A Distributed-Population Multiobjective Genetic Algorithm for Discovering Interesting Classification Rules from Medical Datasets

Basheer M. Al-Maqaleh[1], Waleed A. Omar[2]

[1]Department of Information Technology, Faculty of Computer Sciences and Information Systems, Thamar University, Thamar, Yemen

[2]Department of Computer Science, Faculty of Computer Sciences and Information Systems, Thamar University, Thamar, Yemen

Abstract: Automated discovery of classification rule is considered as one of the most fundamental and important approaches in order to obtain valuable knowledge from medical datasets. Building a comprehensible, accurate and interesting classifier for diseases diagnosis and prediction in medical field is one of the most significant challenges in knowledge discovery and data mining domain. A main concern of the population-based Genetic Algorithms (GA) approaches is how to balance exploration and exploitation during the process of evolutionary searching new solutions to avoid premature convergence to a sub-optimal solution. Distributed Genetic Algorithm (DGA) is considered as the most important classification approach to address the problem of simple GA converging to local optimal solutions. In this paper, a Distributed-Population multiobjective Genetic Algorithm (DPMoGA) approach for discovering interesting classification rules discover from medical datasets is proposed. The DPMoGA approach, has a flexible chromosome representation, an effective multiobjective fitness function, appropriate genetic operators for suggested representation, a new dynamic island model based on distributed population with an efficient migration operator. The DPMoGA approach is validated on several medical datasets from UCI repository, and the experimental results demonstrate the effectiveness of the proposed approach for interesting classification rules mining with significantly higher predictive accuracy rate**.**

***Keywords* - KDD, medical data mining, distributed GA, optimization, classification rule, predictive accuracy.**

1. INTRODUCTION

With rapidly increasing availability and size of the medical datasets in the recent decades have necessitated Knowledge Discovery in Database (KDD) and data mining approaches to extract valuable knowledge from these large datasets [1]. Discovery of knowledge from huge volume of medical datasets in order to help in diseases diagnosis and treatment is a real challenge and has become a research focus [1]. Classification is a fundamental task of data mining to predict a certain outcome based on a given input by constructing the underlying relationship between the attributes set and class label and identifies a model that best fits the training data [2]. A classification rule is a high-level knowledge representation of the form: **If** P **Then** D, where P is a conjunction of predicting attribute values and D is the predicted class. This kind of knowledge representation has the advantage of being intuitively comprehensible to the user. Classification rules mining is a form of classification where solutions to a given problem are classified, according to whether they lead to the desired outcome or not [3]. Medical data mining attempts to solve real world health problems in diagnosis and treatment of diseases, and to improve the accuracy of the discovered knowledge with huge amount of data [4]. It is important for interpreting, analysis and diagnosis, but complicated task that should be performed accurately, efficiently and its automation would be very useful [5]. Evolutionary Algorithms (EAs) and its applications to machine learning and data mining, and specifically to classification problems, have attracted the attention of researchers over the last decade [6,7,8]. EAs are search methods, inspired by natural evolution to find a reasonable solution for data mining and knowledge discovery problems [9]. Genetic Algorithm (GA) is a specialization of EAs, which it searches for good solutions to a problem by maintaining a population of candidate solutions and creating subsequent generations by selecting the current best solutions and using operators like crossover and mutation to create new candidate solutions [9]. Thus, better and better solutions are “evolved” over time. Commonly, the algorithm terminates when either a maximum number of generations has been produced or a satisfactory fitness level has been reached [10]. Generally, there are often problems that require simultaneously optimize more than one objective [8]. In multiobjective optimization the goal is to obtain a set of solutions that all equally fit for the optimum [11]. In classification rule mining, it is often of interest to simultaneously optimize more than one interestingness measure such as accuracy, support, novelty, and so on [12]. Interestingness is quantified via measures that select and rank patterns according to their potential interest to the user [13]. A main concern of population - based evolutionary algorithms is how to balance exploration and exploitation during the process of evolutionary genetic/searching new solutions [10]. Given a solution space, exploration mainly aims to search and evaluate solutions in new regions while exploitation mainly aims to search and evaluate neighbors of previously evaluated solutions [11]. Distributed Genetic Algorithm (DGA) is considered as the most important method to address the problem of simple GA converging to local optimal solutions at times [14]. This is done by special partitioning a huge population into several isolated islands, each evolving in parallel by its own space, and possibly exploring different regions of the search space. This method uses the migratory process that simulates the swapping of individuals belonging to different islands, in such a way to ensure the sharing of good genetic material [15]. An important characteristic on different islands, they are a very convenient structure for parallel or distributed architecture [16]. Therefore, it essential to propose parallelization strategy based on distributed- population multiobjective genetic algorithm approach using island model to handle appropriately larger medical datasets in order to discover interesting classification rules from these datasets with higher predictive accuracy by distributing the large population size into isolated-islands. In this paper, a Distributed Population Multiobjective Genetic Algorithm (DPMoGA) for interesting classification rules discovery from large medical datasets is proposed. In the proposed approach, a flexible chromosome encoding, appropriate genetic operators, a new island model with adequate migration operator, and suitable multiobjective fitness function based on interestingness measures will be developed. The proposed approach would generate interesting knowledge to get perfect diagnosis, enhance medical care and improve the quality of clinical decision support systems.

2. RELATED WORKS

Different data mining techniques have been commonly used in healthcare informatics and diagnosis applications since of their ability to predict new cases such as Support Vector Machines (SVM), ANN, decision trees, Bayesian networks, and regression analysis [17,18,19]. Classification is an important application area for data mining in medical domain [20,21]. An intelligent predictive system using classification techniques for heart disease diagnosis is presented in [19]. This predictive system consists of three classifiers namely; J48 decision tree, Naive Bayes, ANN and the obtained results proved that the J48 decision tree classifier is the best for heart prediction. The main motivation for applying GAs to KDD tasks is that they are robust and adaptive search methods, which perform a global search in the space of candidate solutions, so, GAs could discover interesting patterns that would have been missed by the traditional classification techniques [10,22,23]. Several GA designs have been made to apply GAs in medical domain such as [24,25,26,27]. These methods introduce the application of GA in disease diagnosis, treatment, prognosis, health care management and improve the quality of clinical decision support systems. In this context, a GA is presented in [22], to discover classification rules from datasets to be used in prediction and the discovered rules have a good of predictive accuracy and easy to understand (comprehensible. All GAs share the need to evaluate the fitness of possibly large population, as well as the need to balance their own mechanism for global and local search (i.e. exploration and exploitation) over the search space [12]. Thus, several DGAs have been proposed in [11,29]. A comprehensive survey of the state-of-the-art DEAs and models is presented in [15]. The study of these models helps guide future development of different improved algorithms. The main model of DEAs is the distributed-population model which can be partitioning the population into several semi - isolated nodes, each evolving simultaneously in separation to explore different regions of the search space, this helps EAs to maintains population diversity so as to repel local optimality [30,31]. In [14], a discovery of classification rules using DGA is introduced. In this approach the population is divided into five islands with fully connected graph. The obtained results confirm that the distributed GA discover classification rules with significantly higher predictive accuracy then the traditional GA. The approaches mentioned previously, did not consider the distribution of instances among the classes which affects their predictive accuracy. In the DPMoGA approach, however, the population is divided into varying number of islands depends on the number of class attribute values to maintain population diversity. In this case, all classes have their own islands and would participate in the evolutionary process equally.

3. THE DPMOGA APPROACH

The DPMoGA approach creates varying number of islands *k* depend on number of goal attribute values. It divides the entire population into *p* subpopulations, where each subpopulation is assigned to one island. The distributed model is organized as star topology with a central node and other nodes called islands. The central node works as an intermediate pool to save the migrant individuals between the islands. In each subpopulation, all individuals are associated with the same goal attribute value (island name). In other hand, this goal attribute value is fixed for all individuals of the same subpopulation. Each subpopulation evolves by the same Multiobjective Genetic Algorithm (MoGA) independently from the others (homogeneous island) except for some occasional migrations), which maintains the diversity of the search space. More precisely, the DPMoGA approach is simulated on a single processor, and isolated subpopulations in form of islands evolve locally for few specified numbers of generations (synchronous island, and then migration of rules takes place in-between subpopulations. Isolation among subpopulations is the key behind the island model as to maintain diversity and this is determined by migration process so frequency of migration and migration rate are important factors that influence the performance of any DGA [28]. Consequently, individuals in the DPMoGA approach are migrated after every migrate intervals of generations. Besides, the worst-best migration policy is used. That is, the worst of individuals i.e. Migration Rate (*MR*), in all subpopulations are selected and saved in an intermediate pool, then these individuals are redistributed by choosing the best individuals for each island, and replaced the worst migrant individuals. One advantage of this distributed population approach, with a fixed goal attribute value for each subpopulation, is to reduce the number of crossovers performed between individuals predicting different goal attribute values. Since, crossover is restricted to individuals of the same subpopulation, crossover swaps genetic material of two parents, which represents candidate rules predicting the same goal attribute value. As mentioned above, the DPMoGA approach creates different number of islands depends on the number of goal attribute values so, each class has its own island to participate in the evolutionary process in order to generate interesting classification rules for each class. The DPMoGA approach is specified as follows: -

- Create at random an initial Global Population (*GP*) with *N* chromosomes.
- Create *k* islands $\{l_1,...,l_k\}$ with subpopulations $\{p_1,...,p_k\}$ by choosing the best island for each selected chromosome from *GP* in such a way that maximized the fitness of the selected chromosome. The size of each sub-population is *M* (i.e. $M = GP/k$).
- Execute in parallel the MoGA on the subpopulations of each l_i, $i=1,...,k$ during *G* generations, where *G* is the migration interval (a parameter that controls how often the migration occurs).
- Pause evolution on all islands after *G* generations, and then apply migration operator based on the worst-best migration policy to update all islands with new subpopulations.
- Repeat steps 3 and 4 until no further improvement occur or a fixed maximum number of generations has been reached (*MG*).

In the DPMoGA approach, the training dataset assigned to a local MoGA will be a smaller percentage in size relative to the original dataset. For instance, a dataset using 70% partition for training and 30% for testing and 10 islands as configuration for data distribution will handle in each island such 7% of training data compared with the full dataset. The DPMoGA approach has to discover interesting classification rules by accessing the training set only. Once the training process is finished and the DPMoGA approach has found a set of interesting classification rules for each class, the predictive performance of these rules is evaluated on the test set, which was not seen during training. The remaining details of the DPMoGA approach are given below: -

3.1 Chromosome Representation

To solve an optimization problem, GAs start with the chromosome (string) representation of a parameter set. The search space, or the population, is a set of chromosomes on which genetic operators can perform, thus the most important and complex part of a rule is its condition on which it will execute the action [32]. The DPMoGA approach follows the Michigan approach where each chromosome (individual) represents a candidate classification rule of the form: **If** P **Then** D, where P is the rule antecedent and D is the rule consequent. P consists of a conjunction of conditions, where *z* is number of predictor attributes and each condition is an attribute-value pair of the form A_i = V_{ij}, where A_i is the *i*-th attribute and V_{ij} is the *j*-th value of the domain of A_i. An individual is encoded as a fixed-length string containing genes. Only a subset of the attribute values encoded in the genome will be decoded into attribute values occurring in the rule antecedent. Therefore, the genotype length is fixed, its decoding mechanism effectively represents a variable-length phonotype antecedent. This kind of representation gives a lot of flexibility to the rules being discovered. If the third value of A_i attribute is occurrence (j=3) in antecedent the value of attribute A_i appears in the chromosome is encoded as 3. In case the *i*-th attribute A_i is absent in the antecedent part then this attributes represented by 0. The consequent consist of a single condition of the form $D_k = V_{kl}$, where D_k is the *k-th* goal attribute and V_{kl} is the *l*-th value of the *k*-th goal attribute. Once the rule antecedent is formed, the proposed approach chooses the best consequent for each rule in such a way that maximizes the fitness of the rule. More precisely, it chooses the best possible consequent for the corresponding rule antecedent. In each subpopulation all rules are associated with the same goal attribute value (island name), so, there is no need to encode the consequent part in the chromosome encoding. Fig. 1 shows the chromosome representation.

Gene*1*	…	Gene *i*	…	Gene *z*
Value A*1*	…	Value A*i*	…	Gene Az

Fig. 1 Chromosome representation

For example, consider the Postoperative Patient classification training dataset given in Table 1.

Table 1. Description of the Postoperative Patient dataset.

Attribute	Possible Values	Alleles
L-CORE	High, Mid, Low	'1','2','3'
L-SURF	High, Mid, Low	'1','2','3'
L-O2	Excellent, Good, Fair, Poor	'1','2','3','4'
L-BP	High, Mid, Low	'1','2','3'
SURF-STBL	Stable, Mod-stable, Unstable	'1','2','3'
CORE-STBL	Stable, Mod-stable, Unstable	'1','2','3'
BP-STBL	Stable, Mod-stable, Unstable	'1','2','3'
DECISION	I, S, A	'1','2','3'

Corresponding to the data in Table 1 training dataset a classification rule (phonotype):-

If L-CORE = High ʌ L-SURF =Low ʌ L-BP= Mid ʌ BP-STBL= Stable **Then** DECISION = S.

This rule would be encoded to genotype as:

{1,3,0,2,0,0,1,2}.

3.2 Multiobjective Fitness Function

Quantifying the quality or fitness of an individual is the most important and deciding task in EA as the fitness determines whether an individual is selected to ultimately participate in the search of the optimal solution as the fittest individuals is the one closest to the optimal solution [9]. The multiobjective fitness function used consists of two terms [12]. The first one measures the degree of interestingness of the rule in an objective (data-driven, domain independent), while the other measures its classification accuracy. The term of interestingness consists of two parts. One of them is the interestingness of the antecedent of the rule and the other is interestingness of the consequent of the rule. The degree of the interestingness of the antecedent of the rule is computed by information theoretical measure as under

$$IA = 1 - \left[\frac{\frac{\sum_{i=1}^{n} InfoGain\,(A_i)}{n}}{\log_2(|dom(G_k)|)} \right] \qquad (1)$$

Here, *n* is the number of attributes in the antecedent; $(|dom(G_k)|)$ is the number of possible values of the goal attribute G_k occurring in the consequent. The *log* term is used to normalize the value of *IA* so that this measure takes on a value between 0:1. The *InfoGain* is given by:

$$InfoGain(A_i) = Info(G_k)Info(G_k|A_i) \tag{2}$$

where

$$Info(G_K) = -\sum_{j=1}^{m_k}(Pr(V_{kj})\log_2(Pr(V_{kj}))) \tag{3}$$

$$Info(G_k|A_i) = \sum_{z=1}^{n_i}\left(\Pr(V_{iz})\left(-\sum_{j=1}^{m_k} Pr(V_{kj}|V_{iz})\log_2(Pr(V_{kj}|V_{iz}))\right)\right) \tag{4}$$

Here, m_k is number of possible values of the goal attribute G_k, n_i is the number of possible values of the attribute A_i. *Pr(X)* states the probability of X and *Pr(X|Y)* states the conditional probability of X given Y. The rules whose antecedent contain attributes with low information gain are more interesting than rules whose antecedent contain attributes with high information gain. The computation of the degree of interestingness of the rule consequent (*IC*) is based on the idea that the discovering of minority goal attribute values tend to be more interesting to the user than the discovering of majority goal attribute values as they dispute the existing knowledge and have elements of unexpectedness and interestingness. The *IC* is computed using the following formula:

$$IC = (1 - Pr(V_{kl}))^{1/\beta} \tag{5}$$

where *Pr* (V_{kl}) is the prior probability (relative frequency) of the *l*-th value of the *k-th* goal attribute and *β* is a user-specified parameter, empirically set to 2 in our experiments. The exponent *1/β* in the formula (5) can be regarded as a way of reducing the influence of the rule consequent interestingness in the value of the fitness function. The second part of the fitness function measures is the classification accuracy (*Acc*) of the rule (**If** P **Then** D) as follow: -

$$Acc = \frac{|P\&D| - 1/2}{|P|} \tag{6}$$

where |*P&D*| is the number of instances that satisfy both the rule antecedent and the rule consequent, and |*P*| is the number of instances that satisfy only the rule antecedent. The term $1/2$ is subtracted in the numerator of formula 6 to penalize rules covering few training instances to bias the classifier to pay more attention to the minority classes. The DPMoGA approach represents the discovered knowledge in the form of "**If-Then**", which are symbolic knowledge presentation, easy, and comprehensible (understandable) by the users. Accordingly, the comprehensibility measure has been achieved. So, the multiobjective fitness function is computed as the arithmetic weighted mean of interestingness and classification accuracy as follows: -

$$Fitness = \frac{w_1.\frac{IA+IC}{2} + w_2.Acc}{w_1 + w_2} \tag{7}$$

where w_1, w_2 are user-defined weight, and were set to 1 and 2 respectively relative to the importance of interestingness and classification accuracy. Therefore, the goal is to maximize both the interestingness and classification accuracy as objective interestingness measures at the same time. Furthermore, the above multiobjective fitness function has the advantages of returning meaningful normalize value in the range [0:1].

3.3 Genetic Operators

Genetic operators are one of the most important components of GAs to maintain genetic diversity by introducing new genetic material and to manipulate or recombine the genetic material of candidate rule [22]. In the proposed approach, fitness proportional selection, one-point crossover with crossover rate of 75%, and mutation rate of 1% are used. Furthermore, am elitist reproduction strategy being used, where the best 5 individuals of each generation were passed unaltered to the next generation. A crossover rule determines how often individuals' mate. The two individuals are chosen through the selection operator. Hence, crossover swaps entire rule between individuals, but it cannot produce new gene, the mutation operator accomplishes the creation of new genes. The goal of the mutation operator is to maintain a good diversity that allows a continuous search towards a solution. This is needed to escape from local optima when the algorithm has got stuck in bad region of exploration. This operator randomly transforms the value of an attribute into another (different) value belonging to the domain of that attribute.

3.4 Migration Operator

Migration is the process that guides the exchange of individuals among islands in a DGA [15]. The DPMoGA approach has a migration operator where, from time to time, an individual of a subpopulation is copied into another subpopulation which extends the exploration in this distributed population. In the DPMoGA approach, the migration procedure sends 5% individuals as *MR* every 10 generations (migration interval), and a suitable migration operator is developed for classification rule discovery task. Migration takes place every *G* generations. Each subpopulation sends individuals to all the other subpopulations. More precisely, in each subpopulation p_i, *i = 1,..., k*, the migration procedure chooses *MR* individuals to be migrated, where *MR* is number of the worst chromosomes that will be migrate to other islands, sum of all chromosomes that will migrate from all subpopulations equals to *MR*k* where, *k* is the number of subpopulations or the number of islands, which equals to the number of class attribute values in the dataset being mind. These individuals are collected and saved in an intermediate pool. The accepted individuals by the destination island *i* are the *MR* individuals with the largest multiobjective fitness value. A migration process is applied to allow subpopulations to interact with others for finding potential global solutions from local solutions. This procedure is shown in Fig. 2.

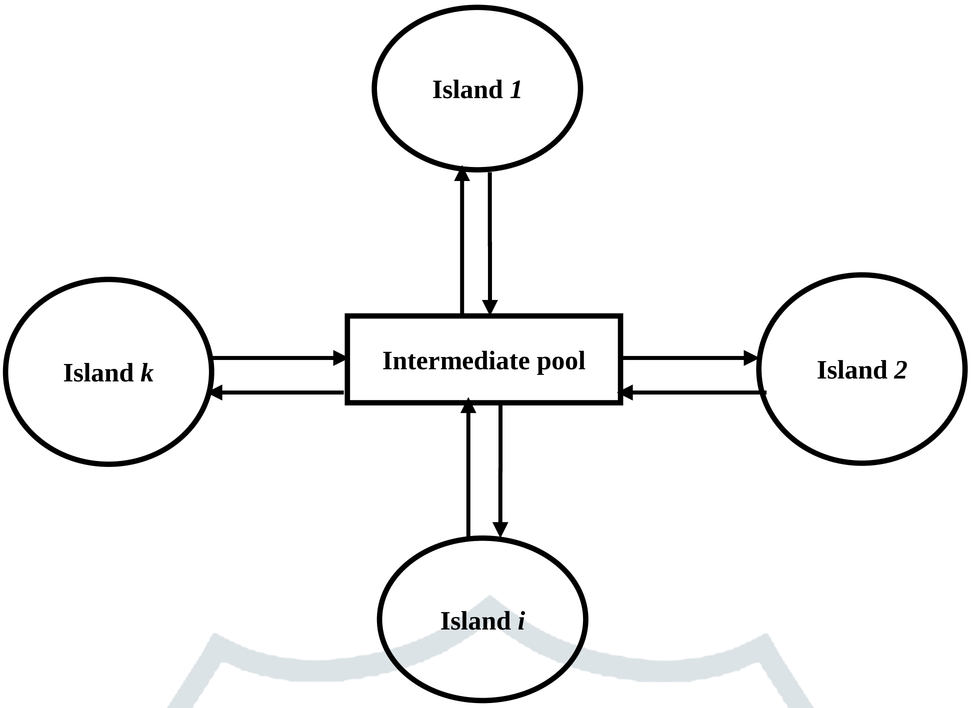

Fig. 2 Migration operator in the DPMoGA approach.

4. Computational Results

The performance of the DPMoGA approach is tested on real-world medical datasets from UCI Machine Learning Repository, which is a collection of widely used benchmark for data mining and KDD community [33]. The performance of the DPMoGA approach is evaluated and compared with a Standard (single-population) Genetic Algorithm (SGA) approach which is implemented in KEEL (a software tool to asses EAs in data mining problems) [34]. All experiments are performed with C# 2010, Windows 10, 4GB RAM and Core 2 Duo 2.00GHz Processor. The DPMoGA approach is developed to cope with nominal dataset. Any dataset contains continuous data is discretized using a public tool called WEKA. WEKA is a collection of machine learning algorithms for data mining task [35]. Also, the instances that had some missing values were removed from the datasets. Each island was run for a maximum of 2000 generations in a synchronous mode, however runs were stopped earlier in case of detecting a convergence i.e. no further improvement occurs in fitness values after 400 generations. For each class attribute value(island), those rules having higher fitness are collected. The parameters setting is an important task in GA based approaches, therefore after doing some experimentation for adjusting the parameters, so the final parameters values that are used in the following experiments are shown in Table 2.

Table 2. Optimized parameters of the DPMoGA approach.

Parameter	**Value**
Training set	75%
Test set	25%
General Population (*GP*)size	100**k*, *k*= number of islands
Population size per island	100
Maximum number of generations	2000
Stagnation limit	400 generations
Selection algorithm	Roulette Wheel Selection
Migration Rate (*MR*)	5%
Migration Interval (*G*)	10 generations
Mutation rate	1%
Crossover rate	75%
Elitism amount	5-elit individuals/ chromosomes

The performance of the DPMoGA approach on different dataset is demonstrated below:

4.1 Breast-Cancer-Wisconsin Dataset

The Breast-Cancer-Wisconsin Dataset (BCWD) was used in this experiment. The total number of instances are 699 and the number of attributes is 10 plus the class attribute. Each instance has one of two possible classes Benign or Malignant tumor. The class distribution for Benign is 458 instances (65.5%) and for Malignant is 241 instances (34.5%). As this dataset contains two class attribute values Benign and Malignant, so the DPMoGA approach would create two islands, one island for Benign and another for Malignant. The DPMoGA approach would discover five interesting classification rules for each island in 452 generations as shown in table 3.

Table 3. Result for the BCWD.

Class/ Island Name	Id	Generation No.	Discovered Interesting Classification Rules	Fitness
Benign	1	452	**If** Bland Chromatin = 1.0 ∧ Mitoses = 1.0 **Then** class = Benign	0.87
	2	452	**If** Clump Thickness = 1.0 ∧ Marginal Adhesion = 1.0 ∧ Mitoses = 1.0 **Then** class = Benign	0.86
	3	452	**If** Marginal Adhesion = 1.0 ∧ Bland Chromatin = 1.0 ∧ Mitoses = 1.0 **Then** class = Benign	0.86
	4	452	**If** Clump Thickness = 3.0 ∧ Marginal Adhesion = 1.0 ∧ Mitoses = 1.0 **Then** class = Benign	0.86
	5	452	**If** Single Epi Cell Size = 2.0 ∧ Bland Chromatin = 1.0 ∧ Mitoses = 1.0 **Then** class = Benign	0.86
Malignant	1	452	**If** Mitoses = 4.0 **Then** class = Malignant	0.87
	2	452	**If** Mitoses = 10.0 **Then** class = Malignant	0.86
	3	452	**If** Clump Thickness = 10.0 ∧ Mitoses = 1.0 **Then** class = Malignant	0.86
	4	452	**If** Normal Nucleoli = 10.0 ∧ Mitoses = 1.0 **Then** class = Malignant	0.86
	5	452	**If** Marginal Adhesion = 8.0 ∧ Mitoses = 1.0 **Then** class = Malignant	0.86

To evaluate the results, the DPMoGA approach validates the discovered interesting rules using the test data and then constructs the confusion matrix for BCWD as shown in Fig. 3.

		Predicted class		**Total Instances**
		Malignant	Benign	
Actual class	Malignant	27	0	27
	Benign	0	172	172

Fig. 3 Confusion matrix for BCWD.

The confusion matrix is a predictive accuracy measurement tool for data mining classification. In this context, the above confusion matrix for test dataset has 27 members of Malignant class and 172 members of Benign class are predicated correctly so, the predictive accuracy is computed using the following formula as under: -

$$\text{Predictive Accuracy } (\%) = (\text{TP} + \text{TN})/(\text{P} + \text{N}) = (27 + 172)/(27 + 172) = 100\%$$

So, the DPMoGA approach is able to set prediction error equal to 0% for the test dataset. Overall the confusion matrix reveals that the constructed classifier can identify correctly about 27 patients, which have cancer diseases and 172 patients which don't have cancer diseases from the test dataset.

4.2 Caesarian Section dataset

The Caesarian Section dataset was used in this experiment. This dataset contains information about Caesarian Section results of 80 pregnant women with the most characteristics of delivery problems in the medical field (6 attributes). As this dataset contains two class attribute values No and Yes, so the proposed approach would create two islands. It would discover five interesting classification rules for each class in 628 generations as shown in Table 4.

Table 4. Result for Caesarian Section dataset.

Class /Island Name	**Id**	**Generation No.**	**Discovered Interesting Classification Rules**	**Fitness**
No	1	628	**If** Heart Problem = Apt ∧ Delivery time = Latecomer **Then** Caesarian = No	0.93
	2	628	**If** Heart Problem = Apt ∧ Delivery time = Premature ∧ Blood of Pressure = Low **Then** Caesarian = No	0.93
	3	628	**If** Heart Problem = Apt ∧ Delivery time = Latecomer **Then** Caesarian = No	0.90
	4	628	**If** Heart Problem = Apt ∧ Delivery time = Timely ∧ Blood of Pressure = Normal **Then** Caesarian = No	0.90
	5	628	**If** Heart Problem = Inept ∧ Blood of Pressure = High ∧ Delivery time = Latecomer **Then** Caesarian = No	0.90
Yes	1	628	**If** Heart Problem = Inept ∧ Blood of Pressure = Low **Then** Caesarian =Yes	0.90
	2	628	**If** Heart Problem =Inept ∧ Delivery time = Premature ∧ Blood of Pressure = High **Then** Caesarian = Yes	0.90
	3	628	**If** Heart Problem = Apt ∧ Delivery time = Timely Blood of Pressure = Normal **Then** Caesarian = Yes	0.90
	4	628	**If** Heart Problem = Inept ∧ Blood of Pressure = Normal **Then** Caesarian = Yes	0.90
	5	628	**If** Heart Problem = Apt ∧ Delivery time = Premature ∧ Blood of Pressure = High **Then** Caesarian = Yes	0.90

The quality of the discovered interesting classification rules is evaluated based on the constructed confusion matrix on test dataset as shown in Fig. 4.

		Predicted class		**Total Instances**
		Yes	No	
Actual class	Yes	13	2	15
	No	0	3	3

Fig. 4 Confusion matrix for Caesarean Section dataset

This confusion matrix shows that, the interesting classification rules are discovered with 88.9% predictive accuracy, and 11.1% prediction error.

4.3 Indian Liver Patients dataset

This dataset contains 416 liver patient records (71.4%) and 167 non-liver patient records (28.6%), which is collected from north east of Andhra Paradesh, India. This dataset contains 441 male patient records and 142 female patient records, with all attributes. The predictor class has two values: 1 If patient has liver disease and 2 if they do not. The distributed population used handles this dataset by creating two islands, one island for each class. The best 5 interesting classification rules for each class during 597 generations are discovered as shown in Table 5.

Table 5. Result for Indian Liver Patients dataset.

Class/Island Name	**Id**	**Generation No.**	**Discovered Interesting Classification Rules**	**Fitness**
1	1	597	**If** Albumin = 1.8 **Then** Dataset = 1	0.89
	2	597	**If** Gender = Male ∧ Direct Bilirubin = 1.3 **Then** patient= 1	0.88
	3	597	**If** Age = 75 ∧ Gender = Male **Then** patient= 1	0.87
	4	597	**If** Direct Bilirubin = 1.3 **Then** patient= 1	0.87
	5	597	**If** Gender = Male ∧ Albumin = 2.7 **Then** patient = 1	0.87
2	1	597	**If** Age = 25 **Then** Dataset = 2	0.90
	2	597	**If** Total Bilirubin = 0.7 ∧ Albumin= 2.7 ∧ Globulin Ratio = 1.3 **Then** patient= 2	0.88
	3	597	**If** Gender = Male ∧ Total Bilirubin = 0.7 ∧ Albumin = 4.2 **Then** patient= 2	0.88
	4	597	**If** Gender = Male ∧ Total Bilirubin = 0.7 ∧ Albumin_ ∧ _Globulin Ratio = 1.3 **Then** patient= 2	0.85
	5	597	**If** Total Bilirubin = 0.7 ∧ Albumin = 4.2 **Then** patient= 2	0.85

To evaluate the performance of the DPMoGA approach on test data the confusion matrix is constructed as shown in Fig. 5.

		Predicted class		Total Instances
		Liver disease	No-Liver disease	
Actual class	Liver disease	90	0	90
	No-Liver disease	10	0	10

Fig. 5 Confusion matrix for Indian Liver Patients dataset.

This matrix shows that, 100 instances are tested, 90 instances are correctly classified as liver disease, and 10 instances are healthy individuals incorrectly identified as liver disease so the predictive accuracy is 90% and prediction error is 10%.

4.4 Hepatitis Dataset

This dataset contains 155 hepatitis patient records and 123 is on the live patient records (79.4%) and 32 patient records is Die (20.6 %). This dataset contains 139 male patient records and 16 female patient records and 19 attributes plus class attribute with two values Live and Die. The discovered interesting classification rules from both islands in 545 generations are reported in Table 6.

Table 6. Result for Hepatitis dataset.

Class/ Island Name	Id	Generation No.	Discovered Interesting Classification Rules	Fitness
Die	1	545	**If** Sex = male ∧ Steroid = no ∧ Malaise = no ∧ Anorexia = yes ∧ Liver big = yes ∧ Protime = low **Then** Class = Die	0.93
	2	545	**If** Sex = male ∧ Malaise = no ∧ Anorexia = yes ∧ Liver big = yes ∧ Protime = low **Then** Class = Die	0.93
	3	545	**If** Sex = male ∧ Steroid = no ∧ Fatigue = no ∧ Anorexia = yes ∧ Liver big = yes ∧ Protime = low **Then** Class = Die	0.93
	4	545	**If** Sex = male ∧ Steroid = no ∧ Fatigue = no ∧ Malaise = no ∧ Anorexia = yes ∧ Liver big = yes ∧ Protime = low **Then** Class = Die	0.93
	5	545	**If** Steroid = no ∧ Antivirals = no ∧ Malaise = no ∧ Anorexia = yes ∧ Liver big = yes ∧ Protime = low **Then** Class = Die	0.93
Live	1	545	**If** Steroid = yes ∧ Anorexia = yes ∧ Liver big = yes ∧ Spleen Palpable = yes **Then** Class = Live	0.93
	2	545	**If** Steroid = yes ∧ Anorexia = yes ∧ Liver big = yes ∧ Varices = yes **Then** Class = Live	0.93
	3	545	**If** Malaise = yes ∧ Anorexia = yes ∧ Liver big = yes ∧ liver firm = yes ∧ Spleen palpable = yes **Then** Class = Live	0.93
	4	545	**If** Sex = male ∧ Malaise = yes ∧ Anorexia = yes ∧ Liver big = yes ∧ Liver firm = yes ∧ spleen palpable = yes **Then** Class = Live	0.93
	5	545	**If** Steroid = yes ∧ Anorexia = yes ∧ Spleen palpable = yes **Then** Class = Live	0.93

The evaluation of discovered rules via confusion matrix using test data is shown in Fig. 6.

		Predicted class		Total Instances
		Die	Live	
Actual class	Die	8	2	10
	Live	0	22	22

Fig. 6 Confusion matrix for Hepatitis dataset.

This confusion matrix shows that the DPMoGA approach discovered the rules with 93.8% predictive accuracy, and 6.2% predictive error.

5. Comparative Study

The comparative study has been done using *MR* and predictive accuracy as demonstrated below: -

5.1 Migration Rate

It is well-known fact that the quality of the solution found by any GA is directly dependent on the value of its parameters [9]. At the centre of such approach lies the migratory process that simulates the swapping of individuals belonging to different islands, in a such way to ensure the sharing of good genetic material. So, the focus is on the characterization of migration process in which the choice of what individuals to migrate and its number affect the results. So, the impact of *MR* on the performance of the proposed approach is analysed. It can be concluded from the experiments that if *MR* is very low, then it works like stansard GA. The performance of the DPMoGA approach also decreases if migration rate is increased above the limit because the diversity decreases at higher *MR*. Big migration tends to have a direct effect on diversity simply because of replacing a larger number of individuals. However, the diversity does not change much in following generations. This mimics the real-world migration phenomena, in which countries allow people who meet certain qualification to get in. The optimal *MR* for the DPMoGA approach is 5 for all datasets. Table 7 shows the impact of *MR* and the predictive accuracy results at different *MR* values.

Table 7 Predictive accuracy with different *MR* values on various datasets.

Dataset	*MR*=1	*MR*=2	*MR*=5	*MR*=10	*MR*=15	*MR*=20
BCWD	89%	88%	100%	96%	95%	90%
Caesarian Section	33%	34%	88.9%	71%	66%	33%
Indian Liver Patients	88%	87%	90%	88%	77%	72%
Hepatitis	90%	88%	93.8%	94%	92%	89%
Average of predictive accuracy	75 %	74.25%	93.18 %	87.25%	82.50%	71 %

Figure 7 summarizes the average of predictive accuracy values with different number of *MR* on various datasets.

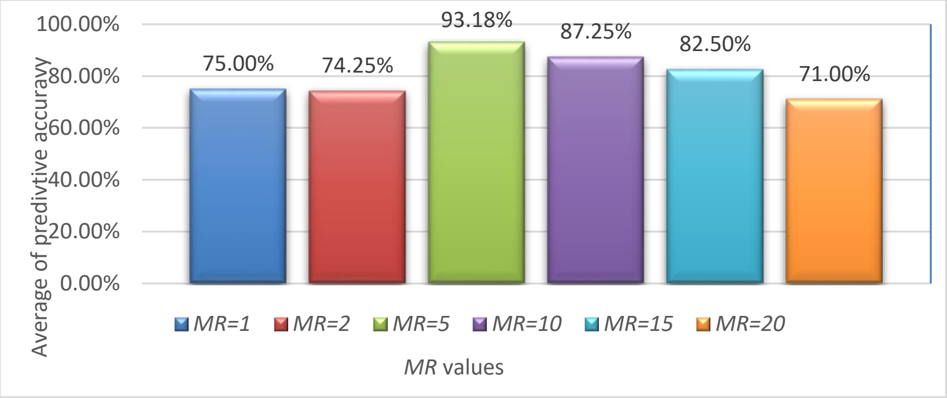

Fig.7 Average of predictive accuracy *vs* different *MR* values.

5.2 Predictive Accuracy

In the context of classification, it is important to evaluate the quality of the discovered rules with respect to rules predictive accuracy [10]. With a satisfactory classification ability, the classifier is used for classifying future / unseen data. This evaluation must be measured on a separate test set, containing data instances that not seen during training, i.e. the ratio of the number of instances correctly classified over total of instances in the test set. A comparison of predictive accuracies obtained on test sets using a SGA approach for discovering classification rules in data mining and the proposed DPMoGA approach is presented in Table 8.

Table 8 Summary of predictive accuracy comparison results.

Dataset	SGA approach	DPMoGA
BCWD	71.5%	100%
Caesarian Section	78%	88.9%
Indian Liver Patients	85%	90%
Hepatitis	84.7%	93.8%

Table 8 shows that, the DPMoGA approach achieves better average predictive accuracy than the SGA approach. The DPMoGA approach outperformed SGA approach due to that, the DPMoGA approach first optimizes the list of best rules and then inter-islands migrations help to optimize to the best set of rules. The SGA approach takes care of attribute interaction only which means that two or more attributes together can affect the class value. The DPMoGA approach adequate for attribute interaction as well as rule interaction i.e. how two or more number of rules work better or worse together. Also, in the DPMoGA approach the application of the selection method and genetic operators are independently performed in each of the subpopulations. Consequently, the number of genetic operators i.e. crossover performed between individuals predicting different goal attributes is reduced as a goal attribute value is fixed for each island. Note that, this is not the case with SGA approach, where genetic operators can apply among parents representing rules predicting different goal attributes. It is expected that, the DPMoGA approach has higher predictive accuracy with less prediction error during the test process on the test data, unlike SGA approach that does not give importance to the minority classes, in turn reduces the predictive accuracy.

6. CONCLUSION AND FUTURE WORK

In this work, the DPMoGA for mining interesting classification rules from medical datasets is proposed. The DPMoGA approach has a flexible chromosome representation, an effective multiobjective fitness function, appropriate genetic operators for suggested representation, a new island model based on distributed population with efficient migration operator. In this paper, the DPMoGA approach is developed as a comparative approach based on an island model by dividing the population into subpopulations and evolved simultaneously in separation to generate valuable knowledge from medical datasets that have both a good classification accuracy, a good degree of interestingness and comprehensibility as the discovered rules are represented in the **If-Then** form. The performance of the DPMoGA approach has been validated using some benchmark medical datasets and the results are evaluated by confusion matrix based on predictive accuracy, and prediction error, which showed the effectiveness of the proposed approach. In general, it can help to reduce the number of FP and FN decisions. Additionally, the performance of the DPMoGA approach depends on migration rate. The optimal migration rate was 5, where the approach gives the best results. The DPMoGA approach is able to discover interesting rules sets with significantly higher predictive accuracy compared to SGA approach. An important direction for future research is developing a method to automated discovery of interesting fuzzy classification rules from medical datasets using Parallel Multiobjective Evolutionary Algorithms (PMoEAs).

References

[1] R. Ghorbani, and R. Ghousi, “Predictive Data Mining Approaches in Medical Diagnosis: A Review of Some Diseases Prediction,” *International Journal of Data and Network Science*, vol. 3, no. 2, pp. 47-70, 2019.

[2] K. S. Rekha, and S. Sumethi, “A Survey of Evolutionary Algorithms and It's use in Data Mining Application,” *International Journal. of Pure and Applied Mathematics*, 119(12), pp. 13593-13600, 2018.

[3] A. Rani, “Analysis of Imbalanced Datasets using Classification Based Techniques,” *International Conference on Advanced Computing (ICAC-2016), Teerthanker Mahaveer University, Moradabad, India*, pp.45-48, 2016.

[4] U. Shafique, F. Majeed, H. Qaiser, et al., "Data Mining in Healthcare for Heart Diseases", *International Journal of Innovation and Applied Studies*, vol. 10, no. 4, pp. 1312-1322, 2015.

[5] A. Ghaheri, S. Shoar, M. Naderan, et al., "The Applications of Genetic Algorithms in Medicine", *Oman Medical Journal*, 30(6), pp. 406-416, 2015.

[6] V. Dhar, D. Chou, and F. Provost, "Discovering Interesting Patterns for Investment Decision Making with GLOWER: A Genetic Learner Overlaid with Entropy Reduction," *Data Mining and Knowledge Discovery*, vol. 4, no. 4, pp. 251-280, 2000.

[7] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease using Risk Factors," *IEEE Conference on Information & Communication Technologies*, *IEEE*, pp. 1227-1231, 2013.

[8] B. Y. Qu, Y. S. Zhu, Y. C. Jiao, et al. "A Survey on Multi-Objective Evolutionary Algorithms for the Solution of the Environmental/Economic Dispatch Problems," *Swarm and Evolutionary Computation*, vol. 38, pp. 1-11, 2018.

[9] M. Mitchell, An Introduction to Genetic Algorithms, *MIT Press*, 1998.

[10] A. A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, *Springer Science & Business Media*, 2002.

[11] C. Brester, and E. Semenkin, “Cooperative Multi-Objective Genetic Algorithm with Parallel Implementation,” *International Conference in Swarm Intelligence*, pp. 471-478, 2015.

[12] B. Alatas, and A. Arslan, “Mining of Interesting Prediction Rules with Uniform Two-Level Genetic Algorithm,” *International Journal of Computaer and Information*, 1(7), pp. 2322-2327, 2007.

[13] J. J. Christopher, K. H. Nehemiah, and K. Arputharaj, "Knowledge-Based Systems and Interestingness Measures: Analysis with Clinical Datasets," *Journal of Computing and Information Technology*, vol, 24, no. 1, pp. 65-78, 2016.

[14] P. Sharma, and Saroj, "Discovery of Classification Rules using Distributed Genetic Algorithm," *Procedia Computer Science*, *Elsevier,ScienceDirect*, vol. 46, pp. 276-284, 2015.

[15] Y. Gong, W. Chen, Z. Zhan, et al., "Distributed Evolutionary Algorithms and their Models: A Survey of The State-of-The-Art", *Applied Soft Computing*, vol. 34, pp. 286-300, 2015.

[16] F. Ferrucci, P. Salza, and F. Sarro, "Using Hadoop Mapreduce for Parallel Genetic Algorithms: A Comparison of the Global, Grid and Island Models," *Evolutionary Computation*, vol. 26, no. 4, pp. 535-567, 2018.

[17] R. R. Patil, "Heart Disease Prediction System Using Naive Bayes and Jelinek-Mercer Smoothing," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 5, 2014.

[18] T. Sharma, A. Sharma, and V. Mansotra, “Performance Analysis of Data Mining Classification Techniques on Public Health Care Data,” *International Journal of Innovative Research in Computer and Communication Engineering*, 4(6), pp. 1-5, 2016.

[19] B. M. Al-Maqaleh, and A. M. G. Abdullah, “Intelligent Predictive System Using Classification Techniques for Heart Disease Diagnosis,” *International Journal of Computer Science Engineering (IJCSE)*, vol. 6, no. 6, pp. 145-151, 2017.

[20] S. A. Lashari, R. Ibrahim, N. Senan, et al., "Application of Data Mining Techniques for Medical Data Classification: A Review," *MATEC Web of Conferences*, vol. 150, pp. 603-609, 2018.

[21] B. M. Bai, B. M. Nalini, and J. Majumdar, “Analysis and Detection of Diabetes using Data Mining Techniques: A Big Data Application in Health Care,” *Emerging Research in Computing, Information, Communication and Applications*, Springer, Singapore, pp. 443-455, 2019.

[22] B. M. Al-Maqaleh, and H. Shahbazkia, “A Genetic Algorithm for Discovering Classification Rule in Data Mining,” *International Journal of Computer Application (IJCA)*, 41(18), pp. 40-44, 2012.

[23] K. K. Gündoğan, B. Alataş, and A. Karci, “Mining Classification Rules by Using Genetic Algorithms with Non-Random Initial Population and Uniform Operator,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 12, no. 1, pp. 43-52, 2004.

[24] S. Iftikhar, K. Fatima, A. Rehman, et al. “An Evolution Based Hybrid Approach for Heart Diseases Classification and Associated Risk Factors Identification,” *Biomedical Research*, vol. 28, no. 8, pp. 3451-3455, 2017.

[25] A. Oztiken, L. Al-Ebbini, Z. Sevkli, et al., "Predicting Quality of Life for Lung Transplant Recipients: A Hybrid Genetic Algorithms-Based Methodology," *Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), IEEE*, pp. 1-6, 2017.

[26] R. Gayathri, "Genetic Algorithm Based Model for Early Detection of Cancer," *Science and Technology*, vol. 3, no. 8, pp. 28-31, 2017.

[27] M. Sari, and C. Tuna, “Prediction of Pathological Subjects using Genetic Algorithms,” *Computational and Mathematical Methods in Medicine*, 2018.

[28] H. A. Hussein, I. Demiroglu, and R. L. Johnston, “Application of a Parallel Genetic Algorithm to the Global Optimization of Medium-Sized Au–Pd Sub-Nanometer Clusters,” *The European Physical Journal B*, vol. 91, no. 2, pp. 34-45, 2018.

[29] J. M. Dadmehr, “The Multi-Objective Genetic Algorithm Based Techniques for Intrusion Detection,” *International Journal of Compute Science and Network Security (IJCSNS)*, vol. 16, no. 3, pp. 39-45, 2016.

[30] X. Sun, L-F. Lain, P. Chou, et al. "On GPU Implementation of the Island Model Genetic Algorithm for Solving the Unequal Area Facility Layout Problem," *Applied Sciences*, 8(9), pp. 1604-1616, 2018.

[31] K. I. Abuzanoureh, "Parallel and Distributed Genetic Algorithm with Multiobjective Algorithm to Improve and Develop of Evolutionary Algorithm," *International Journal of Advanced Computer Science and Application (IJACSA)*, vol. 7, no. 5, pp. 154-160, 2017.

[32] H. Kalia, S. Dehuri, A. Ghosh, et al. "On The Mining of Fuzzy Association Rule using Multi-Objective Genetic Algorithms," *International Journal of Data Mining, Modelling and Management*, vol. 8, no. 1, pp. 1-31, 2016.

[33] C. L. Blake, and M. J. Merz, UCI Repository of Machine Learning Database [http://www.ics.uci.edu/mlearn/MLRepository.html], *Irvine, CA: University of California*, *Department of Information and Computer Science.*

[34] Knowledge Extraction based on Evolutionary Learning: http://www.keel.es.

[35] WEKA:http://www.cs.waikato.qc.nz/m1/weka/index.html.