Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.366 – 372

RESEARCH ARTICLE

COGNITIVE DEVELOPMENT OF EVOLUTIONARY ALGORITHMS IN GENE PATTERN MINING

Helan Cynthiya Y.[1], Anusha M.[2], Dr. J.G.R Sathiaseelan[3]

[1,2,3]Department of Computer Science, Bishop Heber College, Tiruchirapalli, TN, India

[1] cynthi.helan@gmail.com; [2] anusha260505@gmail.com; [3] jgrsathiaseelan@gmail.com

Abstract— ***Pattern mining is one of the predominant areas of data mining where gene patterns can be extracted using different data mining techniques. The three main aspects through which pattern could be mined namely, kinds of patterns to be mined, the mining methodologies and their applications. Loosely speaking, different techniques use different algorithms for finding the interesting patterns from the datasets. This paper aims to present an extensive study on Genetic Algorithms in pattern mining that are used for finding the interesting patterns from both classification and clustering on medical datasets.***

Keywords— ***Gene Pattern Mining, Classification, Clustering, Genetic algorithm***

I. INTRODUCTION

Pattern mining is one of the significant areas of data mining where hidden patterns could be generated with the help of data mining technique [1]. It is also known as Frequent Pattern Mining (FPM) which could discover the patterns that are occurring frequently with some distinctive properties of inherent and valuable. There are three different approaches in pattern mining that includes kinds of patterns mined, mining methodologies and their applications. Firstly, the kinds of patterns to be mined include the basic patterns, multilevel and multidimensional patterns and also extended patterns. There are various pattern mining methods that have been used in various basic mining methods, interesting patterns, and distributed parallel and incremental methods. The patterns that are mined could be item-sets, substructures, subsequence and values. There are certain algorithms such as sequence based, tree, graph and so many that proved their efficacy in high dimensional data. Sequential pattern mining is a type of pattern mining where the patterns are mined from the sequence databases [2]. It uses several algorithms like Generalized Sequential patterns (GSP), Sequential Pattern Discovery using Equivalent class (SPADE), Colspan and many more for mining efficient patterns. Applications of pattern mining [3] include noise filtering and data cleaning in data intensive applications that discovers hidden inherent structures and cluster from the data. Also frequent patterns are effectively used in subspace clustering. The algorithm has been involved in analyzing various types of data namely, spatiotemporal data, video data, time-series data, image data, sequence or structural data, multimedia data and so on. Gene expression patterns have

been taken from microarray datasets that contains several cancer gene patterns, tumor gene patterns and many more.

Classification is a supervised learning of model that describes the different predetermined classes of data [4]. It involves in predicting an outcome based on input. In analyze the training dataset for generating a model based on class label [5]. Several algorithms have been developed for classifying the classes such as C4.5, K-nearest neighbor, Naïve Bayes, Apriori, AdaBoost algorithms. Classification process has been divided in to two steps. Firstly, building the model from training datasets and secondly, values are assigned to the model for obtaining model's accuracy [6].

Clustering is an unsupervised learning of data objects that groups the data of similar characteristics into a cluster by eliminating the irrelevant data objects [4]. It is used to find data segmentation and pattern information. The raw data is allowed to undergo certain clustering techniques in order to form clusters of data. General types of cluster are well-separated cluster, center-based clusters, contiguous clusters, density based clusters and shared property or conceptual clusters [7]. Different algorithms have been carried out for grouping clusters with similarities. Such algorithms are hierarchical, partitioning, grid-based algorithms. Hierarchical algorithms falls into two categories namely agglomerative and divisive methods. Partitioning algorithms includes k-means, medoids, Partitioning Around Medoids (PAM) and CLARA. Apart from these clustering algorithms the evolutionary algorithms such Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) are the recent techniques for providing near optimal solution that the previously mentioned algorithms. These have also been merged with both classification and clustering algorithms for improvising their performance in classifying the data and grouping them into clusters with similar datasets.

This paper is organized as follows: Section II describes the brief study on Genetic Algorithms, Section III discusses a massive survey for Pattern mining based on GA, Section IV depicts the performance analysis of GA and finally Section V draws a conclusion.

II. GENETIC ALGORITHMS

Genetic Algorithms is a heuristic search Algorithms with evolutionary ideas based on natural selection and genetics which has been inspired by Darwin's evolution theory [8]. GA simulate the survival of the fittest among individuals over consecutive generation for solving a problem. GA depends on the genetic structure and behavior of chromosomes within the population of individuals using the following basics such as Individuals in a population compete for resources and mates, The individuals with most successful probability in each 'competition' would produce more offspring than the individuals that perform poorly, Genes from 'good' individuals propagate throughout the population so that two good parents will sometimes produce offspring that are better than either parent, Thus each successive generation will become more suited to their environment.

Population of individuals is maintained within the search space. Each individual is coded as a finite length vector component represented as binary {0,1}. The set of genes forms a chromosomes and a fitness score is assigned to each individuals to compete. The individual with optimal fitness score will be taken into account for mating of parents to produce better offspring. New arrival of offspring replaces the individuals with least fitness score. This paves the way for the best solution in the successive generation. Fig.1 describes the Evolution flow of Genetic Algorithm.

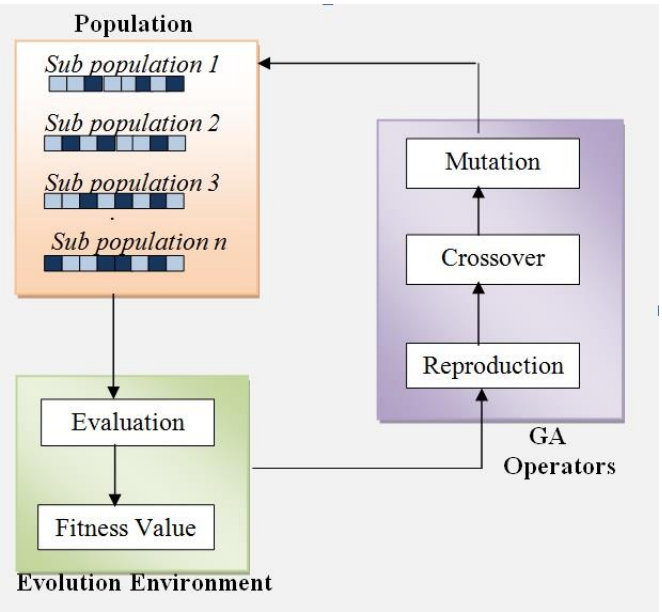

Fig.1. Evolution Flow of Genetic Algorithm

Population repository maintains the maximum number of chromosomes along with their fitness values. Parents with better fitness value are allowed for mating to produce offspring that inherit the characteristics from their parent. Representations of chromosomes, size of population, fitness function, selection, crossover, Mutation are discussed below:

A. Encoding Chromosomes:

Encoding is the process of representing the individual chromosomes in the form of numbers, trees, bits, lists and so on [9]. The Table I. represents the encoding types and their chromosome representation.

TABLE I. ENCODING CHROMOSOMES

S.No	Encoding Types		
	Types	***Chromosome 1***	***Chromosome 2***
1.	Binary Encoding	10110010110100	11010011101001
2.	Permutation Encoding	1 5 3 2 6 4 7 9 8	5 4 7 2 8 9 1 3 6
3.	Value Encoding	1.2324 5.3243	ABDJEIFJDHD
4.	Octal Encoding	07654398	23165749

B. Fitness Function:

The fitness function guides the GA to procure the best solution within the search space. Fitness function assigns fitness value to the individual. It is a problem dependent function. In order to calculate the fitness, firstly chromosome should be decoded and then the objective function is evaluated. This fitness value makes the chromosomes to the near optimal solution. Greater the fitness value is better the solution. The fitness value varies from problem to problem.

C. Selection Operator:

Selection operator selects the parents based on the fitness of each individual. The individual with higher fitness gets the chance for further genetic process viz. mutation and crossover. Several selection methods such as Roulette Wheel Selection (RWS) procure the chance to select the individual that is proportional to its fitness value, Stochastic Universal Sampling (SUS), Linear Rank Selection (LRS), Exponential Rank Selection (ERS) and many more has been used so far.

D. Crossover Operator:

The selected parents from the population with best fitness value are chosen. A crossover site is randomly chosen along the bit string. The position value of the two strings is swapped towards the crossover site. Different crossover techniques such as single point crossover, two point crossover, uniform crossover, and arithmetic crossover are carried out. The process of crossover operator is shown in Fig.2.

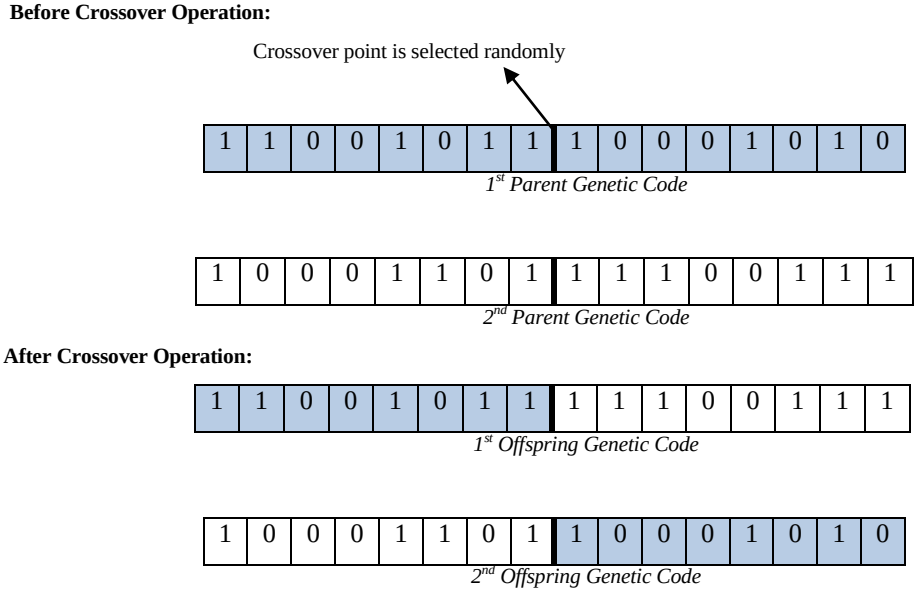

Fig.2. Crossover Operation

E. Mutation Operator

Mutation alters one or more gene values of the new off springs whose bits are flipped at lower probability. Mutation incorporates random walk throughout the search space. It maintains the diversity in the population and inhibits premature convergence. Real valued mutation, Binary mutation are the methods in mutation operator. The following Fig.3 represents the working of mutation operator.

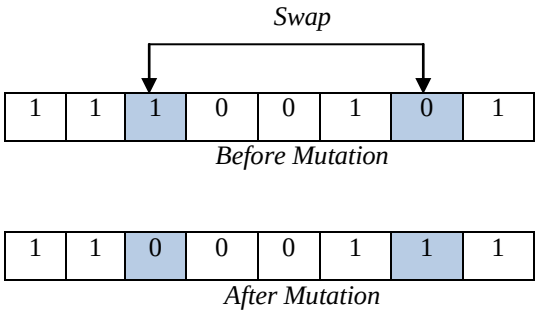

Fig.3. Mutation Operation

F. Pseudo Code for Genetic Algorithm

A general pseudo code for Genetic Algorithm [10] is represented as follows.

PSEUDO CODE FOR GENETIC ALGORITHM

Pseudo Code
1: ***Create*** initial population of individuals
2: ***Compute*** the fitness of each individual
3: ***Repeat***
4: ***Select*** individuals based on fitness value
5: ***Apply*** genetic operators to selected individuals for creating new individuals.
6: ***Compute*** fitness of each of the new individuals
7: ***Update*** the current population with new individuals
8: ***Until*** (stopping criteria)

III. LITERATURE REVIEW

Many works have been done to prove that the Genetic Algorithm plays an important role in mining the interesting patterns.

Kulkarni et al. [11] discussed the comparison between the accuracy of class prediction for two different classifiers namely, Genetic programming and genetically evolved decision tree. The experiment has been done on colon cancer dataset which has been taken from Kent Ridge Biomedical data repository. The best 50, 20, and 10 genes have been selected by two feature selection methods viz., t- statistic with standardized data and mutual information without standardized data. The results have proved that Genetic Programming as classifier with mutual information provides 100% accuracy for all the three best 50,20 and 10 genes. Similarly the accuracy rate of 98.33%, 100% and 98.33% for the best 50, 20 and 10 genes has been predicted using the t- statistic with Genetic Programming as classifier.

Jabbar et al. [12] proposed an efficient associative classification algorithm using genetic approach for the prediction of heart disease. The algorithm uses Gini index for attribute selection, Z-statistic for Hypothesis testing and crossover and mutation operation for accuracy computation. The experiment has been carried out using 6 datasets from SGI machine learning repository and two medical datasets from UCI machine learning repository for accuracy evaluation. The accuracy of heart disease prediction obtained by the proposed algorithm is about 98% that has minimized more random selection.

Alshamla et al. [13] analyzed the performance of Bio inspired evolutionary gene selection algorithm such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) with the traditional algorithms and proved their classification accuracy is superior with minimum number of selected genes than the traditional methods. The author listed the classification accuracy obtained using Bio inspired evolutionary gene selection algorithm and proved by evaluating the Classification Accuracy (CA) with

minimum Number of Genes (NOG) for 4 cancer microarray datasets such as Colon (CA-98.2%; NOG- 4), Leukemia (CA-100%; NOG-4), lung (CA-99.7; NOG- 5) and prostrate (CA-99.7; NOG-4).

Anusha et al. [14] depicted an enhanced K-means Genetic Algorithm for optimal clustering. The author overcomes the drawback of local optima with suitable dataset and also the algorithm fails in computational time. It is inferred that the algorithm produced more than 90% accuracy for real life datasets. This work has been extended by adopting neighborhood learning strategy for optimizing multi objective problems. This algorithm used k-means Genetic Algorithm to find the diversity and the compactness of the clusters. It is noted that the Algorithm could produce minimum shiloutee index value for the maximum datasets. However there is a need for proper feature selection for better more optimal solution [15].

Kabeer et al. [16] depicted a hybrid approach of Boosted Feature Subset Selection (BFSS) and Genetic Algorithm (GA). This approach has been applied on Colon, Acute Lymphoblastic Leukemia and lung cancer datasets. BFSS performs the preprocessing task for generating efficient random population and optimal subset of genes. The two classifiers KNN and SVM have been integrated in BFSSGA. The algorithm has been resulted by providing Classification accuracies of 87.47%, 87.54% and 86.71% for Leukemia, colon and lung cancer respectively with SVM classification method.

Mourad et al. [17] used Genetic Algorithm for mining the sequential patterns from the patient's prescription details using sequential interesting measures on Pharmacy database. The experiment has been done on the heart patient's prescription with1361 transaction and 50 patients taken from King Faisal Specialist Hospital and Research centre in RSA that resulted in the most suitable prescription sequence with crossover and mutation probability 0.7 and 0.001 respectively in order to reduce prescription error. The test result showed that GA procures the solution in less time when the generation increases and has stated that increase in generation also increases the average fitness value while the increase in population size does not guarantee the best performance. The experimental result of the proposed algorithm reduces the time and gives better fitness with increase in generation not in population size. The future work has been stated that this algorithm would be tested using categorical pharmacy database by taking time gap and patients state into consideration.

Korayem et al. [18] presented a hybrid Genetic algorithm and artificial immune system for selecting gens from high dimensional DNA microarray dataset. The proposed algorithm GA/AIS has been performed on colon, leukemia and lymphoma datasets. The best training accuracy for colon dataset is 97.78% with gene subset of 8 genes, 100% for leukemia dataset with 2 genes and 100% for Lymphoma dataset with 3 genes. The best test accuracy for all the three datasets is 100% with gene subset of 12, 2, and 5 genes respectively. This method is composed of two phases namely pre selection phase and Genetic search phase for classifying best gene subset. The average accuracy obtained by this method were 87.7±5.06, 98.33±1.87 and 96.6±2.25 for colon, leukemia and lymphoma cancer datasets respectively.

Dipankar et al. [19] presented a Multi Objective Genetic Algorithm (MOGA) based K-clustering method for optimizing the inter-cluster(separation, s)distance and intra-cluster (Homogeneity, H)distance. The analysis has been done on four datasets taken from UCI machine learning repository. The author compared two clustering algorithm viz. K-means and K-modes with MOGA for their performance analysis and it is concluded that MOGA is much better than these two algorithms by providing lowest DB index of the population.

Marghny et al. [20] depicted an effective evolutionary clustering algorithm for the case study of Hepatitis C. Dataset of Hepatitis has been taken from the machine leaning warehouse of University of California. The author has performed some preprocessing tasks in order to summarize the best data for further evaluation using Genetic algorithm. Firstly the experiment has made to run on three real-life datasets such as vowel data, iris data and crude oil data. Based on the results obtained from the above dataset has provided squared-error by k-means. Later the experiment runs on Hepatitis C dataset with 19 fields in order to find whether the patients with Hepatitis were alive or dead. The result showed the total variability rate as 98% of the continuous dataset.

Peter et al. [21] discussed and analyzed the effectiveness of Multi Objective k-means Genetic Algorithm (MOKGA). The effectiveness of this algorithm has been analyzed by conducting experiments on seven datasets such as GSE12093, GSE9195, NC160 cancer data, Leukemia data, Fig2data, iris and ruspini datasets that has been taken from UCI repository. This proposed algorithm has been developed on the basis of Fast Genetic K-means Algorithm(FGKA) and Niched Pareto Genetic Algorithm. The Total Within-Cluster Variation(TWCV) value for each datasets have been calculated and the optimal number of clusters were obtained as the solution set. The author compared MOKGA with other methods for its performance analysis.

Mukhopadhyay et al. [22] proposed a novel encoding technique for searching best cluster centers using clustering technique through multi objective optimization. The proposed technique Multi Objective Gene Selection and Clustering (MOGSeC) on artificial dataset and two real life datasets viz. brain tumor and lung tumor. Quality of the Clusters has been determined by the percentage of classified pairs(%CP). The %CP value for the brain tumor dataset has been obtained as 86.93% with 5.2 average number of clusters and for the lung tumor dataset the %CP value is 78.33% with 5.2 average number of clusters which is very close to the actual number of cluster value i.e. 5.

IV. PERFORMANCE ANALYSIS

From the above analysis of various paper based on Genetic Algorithm for classifying cancer gene, it is inferred that the Evolutionary Genetic Algorithm plays a vital and efficient role in providing near optimal solution for predicting the cancer genes at the earlier stage.

The performance and the classification accuracy of GA [14], BFFSGA [16] and GA/AIS [18] for colon and leukemia datasets have been depicted pictorially in Fig. 4.

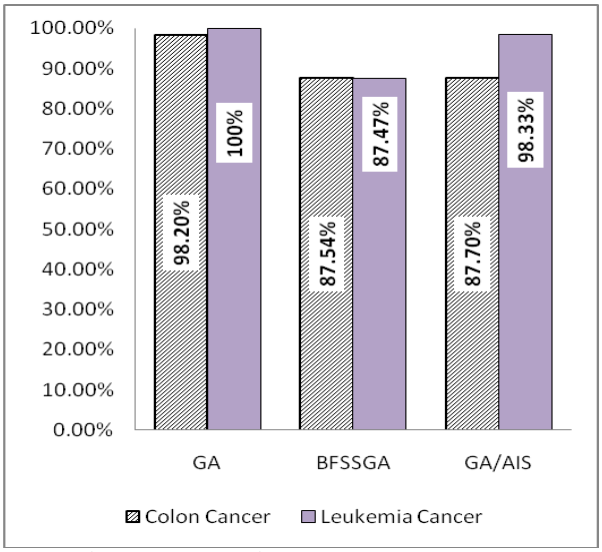

Fig.4. Classification Accuracy of Cancer datasets using Genetic Algorithm.

In the same way, clustering techniques also plays an important part in finding the best clusters of cancer genes. The result from clustering based Genetic algorithm is depicted by the Classified Pair (%CP) value from MOGSeC [22] is described in Fig. 5.

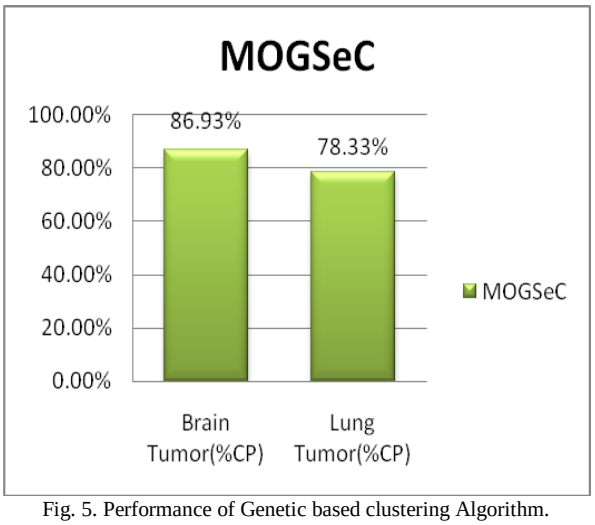

Fig. 5. Performance of Genetic based clustering Algorithm.

V. CONCLUSIONS

Mining frequent perfect patterns is an important and most predominant task in Data mining. In this paper, several Genetic Algorithms based on gene patterns have been extracted and the classification accuracy and the clustering results for different cancer datasets have been discussed and their performance analysis has been portrayed with a pictorial representation. From the analysis, it is inferred that Genetic Algorithm is the most efficient Evolutionary Algorithm for obtaining near optimal solution for any complex datasets. Our future work is based on the classification of cancer genes along with the clustering methods using Genetic Algorithm that could enhance the accuracy of gene patterns in medical dataset.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques",*Elsevier*, 2012.

[2] Vishal S.Motegaonkar and Prof. Madhav V. Vaidya, "A Survey on Sequential Pattern Mining Algorithms", *International Journal of Computer Science and Information Technologies*, ISSN 0975-9646, pp.2486-2492, 2014.

[3] Mohammed Al Hasan, "Summarization in Pattern Mining", *IGI Global*, pp.126-132, 2009.

[4] R.S. Santos, S.M.F. Malheiros, S. Cavalheiro, J.M. Parente de Oliveira, "A Data Mining system for providing analytical information on Brain tumors to public health decision makers", *Computer Methods And Programs in Biomedicine*, ISSN:0169-2607, pp. 296-282, 2013.

[5] Ritika, "Research on Data mining Classification", *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN:2277 128X, pp. 329-332, 2014.

[6] Samir Kumar Sarangi, Dr.Vivek Jaglan, Yajnaseni Dash, "A review of clustering and classification Techniquesin Data Mining", *Intenational Journal of Engineering, Business and Enterprise Applications*,ISSN:2279-0020, pp.140-145, 2013.

[7] R.Roseline, G.Jenitha, J. Henri Amirhtaraj, "Analysis and Application of Clustering Techniques in Data Mining", *International Journal of Computing Algorithm*, pp.910-912, 2014.

[8] Gunjan Verma, Vineeth Verma, "Role and Applicationof Genetic Algorithm in Data Mining", *International Journal of Computer Applications*(0975-888), 2012.

[9] Shaifali Aggarwal, Richa Garg and Dr. Puneet Gaswami, " A Review Paper on Different Encoding Schemes used in Genetic Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN: 2277 128X, pp.596-600, 2014.

[10] Richa Garg, Saurab Mittal, "Optimization by Genetic Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN:2277 128X, pp.587-589, 2014.

[11] Ashwinikumar Kulkarni, B.S.C. Naveen Kumar, Vadlamani Ravi, Upadhyayula Suryanarayana Murthy, "Colon Cancer Prediction with Genetic profiles using evolutionary techniques", *Elsevier*, pp.2752-2757,2011.

[12] M.Akhil jabbar, Dr. Priti Chandra and Dr. B.L Deekshatulu, "Heart Disease Prediction System using Associative Classification ang Genetic algorithm", *International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies*, 2012.

[13] Hala M. Alshamlan, Ghada H. Badr and Yousef A. Alohali, "The performance of Bio-Inspired Evolutionary Gene Selection Methods for Cancer Classification using Microarray Dataset", *International Journal of Bioscience, Biochemistry and Bioinformatics*, pp.166-170, 2014.

[14] M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", *IEEE ICCIC*, pp.580-584, 2014.

[15] M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", *International Journal of Applied Engineering Research*, pp. 228-231, 2015.

[16] Shaikh Jeeshan Kabeer, Moin Mhmud Tanvee, Mohammad Arifur Rahman, Abdul Mottalib, Md. Hasanum Kabir, "BFFS: Enhancing the performance of Genetic Algorithm using Boosted Filtering Approach", International Journal of Computer Applications, pp.29-34, 2012.

[17] Mourad Ykhlef and Hebah ElGibreen, "Mining Pharmacy Database Using Evolutionary Genetic Algorithm", *International Journal of Electronics and Telecommunications*, pp. 427-432, 2010.

[18] Mohammed Korayem, Waleed Abo Hamad, Khaled Mostafa, " A Hybrid Genetic Algorithm and Artificial immunse system for informative gene selection", *International Journal of Computer Science and Network Security*, pp.76-83, 2010.

[19] Dipankar Dutta, Paramartha Dutta, Jaya Sil, "Data clustering with mixed features by Multi Objective Genetic Algorithm", *IEEE*, pp.336-341, 2012.

[20] M.H. Marghny, Rasha M. Abd El-Aziz, Ahmed I. Taloba, "An effective evolutionary clustering algorithm: Hepatits C case study", *International Journal of Computer Applications*, pp.1-6, 2011.

[21] Peter Peng, Omer Addam, Mohamad Elzohbi, Sibel T. Ozyer, Ahmad Elhajj, Shang Gao, Yimin Liu, Tansel Ozyer, Mehmet Kaya, Mick Ridley, Jon Rokne, Reda Alhajj, "Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data", *Elsevier*, pp.108-122, 2014.

[22] Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamithra Bandyopadhyay, "Simultaneous Informative Gene Selection and Clustering through Multi objective Optimization", *IEEE*, 2010.

