVM) for classification of medical disease data. The proposed GA-SVM classifier is used to the relevant subset of features that can improve the SVM classifier. The proposed method achieves better results than other methods. In paper [9] is presented an algorithm for cancer cell disease classification by using genetic an algorithm for feature selection (cells) and then classify these cells into either healthy or cancer. The selected best features are used by support vector machines classifier on the training dataset for classification. The results show the effectiveness of proposed method is gives better accuracy for feature selecting that select 20 features.

3. Genetic Algorithms

Genetic Algorithms are the heuristic search and optimization techniques that mimic the process of natural evolution. The basic concept of genetic algorithms is designed to simulate processes in natural system necessary for evolution [10]. Genetic algorithm exploits historical information to direct the search into the area of better performance within the search space. Strength of genetic algorithm comes from that its can explore the solution space in multiple directions at once, they are well-suited to solving problems having huge search space and they perform efficiently in problems for which the fitness landscape is complex[11].Genetic algorithm contains a set of individuals (population) in every generation. Each individual is a potential solution to the problem. Initially, set of solutions are randomly generated (initial population). The number of solutions in population (population size) depends on the nature of the problem [12].Each individual is measured by fitness function to get its performance. Fitness function is a function which determines how well each individual solves the problem. Selection operation is used to determine which solutions are to be preserved and allowed to reproduce. The main goal of the selection operator is to emphasize the good solutions and eliminate the bad one in a population. There are different techniques of selection such as Tournament selection, Roulette wheel selection, Proportionate selection and Rank selection [13].

Some individuals undergo random transformations by genetic operations to get new individuals. There are two types of transformation [14]:

1. Crossover, which forms new individuals by interchange parts from two individuals. The crossover points of any two chromosomes are selected randomly [15].
2. Mutation, which forms new individuals by making changes in a single individual.

The new individuals are then measured its efficiency. A new population is formed by selecting the more fit individuals from the parent population and children population transformation [14].After many generations, genetic algorithm access to the best individual (solution), this is an optimal or suboptimal solution to the problem. The flowchart of a GA work is presented in the Fig.1 [2].

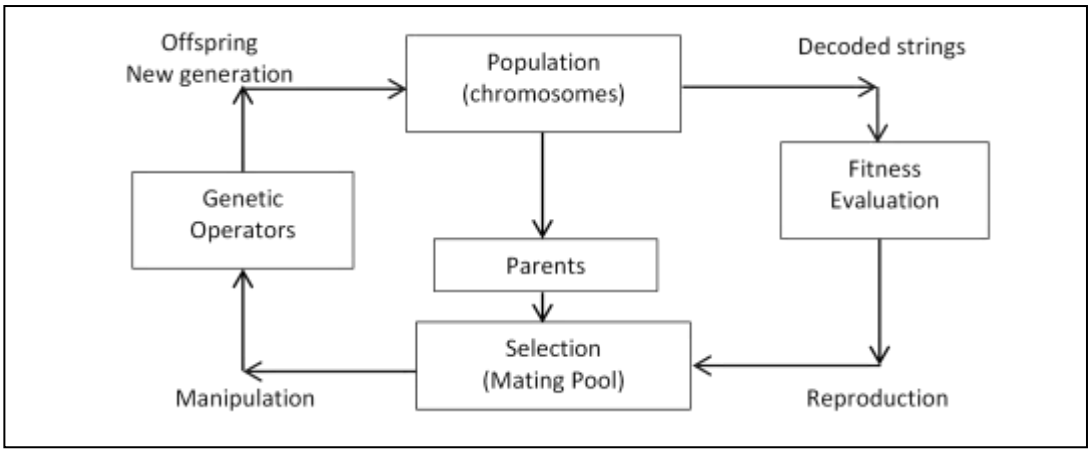

Fig. 1: The Block diagram of Genetic Algorithms.

4. K-Means Clustering Algorithm

K-means clustering algorithm proposed by Mac Queen in 1967 belongs to partitioning methods, which is widely used because of its simpleness and fast convergence [16].The advantage of k-means clustering is simple and easy to implement. Similarly, the drawbacks are of how to determine the number of clusters and decrease the numbers of iteration [17].

The basic step of k-means clustering is simple. In the beginning, determine number of cluster k and assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first k objects can also serve as the initial centroids, then the k-means algorithm will do the steps below until convergence:

1 - Compute the distance between each point from database to center of each cluster.
2 - Grouping the point to cluster with small distance.
3 - Recomputed center of each cluster based on taking the mean for all points in cluster.
4 - Go to step 1, until convergence is achieved.

5- Proposed Algorithm

As mentioned, the geneticalgorithm is technique for multi proposes such as improvement, search and learning. In this work, genetic algorithm using to search the

best data (features) andchoose minimum of it.Thisminimum data used forclassification of medical disease databases (Mirjan hospital diabetes, heartand breast cancer).

The data represents features of state usedby k-means algorithm. Because of large size of features for data, we need to minimize it for desired process; GA used for this purpose. The main steps of this work explained by the following:

1- Generate initial minimum features.
2- Efficiency measuring of the initial features.
3- Explore other features for data.
4- Evaluation of new features.
5- Loop the steps 3-4 until the best chromosome which has minimum features of best classification has been gotten.

The diagram in (Fig.2) explains the works of the proposed system.

5.1Generate Initial Features:

This step represents generating initial population of chromosomes where the chromosome acts the features of one state. Every gene acts as one feature. If feature is active then it is using in classification. The chromosome explained by the following figure (Fig.3).

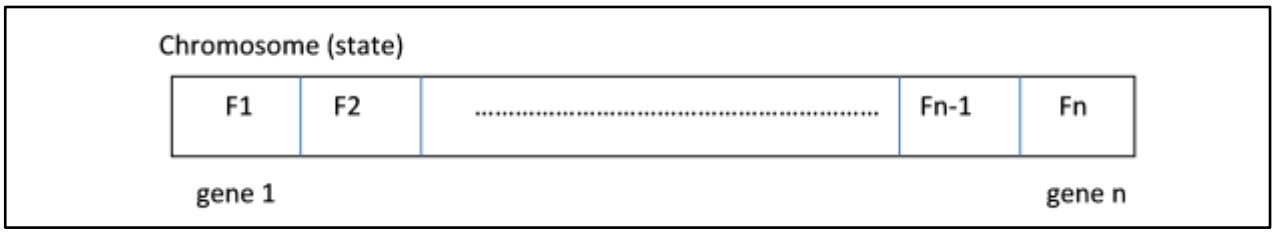

Fig. 3: Chromosome Representation.

Chromosome is encoding as binary. If gene has value 1 then feature is active otherwise feature is inactive. Minimum number of active features in chromosome with best classification is required in this work.

5.2 Selected Feature Evaluation (Chromosome Evaluation)

Every chromosome evaluated by using the k_means classification. This is done to get minimum features. The chromosome with some features and has maximum accuracy of classification represents the best chromosome.

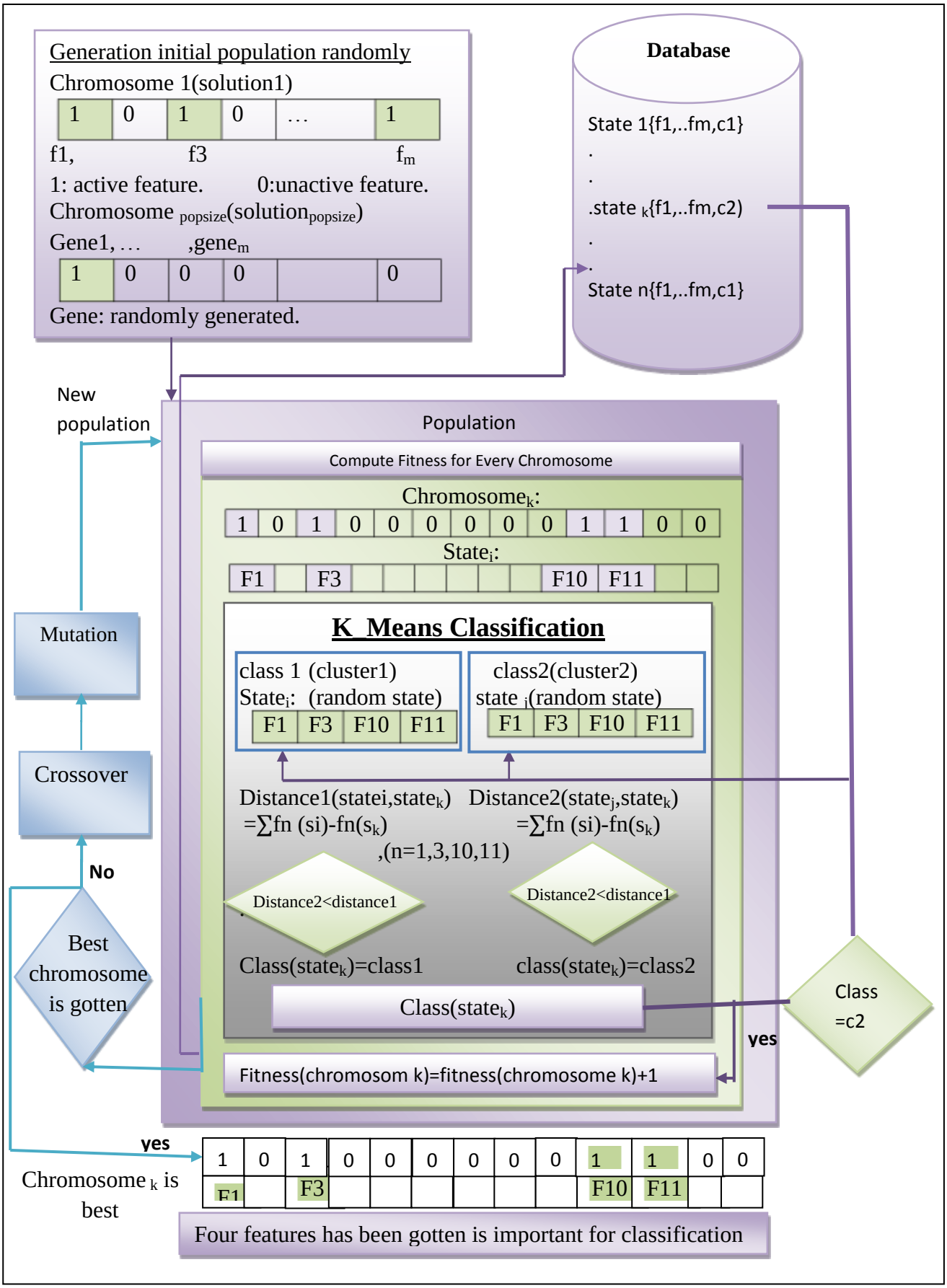

Fig. 2: Diagram of the Proposed System.

Accuracy of classification is computed by classification database using chromosome.
Accuracy of classification= No. of states has correct class / No. of all states * 100%.

5.3 Explore Other Features of Data

In this step, new chromosomes are generated using crossover and mutation operations. New chromosomes have been gotten to get better chromosome with minimum features for classification. 1x crossover is used to get new chromosomes where combination of old chromosomes has been done. Then 2m mutation is applied to improve parent's chromosomes (old solutions).

5.4 Evaluation of New Features

New chromosomes that have been gotten are **evaluated** to get performance of their and access to the best one. This performance is computed such as step 5.2. When best chromosome has been gotten, using it's for minimization the size of database, otherwise explored other new chromosomes. This process is continued until access the minimum number of data.

6. The Experimental Results and Performance Evaluation:

This section describes the experimental results obtained by applying the proposed algorithms to a variety of data sets. For experimentation, three datasets: Mirjan hospital diabetes from center of Marjan hospital in Hilla City/Iraq, also heart and breast cancer data sets) are taken from the UCI machine learning repository as shown in Table 1. The Breast Data set consists of 9 numeric attributes which include, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion Single, Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class respectively. The heart Data set consists of 13 numeric attributes which include, age, sex, chest pain type, resting blood pressure, serumcholesterol in mg/dl, fasting blood sugar> 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy,thal: 3 = normal; 6 = fixed defect; 7 = reversible defect and Class respectively. The Marjan hospital diabetes dataset consists of 26 numeric attributes which include, Age, Gender, Marital , Education, DMDx, Income, Job, BMX, Knowledge, Knowledge Source, HowDMDiagnosed, S1, S2, S2Insuline , S2Drug, S3, S3Hypo, S3Hyper, S3Urine, S3Blood , S4 , S5, S6, S6Foot , S6Habits, S7 and Class respectively.

Table 1: Information of Database.

Database	No. of states	No. of features	Class	No. of patient and not patient
Breast disease	684	9	Integer valued 2 (benign) and 4 (malignant)	444 benign 239 malignant
Heart disease	270	13	0 not patient 1 patient	120 not patient 150 patient
Marjan hospital diabetes	85	26	1 negative 2 positive	38 positive 47 negative

To determine the effect of feature selection on the model, we also established a prediction model without feature selection. Table 2 shows the comparisons of model structure and prediction results between these two models.

Table 2: Comparison between the model with GA-feature selection and model without GA-feature selection.

Database	Accuracy using k_means with all features	Accuracy using k_means with some features	No. of features after selection
Marjan hospital diabetes	80	98%	10 {2,3,6,9,14,16,20,21,22,24}
Heart disease	64%	87%	6 {2 3 9 10 11 12}
Breast cancer	95%	98%	5 {1 2 3 4 7}

A confusion matrix that summarizes the number of instances predicted correctly and incorrectly by a classification model. Three classical evaluation metrics of Precision, Recall and F-score are used to evaluate the efficiency of the proposed method. The three metrics are traditionally defined for classification task with positive and negative classes.

True positive (TP) = number of positive samples correctly predicted.

False negative (FN) = number of positive samples wrongly predicted.

False positive (FP) = number of negative samples wrongly predicted as positive.

True negative (TN) = number of negative samples correctly predicted.

Precision is the proportion of positive predictions that are correct, and recall is the proportion of positive samples that arecorrectly predicted positive. That is:

Precision = TP/ (TP+FP, Recall=TP/(TP+FN)

F-score= (2*precision*recall)/(precision+recall)

Table 3: Evaluation k_ means system using three performance metrics.

Dataset	TP Rate	FP Rate	Precision	Recall	F measure
Breast	224	11	0.95	0.93	1.85
Marjan hospital diabetes	81	**10**	0.75	0.81	1.61
Heart	57	33	0.63	0.47	1.94

Table 4: Evaluation proposed system using three Metrics.

Dataset	TP Rate	FP Rate	Precision	Recall	F measure
Breast	234	12	0.95	0.99	1.97
Marjan hospital diabetes	38	2	0.95	0.97	1.92
Heart	96	12	0.88	0.78	1.55

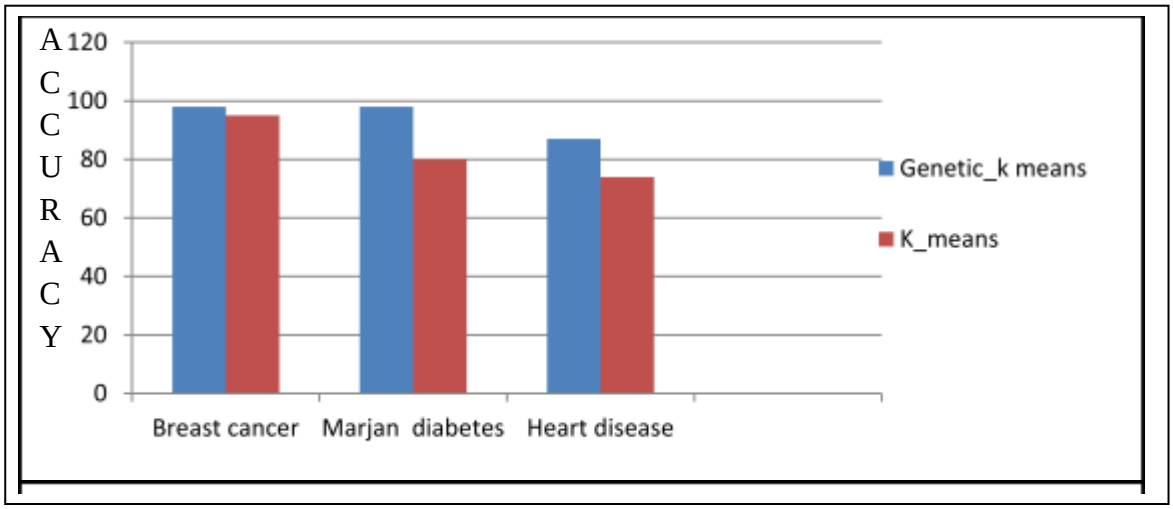

Fig. 5: Compares the Genetic K_means algorithm with K-means classifier.

7. Conclusions and Future works

A major goal of this work is to propose an efficient feature selection method that finding and selecting informative features from small or high dimension data which maximum the classification accuracy according to many performance evaluation metrics.

Feature selection is very important part of pattern recognition and all machine learning techniques. The obtained results show that the proposed method is capable of selecting feature subsets with efficient classification performance, where the features with highest individual performance are selected. The proposed Genetic K-means algorithm is efficiently selected features from different databases with large data and it is minimizing the features that effect on classification of states. Our algorithm selects important features from database by using its affecting on

classification; because it exploits the benefits of the combined algorithms in desired goal. Therefore, the results are promising for biologists as well as the algorithm shows to be perfect and it has high accuracy performance.

In the future, we suggest using multi classifier techniques to enhance the system efficiency and performance and it can be used for many classification fields.

8. References

[1]Vipin Kumar and SonajhariaMinz, Feature Selection: A literature Review, Smart Computing Review, vol. 4, no. 3, June 2014.

[2] Amit K., Artificial Intelligence and Soft Computing, CRC Press LLC, Vol. 2, 2000.

[3] Süveyda Y., Reyyan Y., Alper K., And Uğur S., Feature Selection with Genetic Algorithms on Cardiac Arrhythmia Database, 2007.

[4] E.P.Ephzibah,Cost Effective Approach On Feature Selection Using Genetic Algorithms And Fuzzy Logic For Diabetes Diagnosis, International Journal On Soft Computing (IJSC), Vol.2, No.1, February 2011

[5] MohdSaberiMohamad, SafaaiDeris, Safie Mat Yatim and Muhammad Razib Othman, Feature Selection Method Using Genetic Algorithm For The Classification Of Small And High Dimension Data,First International Symposium on Information and Communications Technologies. October 7-8, 2004.

[6] A. Zibakhsh, M. S. Abadeh, Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function, Engineering Applications of Artificial Intelligence 26, Elsevier, 2013.

[7] Muhammad W.Aslama, Zhechen Zhu, Asoke K. Nandi, Feature generation using genetic programming with comparative partner selection for diabetes classification, Expert Systems with Applications 40, Elsevier, 2013

[8] G.Ravi K., G.A.Ramachandra and K.Nagamani, An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.

[9] Tanzeem K.Mansoori ,Amrit S., and S.K. Mishra,Feature Selection by Genetic Algorithm and SVM Classification for Cancer Detection,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 9, September 2014.

[10] ChetanChudasama, S. M. Shah and Mahesh Panchal, Comparison of Parents Selection Methods of Genetic Algorithm for TSP, Proceedings published by International Journal of Computer Applications (IJCA), 2011.

[11] Amol C. A. and Rajankumar B., Hybrid Genetic Algorithmic Approaches for Personnel Timetabling and Scheduling Problems in Healthcare, International Conference on Technology Systems and Management (ICTSM) , 2011.

[12] Pratibha B. and Manoj K., Genetic Algorithm – an Approach to Solve Global Optimization Problems, Indian Journal of Computer Science and Engineering, Vol. 1 No. 3, 2010.

[13] Rakesh K. and Jyotishree, Effect of Polygamy with Selection in Genetic Algorithms, International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, ISSN: 2231-2307, 2012.

[14] Muhammad T. and M. A. Abido, Assessment of Genetic Algorithm Selection, Crossover and Mutation Techniques in Reactive Power Optimization, IEEE, 2009.

[15] Alessandro V. A. Luciano d. E., Andr´e L. L. A., Luciana P. A. and Carlos H. N. R., Analysis of Selection and Crossover Methods used by Genetic Algorithm-based Heuristic to solve the LSP Allocation Problem in MPLS Networks under Capacity Constraints, International Conference on Engineering Optimization Rio de Janeiro, Brazil, 01 – 05,2008.

[16]R. Harikumar , et.al., Performance Analysis For Quality Measures Using K-means Clustering And EM Models In Segmentation Of Medical Images, Int. J. of Soft Computing and Engineering , Vol. 1 , Issue-6, Jan. 2012.

[17]Moslem M. K., et.al., Applying New Method For Computing InitialCenters Of K-means Clustering With Color Image Segmentation, J.Thi-Qar Sci., Vol. 3, No. 1, 2011.