Australian Journal of Basic and Applied Sciences, 9(27) August 2015, Pages: 583-591

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

An Efficient Model by Applying Genetic Algorithms for Outlier Detection in Classifying Medical Datasets

[1]T Santhanam and [2]M.S. Padmavathi

[1] *Associate Professor and Head, PG and Research Department of Computer Science and Applications, D. G. Vaishnav College,Chennai - 600106, India.*
[2] *Ph.D Scholor, PG and Research Department of Computer Science and Applications, D. G.Vaishnav College, Chennai - 600106, India.*

ARTICLE INFO

Article history:
Received 23 June 2015
Accepted 25 August 2015
Available online 2 September 2015

Keywords:
BayesNet, Genetic Algorithm, Imputation, Inter Quartile Range, K-Means, Outliers, Random Forest, SVM

ABSTRACT

Background: Genetic Algorithms (GA) represent an intelligent exploitation of a random search used to solve optimization problems. Many researchers have used genetic algorithms for feature selection to improve the accuracy of their model. **Objective**: In this research work, GA is used for removing outliers rather than using it for feature selection. The proposed model involves three methods of imputation (mean, median and class mean) and a comparative study of outlier detection techniques using Inter Quartile Range (IQR), K-Means and GA is done. The reduced dataset is then classified using Support Vector Machine (SVM), BayesNet and Random Forest classifiers. **Results**: The experimental result proves that the average outlier detection percentage of GA is 5.1% less and the average classification accuracy is 2.4% more than K-Means. From the tested datasets, it is also proved that the average classification accuracy of SVM (99.57%) is higher than the BayesNet and Random Forest. **Conclusion:** Among the proposed methods at different stages imputation by median followed by GA for outlier removal and SVM for classification is considered as a best framework in classifying medical datasets.

To Cite This Article: T. Santhanam and M.S. Padmavathi., An Efficient Model by Applying Genetic Algorithms for Outlier Detection in Classifying Medical Datasets ***Aust. J. Basic & Appl. Sci.,*** *9(27): 583-591, 2015*

INTRODUCTION

Data cleaning is important to maintain qualitative data. Inconsistent data may lead to wrong decisions and error, therefore data cleaning is a prerequisite in almost all data analysis situations. Data cleaning is a process of determining inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detected errors and omissions (Kalaivany Natarajan *et al*., 2010). This research article involves two steps in data cleaning: 1. Replace missing values 2. Remove outliers. Missing values are tuples having no recorded value for several attributes, and the missing values can be filled by various methods like ignoring the tuple, fill in the missing value manually, use a global constant, use the attribute mean, use the attribute mean for all samples belonging to the same class and use of the most probable value (Jiawei Han *et al*., 2006). This paper focuses on three different methods of imputation (imputation by mean, median, mean belonging to same class) and their effect on outlier detection. Outlier is defined as an observation in a dataset which appears to be inconsistent with the remainder of that set of data (Johnson, 1992). In medical domain, many researchers (Asha *et al*., 2011) have used K-Means clustering to remove the incorrectly classified instances as outliers which has a high data reduction percentage. Outlier removal using IQR (Santhanam *et al*., 2014) has a less data reduction percentage compared with K-Means but the classification accuracy goes down, hence a method has been proposed to overcome the above disadvantage. The main objective is to replace K-Means clustering with GA thereby outlier removal percentage is reduced and classification accuracy percentage is improved.

Classification in data mining is used to understand the whole space available. It is used to predict the members of a group for the given input data. The use of three different classifiers: 1. SVM (mathematical) 2. BayesNet (statistical) 3. Random Forest (tree) is carried out to analyze a better classifier suitable for the proposed work. The new framework proposed is tested for medical datasets with null and non-null values by comparing data reduction and classification accuracy percentage. The rest of the paper is organized as follows: Section 2 shows the literature review followed by data source description in section 3. The preliminaries used are

Corresponding Author: M.S. Padmavathi, Research Scholor, PG and Research department of Computer Science and Applications, DG Vaishnav College, Arumbakkam, Chennai – 106.
E-mail: padmanivas_2002@yahoo.co.in.

given in section 4. Section 5 reports the experimental results followed by conclusion in the last section.

Related Work:

Acuna *et al*. (2004) carried out experiments on four different methods of dealing with missing values: the Case deletion method, mean imputation, median imputation and KNN imputation procedure and their effect on incorrectly classified error rate. Maytal *et al*. (2007) compared the effect of different treatments of missing values on model construction and analyzed treatments for the common case of missing values at prediction time. Vishnu Vishnu Raja *et al*. (2012) detected outliers using genetic algorithm which was accurate in identifying the outliers of the datasets. Chin-Yuan Fan *et al*. (2011) classified liver disorder and breast cancer datasets by proposing a hybrid model to integrate case based data cluster and fuzzy decision tree. Pachgade *et al*. (2012) removed outliers on datasets based on cluster and distance based approach. Asha Gowda Karegowda *et al*. (2012) classified Tuberculosis data with cascading clustering (K-Means) and classification (KNN). Nihat *et al*. (2014) proposed new data preparation method using modified K-Means for removing outliers and SVM for classifying the reduced datasets. Patil *et al*. (2010) proposed a hybrid K-Means and compared the performance of naïve bayes and SVM classification, in which SVM achieved high accuracy percentage. Bruno Fernandes Chimieski *et al*. (2013) compared data mining algorithms related to Classification and Association tasks over medical datasets. : For diagnostics prediction about breast cancer and dermatology issues, BayesNet was proved to be the best classification algorithm. Cuong Nguyen *et al*. (2013) classified breast cancer datasets using random forest classifier and feature selection by weighting. The results proved that 100% accuracy was obtained in the best case.

MATERIALS AND METHODS

Data Source:

Experiments are conducted using the medical datasets from UCI machine repository. Datasets were chosen such that two datasets (Diabetes and WBC) contain missing values and two datasets (WDBC and Heart (stalog)) with no missing values. These datasets were used to check how GA performs outlier detection for real values and imputed values (mean, median and class mean). Pima Diabetes (Radha *et al*., 2014) used to diagnose the presence of diabetes in pregnant women. It contains numerical attributes except for the class variable (0 for sick and 1 for healthy). The WBC (Bichen Zheng *et al*., 2014) dataset was created to prove the usefulness of automation of fine needle aspiration cytological diagnosis. It contains ordinal attributes which takes value from 1-10 except for class (2 for benign, 4 for malignant) and id number. WDBC (Olafit *et al*., 2014) contains features that are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image. It contains numerical attributes except for the class variable (0, 1 for absence and presence of disease). Heart (stalog) dataset (Negar Ziasabounchi *et al*., 2014) contains a combination of real, binary, ordinal and nominal values used to diagnose the presence or absence (1 or 0) of heart disease. The characteristics of the datasets are given in **Table 1**.

Table 1: Dataset Characteristics.

Datasets	No. of Instances	Input variables	Output variable	Missing Values
Diabetes Diagnosis	768	8	1	376
WBC	699	9	1	16
WDBC	569	30	1	NIL
Heart (Stalog)	270	13	1	NIL

Inter Quartile Range (IQR):

Outliers can be detected using IQR and represented using box-and–whisker plot. Any set of data can be described by its five number summaries namely the minimum, first quartile Q_1, median- Q2, third quartile Q_3 and the maximum (Courtnery Taylor). IQR is the difference between the first and third quartiles.

$$IQR = Q_3 - Q_1 \quad (1)$$

The IQR shows how the data is spread about the median. Multiplying the IQR by 1.5 will give us a way to determine whether a certain value is an outlier. Subtracting 1.5 x IQR from the first quartile, any data values that are less than this number are considered outliers. Similarly adding 1.5 x IQR to the third quartile, any data values that are greater than this number is considered as outliers.

K-Means Clustering:

K-Means clustering partitions the data into groups which contain similar objects. The instances which do not belong to any cluster (or) a cluster with fewer data points (or) forced to fit into a cluster are considered as outliers and are removed. It generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited for generating globular clusters. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori (Susana *et al*., 2010). The main idea is to

define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point, it is necessary to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After k new centroids have been chosen, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop, it is clear that the k centers change their location step by step until no more changes are done or in other words centers do not move any more (Velmurugan, 2012). Finally, this algorithm aims at minimizing an objective function known as squared error function given by the following.

$$J(v) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(||x_i - v_j||)^2 \quad (2)$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Genitic Algorithms:

GA is an evolutionary algorithm which offers multi criterion optimization for higher dimensional space problems. It is based on Darwin's theory of natural selection and 'survival of the fittest' (Goldberg, 1989). The basic genetic algorithm operators involved in reproduction are (Robert A. Richards):

- Selection – It deals with the selection of individuals of the population which will reproduce.
- Crossover – It takes a portion of each parent (as described below) and combines the two portions to create offspring.
- Mutation - The random alteration of a string / bit position, to covert 1^s into 0^s or vice versa

It can be used to solve different and diverse types of problems. The algorithm starts with a group of individuals (chromosomes) called a population. Each chromosome is composed of a sequence of genes that would be bits, characters, or numbers. Reproduction is achieved using crossover (2 parents are used to produce 1 or children) and mutation (alteration of a gene or more). Each chromosome is evaluated using a fitness function, which defines which chromosomes are highly-fitted in the environment. The process is iterated for multiple times for a number of generations until optimal solution is reached. The reached solution could be a single individual or a group of individuals obtained by repeating the GA process for many runs (W Li, 2004). The genetic algorithm employs the following process to produce the best individuals from the initial set of populations (Vishnu Raja *et al*., 2012):

Input: Set on N Chromosomes in the search space.

Output: Outliers with lowest fitness value.

Step 1: Generate random population of N individuals.

Step 2: [Fitness] Fitness function f(x) for each chromosome is evaluated.

Step 3: [New Population] Repeat the following steps to create new population

i) [Selection] Select two parents from the population according to their fitness

ii) [Crossover] with the crossover probability crossover the parents to form new offspring. If no crossover is performed the offspring is resulted as parents.

iii) [Mutation] with the mutation probability mutate the offspring at each locus.

iv) [Accept] Place new offspring in the population.

Step 4:[Replace] Use new generated population for the next iteration.

Step 5: [Test] If the termination condition is satisfied, return the best solution.

Step 6: [Result] Sort the fitness value in descending order, the lower value are identified as outliers.

Step 7: [Loop] Go to step 2 for next iteration

Fitness function can be calculated using the below formula:

$$\text{Fitness} = \frac{\text{Total No.of Correctl classified instances}}{\text{Total No.of Training samples}} \quad (3)$$

Support Vector Machines:

SVM is a classier that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. "Support Vectors" are defined as subset of data instances used to define the hyperplane. The distance between the hyperplane and the nearest support vector is called as margin (Vapnik, 1995). The Sequential Minimal Optimization (SMO) algorithm avoids working with numerical quadratic program routines by analytically solving a large number of small optimization sub-problems that involves only two Lagrange multipliers at the time. For any two multipliers α_1 and α_2, the constraints are reduced to the following:

$$0 \leq \alpha_1, \alpha_2 \leq C \quad (4)$$

$$y_1\alpha_1 + y_2\alpha_2 = K \quad (5)$$

This reduced problem can be solved analytically by finding a minimum of a one-dimensional quadratic function. The value of K is fixed in each iteration, it is the negative of the sum over the rest of terms in the equality constraint.

The algorithm proceeds as follows:

1. Find a Lagrange multiplier α_1 that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.

2. Pick a second multiplier α_2 and optimize the pair (α_1, α_2).
3. Steps 1 and 2 are repeated until convergence.

The quadratic programming problem has been solved, when KKT is satisfied by the Lagrange multipliers (John W. Chinneck, 2014). The above algorithm guarantees convergence and heuristic measures are used to choose the pair of multipliers so as to increase the rate of convergence.

Bayesian Network:

Bayesian networks also known as belief networks (or BayesNets), belong to the family of probabilistic graphical models. These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods (Ben-Gall, 2007). Hence, BayesNets combine principles from graph theory, probability theory, computer science, and statistics. A Bayesian network B is an annotated acyclic graph that represents a joint probability distribution over a set of random variables V. The network is defined by a pair B = (G, Ø), where G is the directed acyclic graph whose nodes X1,X2,...,Xn represents random variables, and whose edges represent the direct dependencies between these variables. The graph G encodes independence assumptions, by which each variable Xi is independent of its non descendents given its parents in G. The second component Ø denotes the set of parameters of the network. This set contains the parameter $\theta_{x_i|\pi_i}$ = $P_B(x_i|\pi_i)$ for each realization x_i of Xi conditioned on π_i ,the set of parents of Xi in G (Jie Cheng *et al*., 1997). Accordingly, B defines a unique joint probability distribution over V, namely:

$$P_B\ (X_1 \dots \dots X_n) = \prod_{i=1}^{i=n} P_B\ (X_i|\pi_i)\ = \prod_{i=1}^{i=n}\ \theta_{X_i|\pi_i} \qquad (6)$$

Random Forest :

The random forest algorithms form a family of classification methods that rely on the combination of several decision trees. The particularity of such Ensembles of Classifiers is that their tree based components are grown from a certain amount of randomness. Based on this idea, random forest is defined as a generic principle of randomized ensembles of decision trees (Breiman, 2003). The basic unit of random forest is a binary tree constructed using recursive partitioning (RPART). The random forest tree base learner is typically grown using the methodology of CART. Random forest trees differ from CART as they are grown non-deterministically using a two-stage randomization procedure. In addition to the randomization introduced by growing the tree using a bootstrap sample of the original data, a second layer of randomization is introduced at the node level when growing the tree. Rather than splitting a tree node using all variables, Random forest selects at each node of each tree, a random subset of variables, and only those variables are used as candidates to find the best split for the node. The purpose of this two-step randomization is to de-correlate trees so that the forest ensemble will have low variance, a bagging phenomenon (Patil *et al*., 2010). The construction of RF is described in the following main steps (Andy Liaw *et al*., 2002):

Step 1: Draw n_{tree} bootstrap samples from the original data.

Step 2: Grow a tree for each bootstrap data set. At each node of the tree, randomly select m variables for splitting. Grow the tree so that each terminal node has no fewer than node size cases.

Step 3: Aggregate information from the n_{tree} trees for new data prediction such as majority voting for classification.

Step 4: Compute an Out-Of-Bag (OOB) error rate by using the data not in the bootstrap sample.

An estimate of the error rate can be obtained, based on the training data, by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample (OOB data) using the tree grown
with the bootstrap sample.
2. Aggregate the OOB predictions (On the average, each data point would be out-of-bag around 36% of the
times, so aggregate these predictions). Calculate the error rate, and call it the OOB estimate of error rate.

Experimental Results:

The experiment is conducted using the datasets with null values and no null values. The datasets are cleaned by applying three different imputation methods (mean, median and class mean) and are subjected to outlier removal using Inter Quartile Range, K-Means clustering and GA. The resulting dataset is classified using SVM, BayesNet and Random Forest classifier. To achieve better classification accuracy, 10-fold stratified cross validation method is used. A variation of cross-validation is stratified cross-validation, where the class distribution in each fold is approximately the same as in the initial dataset (Breiman *et al*., 1984). The classification accuracy is calculated as given below:

$$Accuracy\ = \frac{TP+TN}{TP+FP+FN+TN} \qquad (7)$$

Where,

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.

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

During each trail / run, GA always has a tendency to produce different results (outliers). To achieve consistent result for removing outliers using GA, the experiment was repeated 50 times. Among the 50 test runs performed one trail / run which is closer to the average classification accuracy value is selected to compare the classification accuracy obtained by different classifiers used in the proposed method.

Table 2: Number of Instances after outlier detection for datasets with null values.

Dataset	Number of Instances after removing Outliers									
	Total	Imputation by Mean		Imputation by Median			Imputation by Class Mean			
		IQR	K -Means	GA	IQR	K-Means	GA	IQR	K -Means	GA
Diabetes	768	723	503	524	694	502	518	746	645	554
WBC	699	699	672	672	699	670	671	699	671	666

Table 3: Number of instances after outlier detection for datasets with no null values.

Dataset	Total	Number of Instances		
		IQR	K -Means	GA
WDBC	569	514	486	512
Heart (Stalog)	270	269	160	186

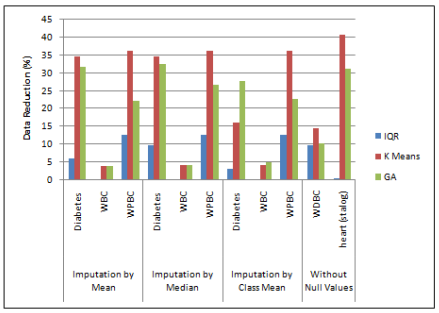

Fig. 1: Outlier detection (%) obtained using IQR, K-Means and GA.

Table 2 and 3 shows the number of instances after removing outliers for datasets with and without null values. A comparative study of outlier detection percentage is given in Fig. 1. From Fig. 1, it is clear that IQR results in less outlier detection percentage and K-Means with more outlier detection percentage. But when comparing with classification accuracy (SVM, BayesNet, Random Forest) obtained by the datasets, GA is better than IQR and K-Means for outlier removal which is evident from the results reported in Tables 4 to 6.

Fig. 2 and 3 compares the classification accuracy of SVM, BayesNet and Random Forest using the tested datasets subjected to GA for outlier removal. It is clear that SVM has highest classification accuracy in all the tested datasets except for diabetes diagnosis. For diabetes dataset, where imputations are more the classification accuracy of SVM is very much nearer to BayesNet or Random Forest. Therefore GA (outlier removal) with SVM (classification) can be considered as a suitable framework for classifying medical datasets. Finally a comparison based on imputation techniques is carried out for the above framework and shown in Fig. 4 and it is obvious that imputation by median leads to a high classification accuracy compared with imputation by mean or class mean.

Table 4: Classification Accuracy (%) for Diabetes Diagnosis.

Technique Used	Imputation by Mean		Imputation by Median			Imputation by Class Mean			
	IQR	K -Means	GA	IQR	K-Means	GA	IQR	K -Means	GA
SVM	77.31	97.61	99.80	69.30	98.40	99.80	79.75	94.41	96.57
BayesNet	75.10	99.60	100	99.85	99.60	100	88.60	99.06	99.81
Random Forest	74.66	99.80	100	99.71	99.80	100	87.80	99.53	99.63

Table 5: Classification Accuracy (%) for WBC.

Technique Used	Imputation by Mean		Imputation by Median			Imputation by Class Mean			
	IQR	K -Means	GA	IQR	K-Means	GA	IQR	K -Means	GA
SVM	96.99	99.55	99.85	96.85	99.25	100	96.85	99.25	99.84
BayesNet	97.28	99.25	99.40	97.13	99.40	99.70	97.13	99.25	99.39
Random Forest	95.99	99.10	99.10	95.42	99.10	99.40	95.70	99.10	99.39

Table 6: Classification Accuracy (%) for dataset with no null values (WDBC and Heart).

Technique Used	WDBC		Heart (Stalog)			
	IQR	K -Means	GA	IQR	K-Means	GA
SVM	96.88	100	100	84.75	91.87	99.46
BayesNet	94.55	100	99.60	84.75	91.87	99.46
Random Forest	94.94	100	98.63	80.29	96.87	98.92

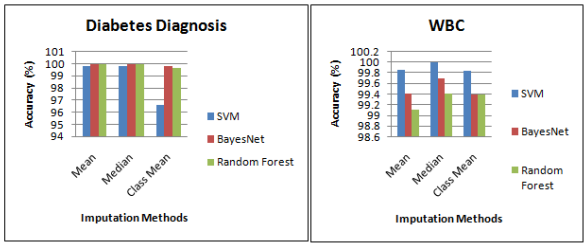

Fig. 2: Classification accuracy (%) for Dataset with null values after removing outliers using GA.

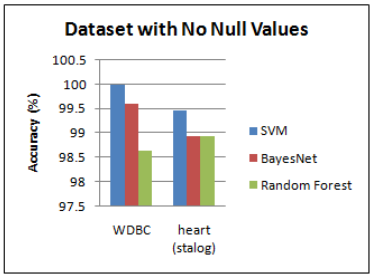

Fig. 3: Classification accuracy (%) for no null value dataset after removing outliers using GA.

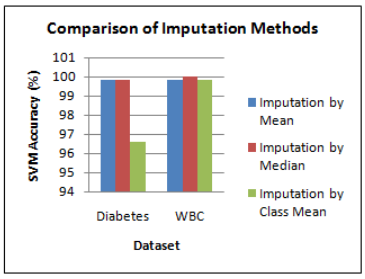

Fig. 4: SVM Classification accuracy (%) for different imputation methods (using GA for outlier removal).

Table 7 shows the comparison of the proposed method with the results reported in the literature for different datasets. It is proved that the accuracy percentage of proposed method shows better accuracy compared with the existing research

Research Findings:

1. Among the imputation techniques used, imputation by median shows a constant rate of classification accuracy in all the experimented datasets.
2. Removing outliers using GA yields better classification accuracy and less data reduction percentage when compared with K-Means.
3. GA can be used for outlier detection irrespective of the datasets containing with and without missing values.
4. For the tested datasets with less or no imputations, SVM provides better classification accuracy than BayesNet and Random Forest.
5. The datasets with imputations are more (nearly 50% of the dataset contains missing values), the use of BayesNet or RandomForest classifiers can be attempted.
6. Imputation by median followed by GA for outlier removal and SVM for classifying the datasets proves to be the best framework to achieve high classification accuracy.

Conclusion:

Performance of classification accuracy depends on data pre-processing. Out of the different data preparation methods proposed, imputation by median followed by GA for outlier detection and SVM to classify the datasets has a better

classification accuracy compared with K-Means and IQR for outlier detection. The proposed method shows improved classification accuracy of 100%, 100%, 100% and 99.46% for the datasets Diabetes Diagnosis, WBC, WDBC and Heart (Stalog). The future work will concentrate on suitable feature selection and dimension reduction techniques to be fitted with this framework. Also, the use of other classifiers in the neural family, fuzzy logic, bagging and boosting can be attempted to enhance the performance. Further, the framework can be tested across well known datasets in different domains such as finance, geography, population etc. reported in the literature.

Table 7. Comparison of the proposed work with the existing works in terms of Accuracy.

Author (Year)	Method	Accuracy (%)
Diabetes Diagnosis		
Radha *et al* (2014)	BLR proves to be better based on computing Time and precision value	75
Ravi *et al* (2014)	Genetic algorithm with SVM classifier	77.3
Veeran Vijayan V. *et al* (2014)	ANFIS algorithm with adaptive KNN	80
Keerthana *et al* (2014)	Density Based Cluster with Naïve Bayes classifier	96.35
Nihat Yilmaz *et al* (2014)	K-Means + SVM	93.65
	Modified K-Means + SVM	96.71
Santhanam T. *et al* (2014)	K-Means+GA+SVM	98.82
Proposed Method	Imputation by Median + GA + RandomForest / BayesNet	100
WBC		
Kharya *et al* (2014)	Bayesian Belief Network	91.83
Vikas Chaurasia *et al* (2014)	Sequential Minimal Optimization(SMO)	96.19
Naveen Chandra Yadav *et al* (2014)	Feed forward neural network model with back propagation learning algorithm	96.34
Bichen Zheng *et al* (2014)	K-Means and SVM	97.38
Omar S. *et al*(2014)	Differential Evolution (DE) and LS-SVM	99.75
Proposed Method	Imputation by Median + GA + SVM	100
WDBC		
Cuong Nguyen *et al* (2013)	Bayesian probability ranking for feature selection and Random Forest as classifier	99.82
Sridevi *et al* (2014)	Rough set K-Means Clustering	99.12
Olfati *et al* (2014)	PCA followed by GA for feature extraction and SVM as classifier	100
Sridevi *et al* (2014)	Rough Set and Correlation feature selection with MLP	100
Proposed Method	Imputation by Median + GA + SVM	100
Heart		
Negar Ziasabounch *et al* (2014)	ANFIS	92.3
Kittipol Wisaeng (2014)	KNN	93
Preeti Gupta *et al* (2014)	Genetic based Artificial Neural Network	97.75
Nihat Yilmaz *et al* (2014)	m.kmeans+SVM	97.87
Deepali Chandna (2014)	Information Gain + ANFIS	98.24
Proposed Method	Imputation by Median + GA + SVM	99.46

REFERENCES

Acuna, E. and C. Rodriguez, 2004. The treatment of missing values and its effect in the classifier accuracy. In D. Banks, L.House, F.R. McMorris,P. Arabie, W.Gaul (Eds).Classification, Clustering and Data Mining Applications, Springer-Verlag Berlin-Heidelberg, 639-648.

Andy Liaw and Matthew Wiener, 2002. Classificationand Regression by randomForest. R News, Vol.2/3.

Asha, T., S. Natarajan and K.N. Balasubramanya Murthy, 2011. A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification. vol. abs/1108.1045.

Asha Gowda Karegowda, M.A. Jayaram and A.S. Manjunath, 2012. Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. International Journal of Engineering and Advanced Technology (IJEAT),ISSN, 2249–8958-1 -3.

Ben-Gal, I., F. Ruggeri, F. Faltin and R. Kenett, 2007. Bayesian Networks. Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons.

Bichen Zheng, Sang Won Yoon and Sarah S. Lam, 2014. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert systems with applications, 41: 1476-1482.

Breiman, L., 2003. Random forests. Machine Learning Journal Paper, 45: 5-32.

Breiman, L., J. Friedman, R. Olshen and C. Stone, 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Breiman, L., P. Spector, 1992. Submodel selection and evaluation in regression. The X-random case. Internat Statist. Rev., 60: 291-319.

Bruno Fernandes Chimieski and Rubem Dutra Ribeiro Fagundes, 2013. Association and Classification Data Mining Algorithms Comparison over Medical Datasets. Artigo Original, www.jhi-

sbis.saude.ws, J. Health Inform. Abril-Junho, 5(2): 44-51.

Chin-Yuan Fan, Pei-Chann Chang, Jyun-Jie Lin and J.C. Hsieh, 2011. A Hybrid model combining Case-based reasoning and Fuzzy Decision Tree for Medical Data Classification. Applied Soft Computing, 11(1): 632–644.

Courtnery Taylor. How do we determine what is an outlier. statistics.about.com.

Cuong Nguyen, Yong Wang and Ha Nam Nguyen, 2013. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. Biomedical Science and Engineering, 6-55.

Goldberg, D.E., 1989. Genetic Algorithm in Search, Optimization, and Machine Learning. Addison Wesley, Boston.

Hemant, P. and T. Pushpavathi, 2012. A novel approach to predict diabetes by Cascading Clustering and Classification. Computing Communication & Networking Technologies (ICCCNT), Third International Conference, 1-7.

Jiawei Han and Micheline Kamber, 2006. Data mining concepts and techniques. Second Eddition, Elsevier.

Jie Cheng, David A. Bell and Weiru Liu, 1997. Learning belief networks from data: an information theory based approach. Proceedings of the sixth international conference on Information and knowledge management, Las Vegas, Nevada, USA, 325-331.

John, W., Chinneck, 2014. Practical Optimization: a Gentle Introduction", http:/ / www. sce. carleton. ca / faculty/ chinneck /po/ Chapter20.pdf.

Johnson, R., 1992. Applied Multivariate Statistical Analysis. Prentice Hall.

Kalaivany Natarajan, Jiuyong Li and Andy Koronios, 2010. Data mining techniques for data cleaning. Engineering Asset Lifecycle Management, 796-804.

Keerhtana, G. and Dr. V. Srividhya, 2014. Performance Enhancement of Classifiers using Integration of Clustering and Classification Techniques",International Journal of Computer Science Engineering (IJCSE), ISSN, 2319-7323, 3-03.

Kharya, S., S. Agrawal and S. Soni, 2014. Using Bayesian Belief Networks for Prognosis & Diagnosis of Breast Cancer. International Journal of Advanced Research in Computer and Communication Engineering, 3-2.

Kittipol Wisaeng, 2014. Predict the Diagnosis of Heart Disease Using Feature Selection and k-Nearest Neighbor Algorithm. Applied Mathematical Sciences, 8(83): 4103-4113.

Li, W., 2004. Using Genetic Algorithm for Network Intrusion Detection, Proceedings of the United States department of Energy Cyber Security Grou, Training Conference, 8: 24-27.

Maytal Saar-Tsechansky and Foster Provos, 2007. Handling Missing Values when Applying Classification Models. Journal of Machine Learning Research, 8:1217-1250.

Naveen Chandra Yadav and Prafull Gajbhiye, 2014. Diagnosis of Breast Cancer Using Neural Network Approach. BMR Bioinformatics & Cheminformatics, 1-1.

Negar Ziasabounchi and Iman Askerzade, 2014. ANFIS Based Classification Model for Heart Disease Prediction. International Journal of Electrical & Computer Sciences IJECS-IJENS, 14-02.

Nihat Yilmaz, Onur Inan and Mustafa Serter Uzer, 2014. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. Springer:Transaction Processing Systems:J Med Syst., 38-48.

Olfati, E., H. Imam Khomeini and M.A Shoorehdeli, 2014. Feature subset selection and parameters optimization for support vector machine in breast cancer diagnosis. Intelligent Systems (ICIS), Iranian Conference onate of Conference:IEEE.

Pachgade, S.D. and S.S. Dhande, 2012. Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach. International Journal of Advanced Research in Computer Science and Software Engineering, 2- 6.

Patil, B.M., R.C. Joshi and D. Toshniwal, 2010. Impact of K-Means on the performance of classifiers for labeled data.Comm. Com.Inf. Sc., 94: 423-434.

Preeti Gupta and Dr. Bikrampal Kaur, 2014. Accuracy Enhancement of Heart Disease Diagnosis System Using Neural Network and Genetic Algorithm. IJARCSSE, 4-8.

Radha, P. and Dr. B. Srinivasan, 2014. Predicting Diabetes by cosequencing the various Data Mining Classification Techniques. IJISET-International Journal of Innovative Science, Engineering & Technology, 1-6.

Ravi Kumar, G., Dr. G.A Ramachandra and K. Nagammai, 2014. An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets. International Journal of Advanced Research in Computer Science and Software Engineering, 4-2.

Robert, A., Richards. zeroth-order shape optimization utilizing a learning classifier system. Classifier Systems and Genetic Algorithms, from – http:// www.stanford.edu / ~buc/ SPHINcsX / book.html.

Santhanam, T. and M.S. Padmavathi, 2014. Comparison of K-Means Clustering and Statistical Outliers in Reducing Medical Datasets. International Conference on Science Engineering and Management Research (ICSEMR), IEEE, 1-6.

Sridevi, T. and A. Murugan, 2014. An Intelligent Classifier for BreastCancer Diagnosis based on K-Means Clustering and RoughSet. International

Journal of Computer Applications (0975–8887), 85-11.

Susana, A., J. Leiva-Valdebenito Francisco and Torres-Aviles, 2010. A Review of the Most Common Partition Algorithms in Cluster Analysis: A Comparative Study. Colombian Journal of Statistics, ISSN: 0120-1751, 33(2): 321-339.

Vapnik, V.N., 1995. The Natural of Statistical Learning theory. Springer–Verleg, New York,USA.

Veeran Vijay, V. and Aswathy Ravikumar, 2014. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus. International Journal of Computer Applications (0975–8887), 95–17.

Velmurugan, T., 2012. Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points. Int.J.Computer Technology & Applications, 3(5): 1758-1764.

Vikas Chaurasia and Saurabh Pal, 2014. A Novel Approach for Breast Cancer Detection using Data Mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering, 2-1.

Vishnu Raja, P. And Dr.V.Murali Bhaskaran, 2012. An Effective Genetic Algorithm for Outlier Detection. International Journal of Computer Applications (0975–8887), 38–6.