Volume 7, Issue 3, March 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Medoid Based Approach for Missing Values in the Data Sets Using AANN Classifier

Vinotha P G, Uthra V, Dr Anto S

Department of CSE, Sri Krishna College of Technology, Coimbatore,
Tamilnadu, India

DOI: 10.23956/ijarcsse/V7I3/0102

Abstract— The imputation of the missing values in real life databases has been realized only through traditional and conventional means. It weakens the whole approach for raw and unstructured data. This paper proposes a Medoid based approach to overcome this problem. The clusters are formed based on the similarities of missing values in the dataset in an iterative manner. The missing values are imputed based on an accurate mathematical model using Partitioning Around Medoids (PAM) algorithm. The datasets are then trained and tested using Auto Associative Neural Network (AANN) classifier that approximates the error function. The dataset used for this paper is Pima Indian Diabetes (PID) dataset retrieved from University of California, Irvine (UCI) machine learning repository. The result shows that our proposed method reduces the error rate in lesser computation time.

Keywords— Missing values, Medoids, Partitioning Around Medoids, Auto Associative Neural Network classifier, Pima Indian Diabetes dataset.

I. INTRODUCTION

Artificial Intelligence (AI) has a profound influence in the field of medical expert systems and pattern recognitions. The clinical analysis to find the root cause and after math of a disease is important in the field of medicine. The maintenance of these medical data sets is a tedious task wherein the details are maintained with utmost precision [1]. AI helps to design an expert system that makes decision and acts like human. Some patients have privacy concerns with their details so this system is helpful in achieving it as the decision is monitored only by the system and so the data is dealt with security. Missing values are mainly due to human error as the observer and the respondent (patient) may do miscommunication while taking the survey. Since hospitals have sample data bases for each disease it is impossible for humans to maintain it all single handedly. The data base software is also prone to error so it is not reliable [1]. The solution obtained from such type of data sets is not optimal hence results in premature convergence. This is a condition where the system converges when it encounters missing values thus it is a hindrance for research. The imputation of the missing values is a mandatory step that has to be done before the commencement of the actual work. The k-Medoid based approach combined with PAM (Partitioning Around Medoids) algorithm is used for imputation. Machine Learning (MI) is a branch of AI that uses a supervised learning methodology. The data samples are divided based on 80(Training) -20(Testing) rule. Based on the number of samples that passes the testing phase analysis is made and its effectiveness is improved further.

Data mining is a stream in computer science that helps to mine and extract the knowledge rather than the data set itself. Data mining approach plays a vital role in AI in cleaning the unstructured data. The process involved in the preprocessing steps helps to structure the raw data available. It helps to remove redundancy and reduce dimensionality [2]. It normalizes and makes it simpler for research. The Genetic Algorithm (GA) is a process of selection based on the fitness value calculated for each feature and filters the least significant. This technique is useful to select only the best fit features and avoid the rest to make the work simpler.

The classification approach is aided using Auto Associative Neural Network (AANN). Some tuples may fail in the testing stage due to uncertainties, thus increases the error rate and makes the expert system to collapse.

The specialty of this network is that it reduces the error rate with less number of nodes..The following sections consist of the previous works related to this field, followed by the proposed system and the performance analysis done to compare this approach with the existing system.

II. RELATED WORK

For the imputation of missing data, Pedro J.CarciaLaencina et al (2009) have Proposed a novel KNN imputation procedure using a feature weighted distance metric based on mutual information(MI) [3]. This method provides the missing data estimation aimedat solving the classification task. For acquiring the required knowledge in the form of production rules Mu-Chin Su have proposed a novel class of neural networks to articulate the knowledge it learned from a set of examples in order to provide an appealing solution to the problem of knowledge acquisition. For simplifying the application of Artificial Neural Networks(ANNs) for solving wider range of problems, Arso M. Vukicevic et al, have proposed an expert system that reduces the complexity of ANNs development compared to the previous studies that required manual selection of the optimal features and/or ANN configuration [4]. Thus the expert system represents a robust and user-friendly framework. In order to find the best medical Expert System(ES) Dinesh Grover et al, have presented a comprehensive study of medical ES for diagnosing various diseases. Thus it gives an précised view diagnostic ES and provides an analysis about already existing studies. For providing a good medical services, Moshood A.Hambali, have proposed Medical Expert System that addresses the various method for diagnosing a disease [5]. The MES comprises of 46 rules for diagnosing and is capable of assisting medical experts in diagnosing diseases and to provide good health services to their patients. Diabetes diagnosis is also dealt using Multi-Layer Neural Network by Hasan Temurtas et al. The accuracy of the system is comparatively low. The key problem in a Neural Network (NN) is determining the number of hidden nodes which affects accuracy [6]. To overcome this problem, the proposed system uses ELM on Single Hidden Layer Feed Forward Network, in which the hidden nodes are randomly selected and the optimal number of hidden nodes is determined by SA.Kayaer and Yidirim (2003) presented a classifier for PID dataset classification, using general regression neural networks, which produced an accuracy of 80.21%.Hasan et al. (2009) have used Levenberg-Marquardt (LM) algorithm for training a multilayer neural network structure to diagnose diabetes [7]. An accuracy of 82.37% is obtained which is low when compared to other existing diabetes diagnosis systems.

III. UCI MACHINE LEARNING REPOSITORYAND DISEASE DATASET

The University of California, Irvine (UCI, UC Irvine, or Irvine), is a public research university located in Irvine, California, United States, and one of the 10 campuses in the University of California (UC) system. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [4].

The Pima Indian Dataset (PID) is used from UCI. The PID dataset consists of 768 instances with 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’, where ‘1’ means a positive test for diabetes and ‘0’ is a negative test for diabetes. This dataset consist of 8 features namely, number of times pregnant, plasma glucose concentration (PGC), diastolic blood pressure (DBP), triceps skin folds thickness (TSFT), insulin, body mass index (BMI), diabetes pedigree (DP) function and age. All the patients of this dataset are females with at least 21 years of Pima Indian heritage. This dataset has missing values and missing value imputation is not done since it has less impact on classification accuracy.

IV. PROPOSED METHOD

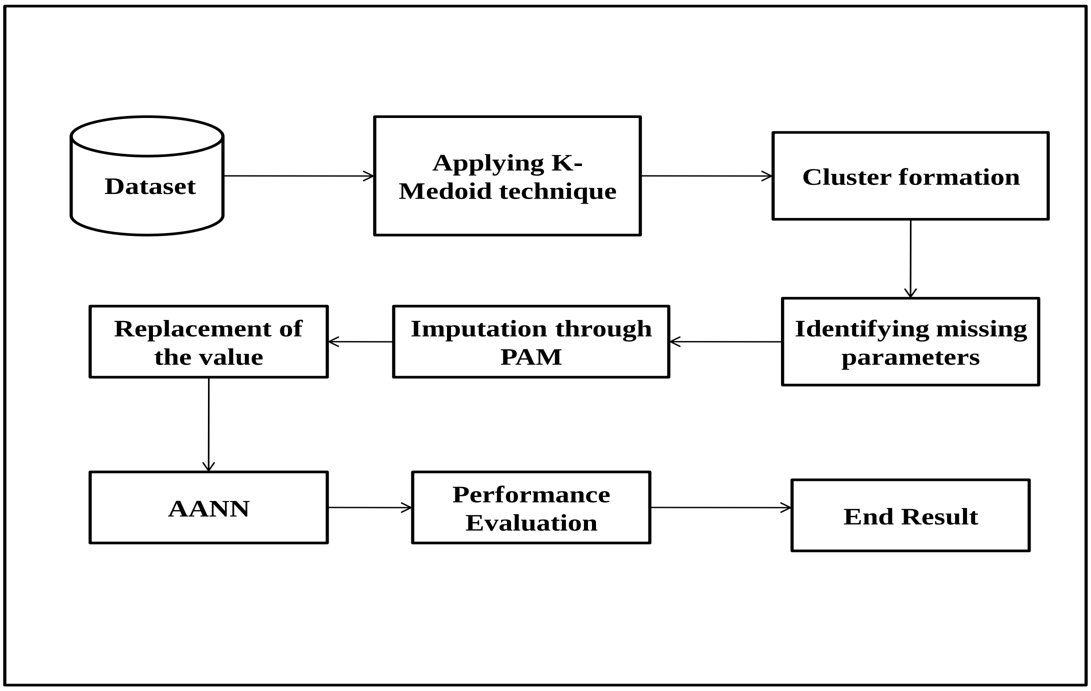


Fig 1. Block diagram of the proposed approach

The proposed method consists of following phases.

1) K Medoid Approach
2) Imputation through PAM Algorithm
3) Auto Associative Neural Network Classifier

A. K Medoid Approach

A Medoid is the representation of data objects in a data set. The k-Medoid algorithm is based on clustering and partitioning technique. On comparison with K-Means this approach is more robust. It partitions the data objects into K clusters. The algorithm flow works in such a way that it selects the data points in the dataset as Medoids in a random manner and the distance between the similar groups are computed after which the clusters are formed.

B. Imputation through PAM Algorithm

The Partitioning Around Medoids (PAM) algorithm imputes the missing values. The clusters formed in the previous module helps to categorize the values based on the missing terms, after which it is replaced by the value obtained using the formula given below.

$$\sum_{i=1}^{n} C_i \; / \; (n-m)$$

n=9 (Total datas in a row)
Missing values (0)= m
C= value of each data in a row

This is a mathematical model based approach where the values replaced are accurate rather than just making random guesses or considering only the nearest value and replacing it.

C. Auto Associative Neural Network Classifier

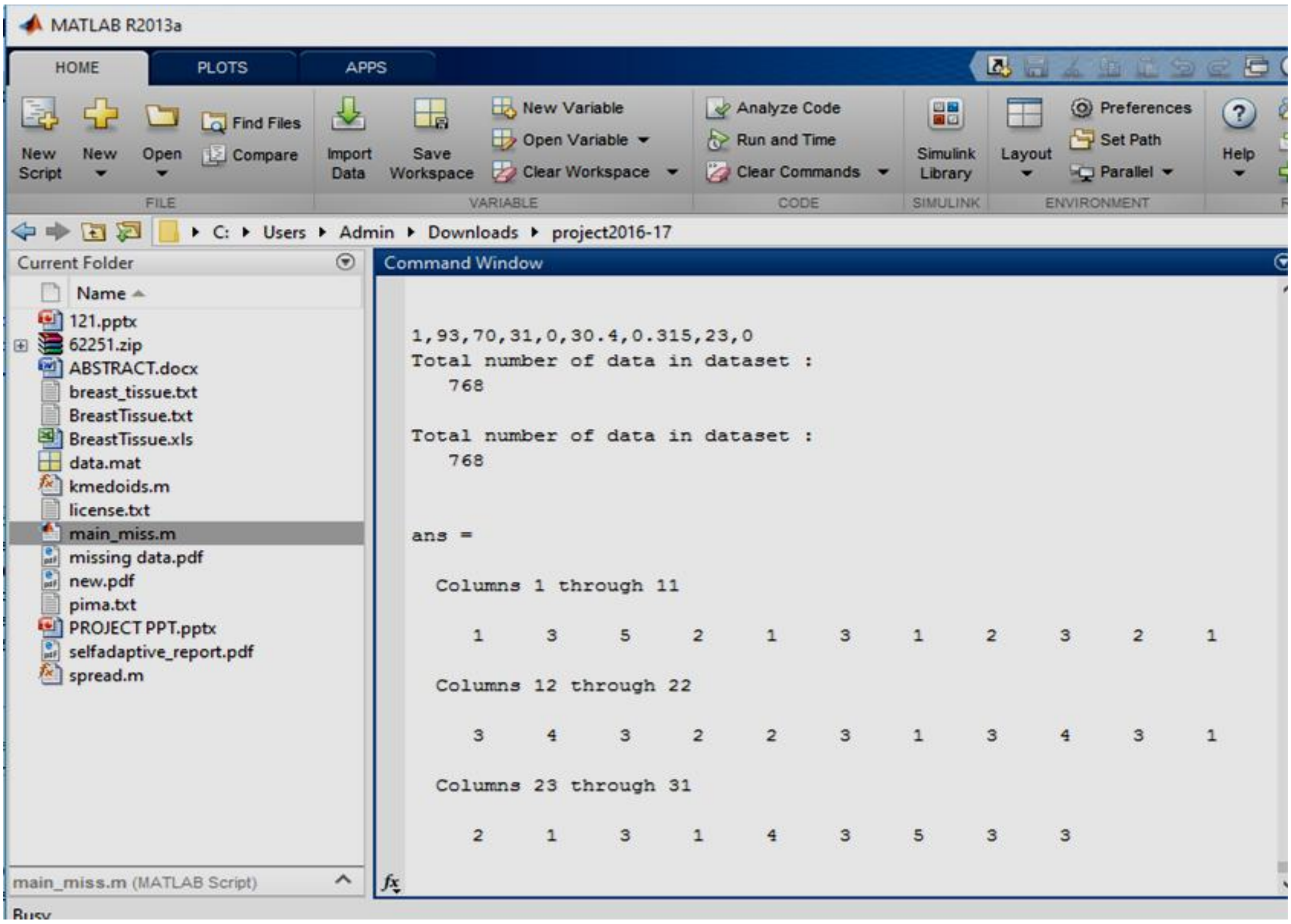
The neural network consists of an input and an output layer in between which is the hidden layer present. The number of nodes and weights forms the parameters of a neural network. The Auto Associative Neural Network is more efficient in making the hidden layer to reduce the error rate with minimal number of nodes. This hidden layer recognizes only the non-redundant data, thus making the computation time more robust. It captures the non-linear dependencies. It improves the expressiveness of the network and helps to represent more complex models in a simpler way. The effectiveness in approximating the error function depends on the number of nodes in the hidden layer. The neural network requires more number of runs to find the best solution. The hidden layer must be trained in such a way that the target is achieved with lesser number of runs.

V. EXPERIMENTATION AND RESULTS

The proposed model is implemented using MATLAB 8.0.1(R2013a).

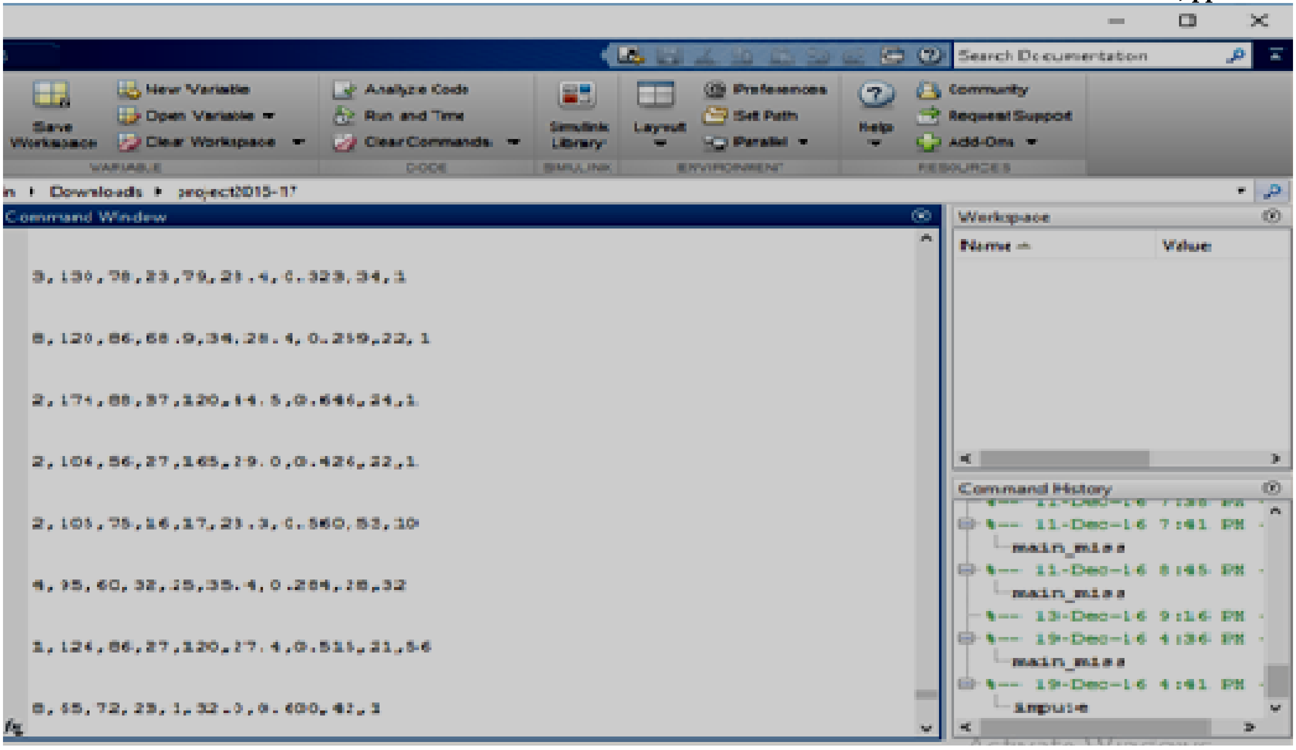
INPUT - Pima Indian Diabetes Data sets with missing values
METHOD - K- Medoid
OUTPUT - Cluster formation based on similarities in the values missing in each row

The data sets obtained contained 768 values/instances in total along with missing values. The data sets are imported into the workspace for which medoids are formed at random. Three medoids or cluster groups are formed dividing the data set as columns of say, 1 through 11, 12 through 22 and 23 through 31. Five layers are formed. The first layer is occupied by rows with no missing values, second layer by rows with one missing value and so on till layer 5.



```
1,93,70,31,0,30.4,0.315,23,0
Total number of data in dataset :
   768

Total number of data in dataset :
   768


ans =

  Columns 1 through 11

     1     3     5     2     1     3     1     2     3     2     1

  Columns 12 through 22

     3     4     3     2     2     3     1     3     4     3     1

  Columns 23 through 31

     2     1     3     1     4     3     5     3     3
```

Fig 2. Formation of Clusters using K Medoids



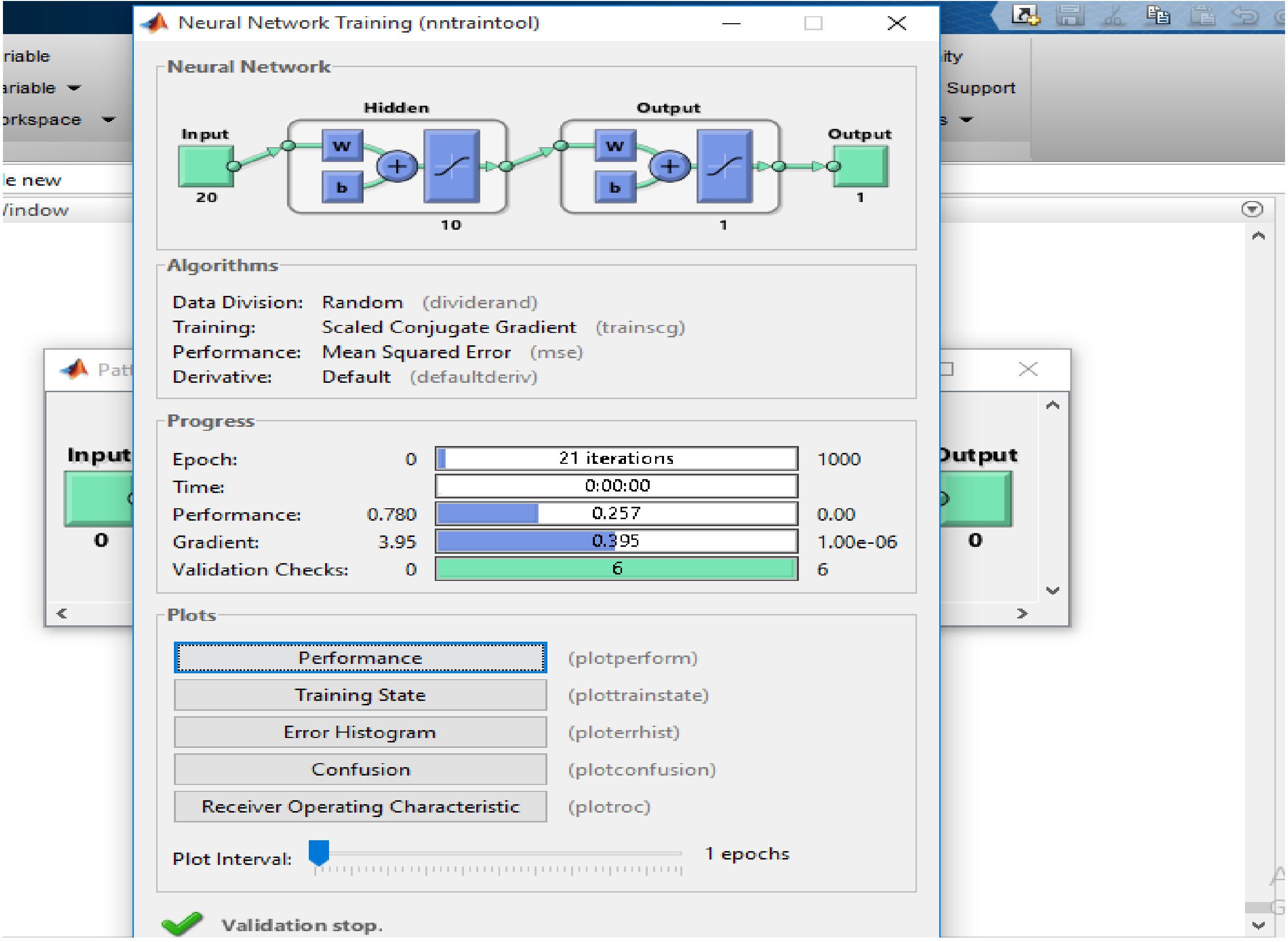
```
3, 130, 78, 23, 79, 28.4, 0.323, 34, 1

8, 120, 86, 68 .9,34.28.4, 0.259,22, 1

2, 174, 88, 37, 120, 44.5, 0.646, 24, 1

2, 104, 56, 27, 165, 29.0, 0.426, 22, 1

2, 105, 75, 16, 17, 23.3, 0.560, 53, 10

4, 95, 60, 32, 25, 35.4, 0 .284, 28, 32

1, 124, 86, 27, 120, 27.4, 0.511, 21, 56

8, 65, 72, 23, 1, 32.0, 0.600, 42, 1
```

```
%-- 11-Dec-16 7:41 PM --%
   main_miss
%-- 11-Dec-16 8:45 PM --%
   main_miss
%-- 13-Dec-16 9:16 PM --%
%-- 19-Dec-16 4:36 PM --%
   main_miss
%-- 19-Dec-16 4:41 PM --%
   impute
```

Fig 3. Imputed values using PAM

A. Testing-Training

80% of the samples are selected and trained and the rest 20% are tested, it shows a varied results and a few tuples have even failed in the testing phase. The hidden layer is trained with 10 nodes kepf fixed and the process is windowed to 21 iterations and at the 15 th iteration the error is approximated to 0.15 where the training and testing converges at the best fit and all the tuples have cleared the testing phase.



Algorithms

Data Division: Random (dividerand)
Training: Scaled Conjugate Gradient (trainscg)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch: 0 21 iterations 1000
Time: 0:00:00
Performance: 0.780 0.257 0.00
Gradient: 3.95 0.395 1.00e-06
Validation Checks: 0 6 6

Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Confusion (plotconfusion)
Receiver Operating Characteristic (plotroc)

Plot Interval: 1 epochs

✓ Validation stop.

Fig 4. Training of Neural Network

B. Performance metrics

The comparison is made among the existing approach to our proposed approach, it clearly shows that our approach has approximated the error function to a greater extent. The comparison is given below.

Table I Comparison of Error Approximation

METHOD	K Means	K-Medoid-RBFN classifier	K-Medoid-AANN classifier
ERROR VALUE	1.265	1.015	0.15191

VI. CONCLUSION

Occurrence of missing values is very common in many datasets belonging to various domains. The imputation based methods have been used to find appropriate replacements for missing values. Moreover, the project phase is completed successfully by forming Clusters and replacement for missing values using K Medoid approach and error rate is approximated by training the neural network with lesser number of runs.

REFERENCES

[1] Ch. Sanjeev Kumar Dash (2016) ‘Design of Self-Adaptive and Equilibrium Differential Evolution Optimized Radial Basis Function Neural Network ClassiÞer for Imputed Database’, Pattern Recognition Letters (2016).

[2] Kayaer, K. and Yidirim, T. (2003) ‘Medical diagnosis on Pima Indian diabetes using general regression neural networks’, Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing, pp.181–184.

[3] Hasan, T., Nejat, Y. and Feyzullah, T. (2009) ‘A comparative study on diabetes disease diagnosis using neural networks’, Expert Systems with Applications, Vol. 36, No. 4, pp.8610–8615.

[4] M. Forina., 1991, “Pima Indian Diabetes Dataset”, URL:http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[5] E. I. Mohamed, R., Linderm, G. Perriello, N. Di Daniele, S. J. Poppl, A. De Lorenzo., 2002, “Predicting type 2 Diabetes using an electronic nose-base artificial neural network analysis”, Diabetes Nutrition & Metabolism, Vol. 15, pp.215-221.

[6] Eleni l Georgia, Online prediction of glucose concentration in type 1 diabetes using extreme learning machines, Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE

[7] Sumi Alice Saji; K. Balachandran, Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction,2015 International Conference on Advances in Computer Engineering and Applications