Dynamic genetic algorithm-based feature selection and incomplete value imputation for microarray classification

R. Devi Priya[1,*] and R. Sivaraj[2]

[1]Department of Information Technology, Kongu Engineering College, Erode 638 052, India
[2]Department of Computer Science and Engineering,
Velalar College of Engineering and Technology, Erode 638 002, India

Large microarray datasets usually contain many features with missing values. Inferences made from such incomplete datasets may be biased. To address this issue, we propose a novel preprocessing method called dynamic genetic algorithm-based feature selection with missing value imputation. The significant features are first identified using dynamic genetic algorithm-based feature selection and then the missing values are imputed using dynamic Bayesian genetic algorithm. The resulting complete microarray datasets with reduced features are used for classification, which results in better accuracy than the existing methods in eight microarray datasets.

Keywords: Microarray dataset, feature selection, missing values, genetic algorithm.

THE microarray technology has emerged as an important one in the medical field, which helps clinicians in analysing a large number of genes together to draw inferences regarding the functionalities of cells. The gene information is used to predict the type and criticality of the diseases. However, the microarray datasets are usually large and contain many features (attributes) with missing values. In order to minimize the time taken for processing unnecessary data, feature selection methods are commonly used to identify the required features. Information is mostly obtained from clinical experiments or surveys where some data may remain missing due to technical or human difficulties in assessing the study parameters. Most of the analysts delete records with missing values and perform feature selection on the remaining dataset[1]. This will in turn result in loss of valuable information and its statistical power. If accurate inferences should be made, the missing values should be estimated and included for analysis. This need motivates researchers to introduce different methods to treat missing values in microarray datasets.

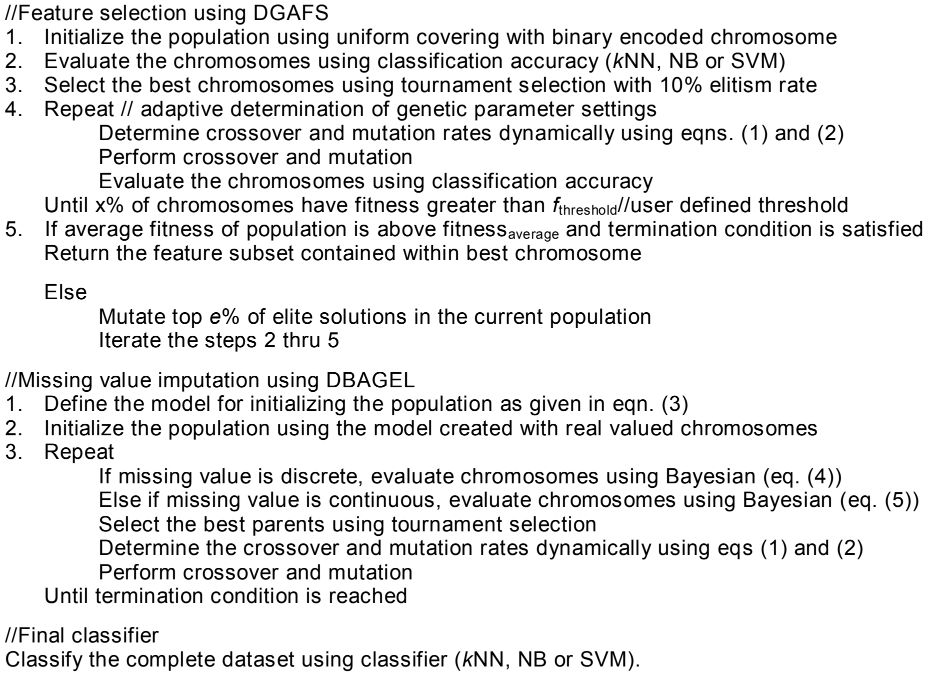
In order to reduce the computational complexity in analysing irrelevant features, here we use dynamic genetic algorithm-based feature selection (DGAFS) method to select the significant features. Then, a novel methodology called dynamic Bayesian genetic algorithm (DBAGEL), which combines genetic algorithm (GA) and Bayesian principles is introduced to estimate non-ignorable missing values which often occur in real datasets. DBAGEL is designed by enhancing the principles of Bayesian Genetic Algorithm (BAGEL) proposed by Devi Priya and Kuppuswami[2]. The imputation accuracy of DBAGEL at missing rates ranging from 5% to 40% is better when compared with that of existing techniques and the classification accuracy of the reduced complete feature subset is encouraging in all the datasets.

Microarray data classification is usually done to classify genes in the datasets, which are then used to make vital clinical decisions. Some of the recent studies on microarray classification include fuzzy rough set approach[3], generalized radial basis function neural networks[4], genetic swarm algorithm[5], ensemble classifiers[6], particle swam optimization (PSO)-based decision tree classifier[7], etc.

Gene (feature) selection should be done prior to gene classification or any other analysis[8]. Methods used for feature selection are grouped into (i) filter, (ii) wrapper and (iii) embedded methods. Filter methods depend on intrinsic characteristics of genes to discriminate them using statistical techniques. They are fast, classifier-independent and suitable for large datasets[9]. In wrapper methods, the gene subsets are initialized using heuristic information and then evaluated by the corresponding classifier[10]. In spite of the increased computational complexity, wrapper methods are preferred over filter methods, since they utilize heuristic information and appropriate classifier to select optimal set of features with training and testing sets. In the filter methods, the feature subsets selected are not aligned with the predictive model and produce only general results with less performance than the wrapper methods. The embedded methods utilize classifiers to initialize and select features iteratively and provide a balanced trade-off between the filter and wrapper methods[11].

At times, specific values of some attributes cannot be measured or recorded and the scenario is called not missing at random (NMAR). The missing values themselves cause the missingness, and other attributes in the dataset do not have any influence on it. These significant missing values cannot be ignored and are defined as non-ignorable missing values. Right assumptions are required and models have to be developed based on prior knowledge to depict corresponding missingness in the dataset. The NMAR values cannot be imputed by considering the missing attribute alone; rather related attributes should also be included in the model. Researchers have introduced different models for imputing NMAR data[12]. Selection and pattern mixture models are commonly used[13]. Calibration weighting[14], pseudo empirical likelihood[15] and GAs[16] are some other methods which have been proposed for treating NMAR values. Mean or mode value substitution and complete case analysis are simple methods, but less

*For correspondence. (e-mail: scrpriya@gmail.com)



```
//Feature selection using DGAFS
1. Initialize the population using uniform covering with binary encoded chromosome
2. Evaluate the chromosomes using classification accuracy (kNN, NB or SVM)
3. Select the best chromosomes using tournament selection with 10% elitism rate
4. Repeat // adaptive determination of genetic parameter settings
          Determine crossover and mutation rates dynamically using eqns. (1) and (2)
          Perform crossover and mutation
          Evaluate the chromosomes using classification accuracy
   Until x% of chromosomes have fitness greater than $f_{\text{threshold}}$//user defined threshold
5. If average fitness of population is above $\text{fitness}_{\text{average}}$ and termination condition is satisfied
   Return the feature subset contained within best chromosome

   Else
          Mutate top e% of elite solutions in the current population
          Iterate the steps 2 thru 5

//Missing value imputation using DBAGEL
1. Define the model for initializing the population as given in eqn. (3)
2. Initialize the population using the model created with real valued chromosomes
3. Repeat
          If missing value is discrete, evaluate chromosomes using Bayesian (eq. (4))
          Else if missing value is continuous, evaluate chromosomes using Bayesian (eq. (5))
          Select the best parents using tournament selection
          Determine the crossover and mutation rates dynamically using eqs (1) and (2)
          Perform crossover and mutation
   Until termination condition is reached

//Final classifier
Classify the complete dataset using classifier (kNN, NB or SVM).
```

Figure 1. Basic outline of dynamic genetic algorithm based feature selection – missing value imputation.

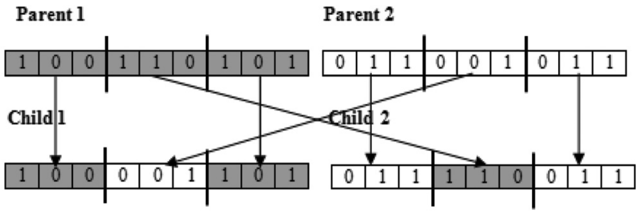


Figure 2. Two-point crossover in dynamic genetic algorithm based feature selection (DGAFS).

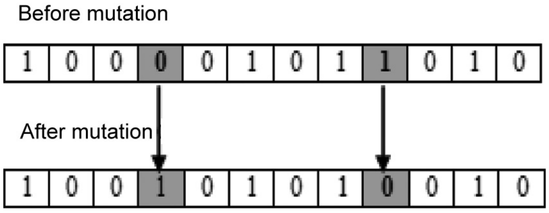


Figure 3. Mutation in DGAFS.

efficient and rarely preferred[17]. Bayesian methods like approximate Bayesian bootstrap (ABB) integrated with multiple imputation[18] and non-parametric Bayesian-based multiple imputation (Bay-MI)[19] are preferred since the assumptions and constraints required for imputation can be easily incorporated in the Bayesian rule. With the knowledge that performance of Bayesian methods can be enhanced when hybridized with optimization methods, DBAGEL is proposed by combining Bayesian methods and GA, where the parameter values of GA are dynamically adapted.

DGAFS-MI algorithm consists of three main steps which are discussed in detail below. Figure 1 shows the basic outline of the algorithm.

In DGAFS, binary encoding is used where 1 and 0 represent presence and absence of genes in the corresponding positions of chromosomes. The required chromosomes are initialized through uniform covering initialization, where the density of 1 is randomly selected in the interval [0, 1] and then 1 is placed in the chosen positions with the selected probability.

The initialized chromosomes are then evaluated using fitness function. Classifiers like k nearest neighbour (kNN), naïve Bayes (NB) or support vector machine (SVM) can be implemented, and the obtained classification accuracy can be used as fitness value of the corresponding chromosome. Then, chromosomes with good fitness values are selected as parents for crossover operation. Parents are selected through appropriate mechanisms like roulette wheel, tournament or rank selection. In two-point crossover, two random crossover points from the selected parents are chosen and genes from two parents are exchanged to form new children (Figure 2). The children will have new recombined characteristics which increase the explorative capability of the algorithm. The resulting chromosomes are then mutated to enhance their exploitative capabilities, where the genes are flipped (for example 0 to 1 and 1 to 0 in Figure 3).

The significant contribution of DGAFS is dynamic determination of crossover and mutation probabilities. They indicate the number of chromosomes to be crossed over and the number of genes to be mutated respectively.

The crossover probability P_c is defined using the formula:

$$P_c = \begin{cases} x\left[1 - \dfrac{\text{fitness}_{\text{crs}} - \text{fitness}_{\text{avg}}}{\text{fitness}_{\max} - \text{fitness}_{\text{avg}}}\right], \text{fitness}_{\text{cros}} > \text{fitness}_{\text{avg}}, \\ x, \text{otherwise}, \end{cases} \tag{1}$$

where $\text{fitness}_{\text{crs}}$ indicates the largest fitness value of two chromosomes chosen for crossover, while $\text{fitness}_{\max}$ and $\text{fitness}_{\text{avg}}$ refers to maximum and average fitness of the population. The mutation probability P_m is given by

$$P_m = \begin{cases} y\left[1 - \dfrac{\text{fitness}_{\text{mut}} - \text{fitness}_{\text{med}}}{\text{fitness}_{\max} - \text{fitness}_{\text{med}}}\right], \\ y, \text{otherwise}, \end{cases}$$

$$\text{if fitness}_{\text{mut}} > \text{fitness}_{\text{med}}, \tag{2}$$

where $\text{fitness}_{\text{mut}}$ refers to the fitness value of the mutated chromosome and $\text{fitness}_{\text{med}}$ is the median fitness value in the population. The constants x and y are set by the user in the range [0, 1]. The crossover and mutation operations are repeated until n% of chromosomes have fitness value greater than $\text{fitness}_{\text{threshold}}$ set by the user. This iterative determination of crossover and mutation rates and operations controlled by threshold value help in achieving better exploration and exploitation and prevent the algorithm from being tapped in local optimum.

The algorithm is repeated until the termination point is reached or the population gets the required chromosomes with predefined average fitness value, $\text{fitness}_{\text{average}}$. If the defined threshold is not reached, top e% of elite solutions in the current population are mutated and introduced into the new population for the next generation. The genes (features) present in the final optimal solution are included in the feature subset.

In the feature subset selected by DGAFS, the missing values are imputed by DBAGEL. GA and Bayesian methods are integrated with dynamic determination of parameters in DBAGEL in order to impute non-ignorable, discrete and continuous NMAR values. In both model creation and fitness estimation of GA, Bayesian principles are used.

In a dataset with N instances, let x and y represent the target and covariate attributes respectively. The missing variable M_j is set to 1 if jth record is complete and 0, if its values are missing. The complete instances are selected in the subset P in which the given factorization as shown in eq. (3) is applied on all its samples.

$$f_{\text{p}}(x_j|y_j, M_j) = \frac{P(M_j = 1|x_j, y_j) f_P(x_i|y_i)}{P(M_i = 1|y_i)}, \text{ for } i = 1, 2, \ldots, N, \tag{3}$$

where $f_{\text{P}}(x_j \mid y_j)$ indicates conditional probability function (pdf) of target variable x_j. Appropriate pdf is chosen depending upon data distribution in the dataset. Univariate, bivariate or multivariate distributions can be implemented based upon the number of variables involved. Normal distribution is chosen for the example, since it is commonly assumed in Bayesian models. Even if the real distribution of data is not known, analysts usually prefer normal distribution. If the dataset has some other type of distribution, it can also be modelled in the same way. This Bayesian model can be implemented for handling both homogeneous and heterogeneous missing values. The samples are sorted in decreasing order of their pdf values and the top ones are selected for the initial population. This model thus successfully inserts good chromosomes into the population with integer encoding, where the gene values are depicted as such in the chromosome.

Fitness of chromosomes in the population is then evaluated using Bayes' rule. Equation (4) is used to calculate the Bayesian probability for discrete attribute, where X_{mis}, Y and α indicate the missing target attribute, covariates and model parameters.

$$P(X_{\text{mis}}|Y, \alpha) = \frac{P(Y, \alpha| X_{\text{miss}}) \cdot P(X_{\text{mis}})}{P(Y, \alpha)}. \tag{4}$$

Probability is replaced by pdf, if the missing attribute is continuous as given in eq. (5). $f(X_{\text{miss}} \mid Y, \alpha)$ represents the joint probability of the missing attribute with covariate and model parameters. The function $f(.)$ indicates the pdf with normal distribution. The implementation of pdf takes more time than probability estimation since it involves integral calculation of the variables involved.

$$f(X_{\text{miss}}|Y, \theta) = \frac{f(Y, \alpha|X_{\text{miss}}) \cdot f(X_{\text{miss}})}{f(Y, \theta)}. \tag{5}$$

Based on fitness values of the chromosomes, best parents are selected for reproduction operation. One-point, two-point or uniform crossover can be used. Since the selected feature subset contains limited number of significant features, one-point crossover will suffice where one random crossover point is chosen and the genes are

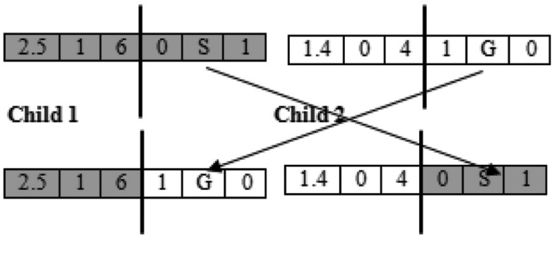


Figure 4. One-point crossover in DBAGEL.

Table 1. Performance results of dynamic genetic algorithm based feature selection (DGAFS) with different classifiers

Dataset	DGAFS- naïve Bayes (NB)	DGAFS- *k*-nearest neighbour (*k*NN)	DGAFS- support vector machine (SVM)
Colon	4 (93.7)	6 (93.5)	4 (94.2)
DLBCL	4 (94.8)	6 (94.3)	2 (95.6)
CNS	3 (74.4)	5 (75.3)	3 (78.6)
ALL/AML	4 (94.5)	4 (95.2)	4 (95.8)
Lung	3 (99.1)	2 (99.1)	2 (99.5)
Prostate	19 (72.3)	20 (73.3)	17 (75.1)
Ovarian	3 (99.1)	5 (99.1)	3 (99.3)
Breast	6 (82.9)	15 (83.6)	6 (84.2)

Note: Each entry represents the number of selected features (classification accuracy %).

Table 2. RMSE of BAGEL and other methods at different missing rates

		Missing rate			
Dataset	Algorithm	10%	20%	30%	40%
Colon	Mean	3.64	3.83	4.31	6.70
	ABB	3.17	3.96	4.25	6.32
	Bay-MI	2.34	3.45	3.83	5.27
	DBAGEL	1.47	2.98	2.98	3.94
DLBCL	Mean	5.75	6.45	7.14	7.98
	ABB	4.23	3.38	3.02	7.48
	Bay-MI	3.48	2.48	2.84	5.61
	DBAGEL	1.47	2.05	2.79	4.35
CNS	Mean	4.20	5.97	6.32	7.21
	ABB	4.13	5.55	5.45	6.84
	Bay-MI	3.75	4.15	5.02	6.23
	DBAGEL	2.46	3.30	4.87	5.47
ALL/AML	Mean	4.81	5.46	7.59	8.63
	ABB	3.26	3.84	4.57	5.31
	Bay-MI	2.36	2.94	3.94	5.81
	DBAGEL	1.45	2.63	3.27	4.18
Lung	Mean	2.13	3.46	4.97	6.18
	ABB	2.91	3.78	4.79	6.75
	Bay-MI	2.56	3.14	4.67	5.78
	DBAGEL	1.23	2.28	3.56	4.13
Prostate	Mean	2.05	2.90	3.14	5.02
	ABB	4.17	5.23	5.61	7.17
	Bay-MI	3.14	4.26	5.15	6.38
	DBAGEL	1.97	2.34	3.25	4.76
Ovarian	Mean	4.56	4.58	5.19	6.23
	ABB	3.24	3.45	3.97	4.28
	Bay-MI	2.97	3.12	3.96	4.27
	DBAGEL	2.31	2.84	3.25	4.89
Breast	Mean	5.33	5.85	6.34	7.28
	ABB	4.36	4.91	5.32	6.18
	Bay-MI	4.18	4.76	5.21	6.04
	DBAGEL	3.56	3.84	4.29	4.37

ABB, Approximate Bayesian bootstrap; Bay-MI, Bayesian based multiple imputation; DBAGEL, Dynamic Bayesian genetic algorithm.

exchanged among parents (Figure 4). Mutation is performed on the new offspring, where genes in different positions are swapped (Figure 5).

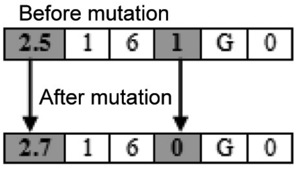


Figure 5. Mutation in dynamic Bayesian genetic algorithm.

If the stopping condition is reached, chromosome with the best fitness value is returned as the optimal solution which contains the value to be replaced in the missing place. Otherwise, the steps from fitness estimation to genetic operations are repeated. The algorithm is run until the chromosomes remain consistent in *n* successive generations.

Finally, classifiers like NB, *k*NN and SVM are implemented with ten-fold cross validation to test the classification performance in the processed datasets.

DGAFS-MI is implemented on eight microarray datasets taken from public repositories and its classification accuracies after feature selection and missing value imputation are estimated.

In the original datasets, the relevant features are first selected using DGAFS with the following genetic parameters.

Population size, 40; Encoding, Real encoding; Selection, Tournament selection; Crossover, Two-point crossover; Mutation, Flip mutation; Elitism, 10%; Crossover rate (P_c), Determined dynamically using eq. (2); Mutation rate (P_m), Determined dynamically using eq. (3).

The fitness of chromosomes is given by the classifier accuracy obtained from classifiers like NB, *k*NN and SVM. Ten-fold cross validation is done on the results obtained by implementing classification mechanisms in the datasets. The optimal feature subset obtained from DGAFS with the best classification accuracy is returned as the solution in every run. The algorithm is executed for 50 runs and the best result obtained among all the runs is given in Table 1, showing the number of selected features and their corresponding classification accuracies.

Table 3. Classification accuracy% of DGAFS-MI with different classifiers

Dataset	DGAFS-MI-NB	DGAFS-MI-*k*NN	DGAFS-MI-SVM
Colon	98 ± 1.02	98 ± 1.51	99 ± 0.08
DLBCL	97 ± 1.64	97 ± 1.35	98 ± 0.85
CNS	93 ± 1.17	92 ± 0.14	94 ± 0.63
ALL/AML	98 ± 1.64	97 ± 1.28	98 ± 0.45
Lung	99 ± 0.20	99 ± 0.35	99 ± 0.50
Prostate	89 ± 0.61	90 ± 1.58	92 ± 1.04
Ovarian	99 ± 0.40	99 ± 0.50	99 ± 0.50
Breast	90 ± 1.63	90 ± 1.82	91 ± 1.34

NB, Naïve Bayes; kNN, *k*-nearest classifier; SVM, support vector machine.

Both lung and ovarian cancer datasets, DGAFS results in more than 99% classification accuracy. In ovarian dataset, NB and SVM classifiers achieve best classification accuracy with less number of features than *k*NN. In lung dataset, *k*NN and SVM are significant than NB. In general, it is observed that NB selects few features than *k*NN. But the classification accuracy of *k*NN is better than that of NB. A compromise has to be made on these two factors when the two classifiers are considered. In all the datasets, SVM achieves better classification accuracy than the other two classifiers because of the kernel functions used in it. But its computational complexity is higher than the others due to more calculations involved. The final performance depends upon the initial set of features selected. Hence utmost care must be taken to induce right features into the subset. Since SVM is better in terms of both features and classification accuracy, it is used in further experiments.

The reduced feature set contains only the relevant features which are essential for efficient classification. From the complete dataset, non-ignorable missing values are simulated at different missing rates of 5%, 10%, 20%, 30% and 40%. DBAGEL is then used to impute them and root mean square error (RMSE) is evaluated. In DBAGEL, population size is an important parameter affecting the results and hence needs careful attention. If the number of chromosomes is too few, efficient chromosomes may not get a chance in the population. If it is high, convergence of the algorithm is delayed. Twenty-five trials are conducted with different population sizes of 30, 40, 50 and 60. It is found that 40 chromosomes produce better results and hence the population size is fixed to 40. The genetic parameters of DBAGEL are the same as those of DGAFS, except that one-point crossover is used in DBAGEL instead of two-point crossover in DGAFS.

In Table 2, RMSE of DBAGEL, mean imputation, approximate Bayesian bootsteap (ABB) used in Siddique and Belin[18] and Bay-MI used in Si[19] are reported. Mean imputation is easy, but it substitutes the mean value in all missing holes and does not try to impute the exact values. It is less preferred by researchers since it produces only biased results than other imputation methods. ABB and Bay-MI are both Bayesian-based methods. In ABB, bootstrap technique is integrated with Bayesian principle. It is observed that it produces better results than mean imputation, but underperforms when compared with Bay-MI and DBAGEL. Since Bay-MI hybridizes the Bayesian method with multiple imputation which is already a standard technique for missing value imputation, it is able to produce good results in all the eight datasets at different missing rates. DBAGEL outperforms the other three methods due to its dynamic adaptation of genetic parameters as in DGAFS. Even at the missing rate of 40%, DBAGEL shows RMSE only within 5% in all the datasets.

After the missing values are estimated using DBAGEL, the dataset can be used for classification. Classifiers like NB, *k*NN and SVM are implemented in the processed microarray datasets and their accuracies in classifying the datasets are estimated again and compared. Significant difference is observed between the classification accuracies of Table 1 (without missing value imputation) and Table 3 (after missing value imputation). Average of 7% performance improvement is seen in the results, which is encouraging. The main factor behind the improvement in Table 3 is because the missing values are efficiently imputed and the dataset is made complete retaining useful information. When the missing values are ignored, useful information required for strategic clinical decisions are left out. For example, if there are 200 instances and 50 features in the dataset with 2% missing values in different instances, 50% of the instances will be incomplete and analysis made from this will not be accurate. After implementing DGAFS-MI on large lung and ovarian datasets with 12,533 and 15,154 genes respectively, classification accuracy observed is close to 100% with all three classifiers. SVM is better in terms of classification accuracy. But its computational complexity is higher than that of NB and *k*NN. If accuracy is required, SVM can be used and if the implementation needs to be simple, NB or *k*NN can be used. But again some heuristic procedure is required to choose the value of *k* in *k*NN.

Classification in microarray datasets becomes difficult due to the presence of many irrelevant attributes and missing values. Here we propose DGAFS-MI by selecting significant features and imputing missing values. The threshold maintained for dynamically updating population size, crossover and mutation probabilities restricts the unwanted attributes and retains only optimal features in the population. BAGEL supports this process by efficient imputation of missing values. The proposed algorithm is implemented on real datasets. The results show that the classification accuracy obtained on the processed datasets is better than other existing algorithms. DGAFS-MI can thus reduce the burden of clinicians and help them in efficient analysis of microarray datasets.

1. Lee, C.-P. and Leu, Y., A novel hybrid feature selection method for microarray data analysis. *Appl. Soft Comput.*, 2011, **11**, 208–213.
2. Devi Priya, R. and Kuppuswami, S., A genetic algorithm-based approach for imputing missing discrete values in databases. *WSEAS Trans. Inf. Sci. Appl.*, 2012, **9**, 169–178.
3. Pramod Kumar, P., Prahlad, V. and Poh, A. L., Fuzzy-rough discriminative feature selection and classification algorithm with application to microarray and image datasets. *Appl. Soft Comput.*, 2011, **11**, 3429–3440.
4. Fernandez-Navarro, F., Hervás-Martínez, C., Ruiz, R. and Riquelme, J. C., Evolutionary generalized radial basis function neural networks for improving prediction accuracy in gene classification using feature selection. *Appl. Soft Comput.*, 2012, **12**, 1787–1800.
5. Ganesh Kumar, P., Aruldoss Albert Victoire, T., Renukadevi, P. and Devaraj, D., Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Exp. Syst. Appl.*, 2012, **39**, 1811–1821.
6. Reboiro-Jato, M., Díaz, F., Glez-Pena, D. and Fdez-Riverola, F., A novel ensemble of classifiers that use biological relevant gene sets for microarray classification. *Appl. Soft Comput.*, 2014, **17**, 117–126.
7. Chen, K.-H., Wang, K.-J., Wang, K.-M. and Angelia, M.-A., Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Appl. Soft Comput.*, 2014, **24**, 773–780.
8. Bolon-Canedo, V., Sánchez-Marono, N. and Alonso-Betanzos, A., Distributed feature selection: an application to microarray data classification. *Appl. Soft Comput.*, 2015, **30**, 136–150.
9. Su, Y., Murali, T. M., Pavlovic, V., Schaffer, M. and Kasif, S., RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, 2003, **19**, 1578–1579.
10. Li, L. *et al.*, A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, 2005, **85**, 16–23.
11. Zibakhsh, A. and Abadeh, M. S., Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function. *Eng. Appl. Artif. Intell*,. 2013, **26**, 1274–1281.
12. Xie, H., Adjusting for nonignorable missingness when estimating generalized additive models. *Biometrika. J.*, 2010, **52**, 186–200.
13. Muthen, B., Asparouhov, T., Hunter, A. and Leuchter, A., Growth modeling with non-ignorable dropout: alternative analyses of the STAR*D antidepressant trial. *Psychol. Meth.*, 2011, **16**, 16–33.
14. Kim, J. K., Calibration estimation using empirical likelihood in survey sampling. *Stat. Sin.,* 2009, **19**, 145–157.
15. Fang, F., Hong, Q. and Shao, J., Empirical likelihood estimation for samples with non-ignorable nonresponse. *Stat. Sin.,* 2010, **20**, 263–280.
16. Devi Priya, R. and Kuppuswami, S., Drawing inferences from clinical studies with missing values using genetic algorithm. *Int. J. Bioinformat. Res. Appl.*, 2014, **10**, 613–627.
17. Kaciroti, N. and Raghunathan, T. E., Bayesian sensitivity analysis of incomplete data using pattern-mixture and selection models through equivalent parameterization. *Ann. Arbor.*, 2009, **1001**, 48109.
18. Siddique, J. and Belin, T. R., Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Comput. Stat. Data Anal.*, 2008, **53**, 405–415.
19. Si, Y., Non-parametric Bayesian methods for multiple imputation of large scale incomplete categorical data in panel studies. Ph D dissertation, Duke University, USA, 2012.

ACKNOWLEDGEMENT. We thank the reviewers for their useful comments that helped improve the manuscript.

Received 19 November 2015; revised accepted 12 August 2016

doi: 10.18520/cs/v112/i01/126-131

Exposure to particulate matter in different regions along a road network, Jharia coalfield, Dhanbad, Jharkhand, India

Shiv Kumar Yadav and Manish Kumar Jain*

Department of Environmental Science and Engineering,
Indian School of Mines, Dhanbad 826 004, India

Occupational particulate matter (PM) concentrations were measured during November 2014 along a road network in the mining and non-mining areas at Jharia coalfield, Dhanbad, Jharkhand, India. The monitoring was conducted for a week in the peak time using a portable GRIMM (model 1.109) aerosol spectrometer. Measured PM was designated as inhalable, thoracic and alveolic particles for aerodynamic diameter 10–34, 4–10 and less than 4 μm respectively. The main sources of PM along the roadside in the study area were mining operations as well as heavy traffic and resuspension of road dust. Concentration of inhalable particles was maximum at Bankmore (BMO), whereas concentration of thoracic and alveolic particles was maximum at Katrasmore (KMO) in the mining area. Concentration of all three types of particles was minimum at the Indian School of Mines in the non-mining area. The distribution curves of inhalable particles were positively skewed and platykurtic in nature, whereas for thoracic and alveolic particles these curves were positively skewed at all locations, except BMO and also platykurtic in nature, except Godhar (GDR). Contribution of alveoli particle sizes for 0.375

*For correspondence. (e-mail: manishjkm@gmail.com)