Genetic Chromo Dynamic Framework for Imputation of Missing Data in Type II Diabetes Databases

Ilango Paramasivam
PhD Research Scholar, Department of Computer Applications
National Institute of Technology
Tiruchirappalli, India
Assistant Professor(Sr.), School of Computing Sciences
VIT University, Vellore, India
ilangosarojini@yahoo.com

Hemalatha Thiagarajan
Professor, Department of Mathematics
National Institute of Technology
Tiruchirappalli, India
hema@nitt.edu

and

Nickolas Savarimuthu
Assistant Professor, Department of Computer Applications
National Institute of Technology
Tiruchirappalli, India
nickolas@nitt.edu

***Abstract*- Data Mining approaches have been widely applied in the field of healthcare to facilitate the quality diagnosis. The physiologic state and the complexity of disease of a patient is monitored through diverse variety of symptoms and laboratory test measurements. Data acquisition in healthcare systems is voluminous; they come from many different sources, not all commensurate structure or quality. The performance of the data mining algorithms highly depends on the quality data. In medical domain, missing data might occur as the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns or it may be unfeasible for the patient to undergo the clinical tests, equipment malfunctioning, etc. Methods for resolving missing values are therefore needed in healthcare systems to enhance the quality of diagnosis. In this paper, Genetic Chromo dynamic framework GCFIT_MIS_IMPUTE is proposed for the imputation of missing data. The framework is experimented on Pima Indian Type II Diabetes Dataset and the performance is evaluated using Average Imputation Error.**

***Keywords*- data mining, missing data, imputation methods, Genetic Chromo dynamic Framework, Average Imputation Error(AIE), PIMA dataset.**

I. INTRODUCTION

With the exponential growth of information technology investments in healthcare systems and widespread diffusion of medical data repositories, large volumes of medical data are generated and collected by medical institutions. The interest in systems for autonomous decisions in medical applications is growing, as data is being generated almost daily, in huge repositories with heterogeneity in nature. This volume of data is devoted to the relatively young and growing field of medical data mining and knowledge discovery. Central to leveraging the rapidly expanding corpus of medical data is the discovery of the knowledge, data regularities or high-level information essential to supporting medical diagnostic decisions, improving the quality of patient care, etc. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database increasingly becomes necessary. Data mining in medicine can deal with this problem. It can also improve the management level of hospital information and promote the development of telemedicine and community medicine. Because the medical information is characteristic of redundancy, multi-attribution, incompletion and closely related with time, medical data mining differs from other one.

Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. Survival analyses is a field in medical prognosis that deals with application of various methods to historic data in order to predict the survival of a particular patient suffering from a disease over a particular time period. With the increased use of computers powered with automated tools, storage and retrieval of large volumes of medical data are being collected and are being made available to the medical research community who has been interested in developing prediction models for survivability. As a result, new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets [1][2].

Data mining approaches have been widely applied in the field of Health care [3][4][5]. By interpreting the patient data using the mined knowledge, the healthcare personnel will be assisted to make a guided decision about the clinical case at hand. Many studies in health informatics literature have investigated the effectiveness of the clinical decision support systems and concluded that these systems are indeed helpful [6] for effective medical diagnosis.

Typically, a data mining process comprises of six steps: understanding the problem domain, understanding the data, preparing the data, data mining, discovering knowledge evaluation, and finally using the discovered knowledge [7]. It is estimated that about 20 per cent of the effort is spent on business objective determination, about 60 per cent on data preparation and about 20 per cent on data mining and analysis of knowledge and knowledge assimilation steps, respectively [8]. It is understood that more time is spent on data preparation. Actually, the real-world datasets lack a lot of data quality problems like incompleteness, redundancy, inconsistency, or noisy data. These serious quality problems if not addressed will certainly reduce the performance of data mining algorithms [9]. Hence in many cases, a lot of effort and time is spent on data pre-processing phase. The application of efficient and sound data pre-processing procedures can reduce the amount of data to be analyzed without losing any critical information, improve the quality of the data, enhance the performance of the actual data mining algorithms and reduce the execution time of mining algorithms [9]. A number of widely used and effective data pre-processing techniques that proved to be useful in practice include: data cleaning, integration, and transformation [10]. In addition to these, feature selection, extraction, construction and discretisation are also widely applied [10] [11].

In a real-world environment, there are many possible reasons why the inaccurate or inconsistent data occur in a database, e.g., equipment mal functioning, the deletion of data due to inconsistency with other recorded data, not entering data due to poor understanding, considering the data as unimportant at the time of entry, etc. This may cause numerous missing data in the database, which can impact the performance of the data mining algorithm. Generally the presence of 1% missing data in the database is considered trivial and 1% to 5% is manageable. However, sophisticated methods and tools are required to handle 5-15% missing data.

Many methods for dealing with missing data from ignoring the instances containing missing data to impute the missing data are found in the literature [12]. Among them, the imputation methods fill-in the missing values by attributing them to other available data. Imputation minimizes Bias, and uses 'expensive to collect' data, that would otherwise be discarded [12]. This paper proposes a framework GCFIT_MIS_IMPUTE to impute the missing value by selecting the best chromosomes which matches the missing records and imputing the missing attribute value using the attribute values available in the selected best chromosomes. The approach is tested on Pima dataset [13] and the experiments were performed using Matlab. The results are validated by quantifying the error for the percentage of missing data, the average imputation error [14]. The proposed approach is compared against five methods namely 10-NN method, Mean-based imputation, two correlation-based methods known as LSImpute_Rows and EMImpute_Columns [15], and a multiple imputation (MI) method referred to as NORM [16].

The review of the different methods to handle missing data in various data mining applications is discussed in section II. Section III describes the dataset to evaluate the proposed approach. The objective and problem definition of the proposed system is given in section IV. Section V and VI discusses the experimental results and the performance analysis.

II. RELATED WORK

Diabetes is a disease that results in severe complications such as blindness, kidney failure, amputation and CVD. Data mining may support the discovery of relevant predictive patterns in diabetes databases, and may be applied to predict complications. Previous research includes the application of data mining techniques, e.g. clustering and rule-based systems, to predict CVD risk scores using different clinical factors [17, 18]. Cardiovascular Disease (CVD) is the leading cause of death among people with diabetes, accounting for at least two out of three diabetes-related deaths [19]. The design of domain-specific, knowledge-driven data mining tools for CVD complication prediction in diabetes datasets deserves deeper investigations.

In healthcare systems, clinical databases often contain a substantial amount of missing data, due to the lack of test results and due to certain interventions or administrative inaccuracies. It is not uncommon to encounter databases that have up to a half of the entries missing, making it very difficult to mine them using data analysis methods that can work only with complete data. A common way of dealing with this problem is to impute (fill-in) the missing values. The data may be missed in medical databases due to procedural errors, refusal of response or non-applicable of responses [12]. When using such a database for any data mining process, it is important to have a complete dataset as possible. If data are imputed, the validity of these values should be assessed. More importantly, research [12] indicates that a meaningful treatment of missing data shall always be independent of the problem being investigated.

Many methods for dealing with missing data are found in the literature [12]. One of the simplest methods to handle missing data is eliminating instances that contain missing values [9], but it lowers the quality of the mining process [10]. Another alternative approach is to replace them by using the average value replacing missing values with a global constant or attribute mean [9]. But these methods can be used only when the percentage of missing data is below 5%. A typical way is imputation of missing data - to fill in missing values by attributing them to other available data. Imputation is the process of estimating missing data of an observation based on valid values of other variables [12]. However, these techniques can only be applied when the attribute consists of numeric data because the statistical measures used are defined only on numerical values.

Several methods to deal with missing data have been proposed. One general approach to handle missing data is to create data mining algorithms that ‘‘internally’’ handle missing data and still produce good results. For example, the CART decision-tree learning algorithm [23] internally handles missing data essentially using an implicit form of imputation based on regression. Another common technique for dealing with missing data is to create a new data value (similar to ‘‘missing data’’) and use it to represent missing data. However, this has the unfortunate side effect that data mining algorithms may try to use “missing” as a legal value, which is likely to be inappropriate. It also sometimes has the effect of artificially inflating the accuracy of some data mining algorithms on some datasets [24]. Some of the simpler pre-processing techniques for handling missing data have limited applicability or introduce bias into the data [25].

Imputation techniques range from fairly simple ideas (such as using the mean or mode of the attribute as the replacement for a missing value [26][27) to more sophisticated ones that use regression [28], Bayesian networks [29], and decision-tree induction [30]. Using the mean or mode is generally considered a poor choice [30], as it distorts other statistical properties of the data (such as the variance) and does not take dependencies between attributes into account. Hot-deck imputation [31] fills in a missing data using values from other rows of the database that are similar to the row with the missing data. Regression imputation [28] imputes missing data with predicted values derived from a regression equation based on variables in the dataset that contain no missing data. Regression assumes a specific relationship between attributes that may not hold good for all datasets. Some of the missing data imputation techniques like K-Nearest Neighborhood (KNN)[32], LSImpute_Rows, and EMImpute_Columns [33] are found in the literature.

In recent years, machine learning (ML) algorithms were introduced to develop imputation methods [22][34]. In contrast to statistical methods, ML algorithms generate a data model from data that contain missing values, and next the model is used to perform classification that imputes the missing values. Several different types of ML algorithms were used, such as decision trees, probabilistic, and rule-based methods [20], however the underlying methodology was the same.

Most healthcare datasets contain a lot of missing values [12]. Numerous case studies are found in the literature regarding the imputation of missing data in medicine [35] [36]. In medical data analysis two factors make the interpretation of predictive results very difficult. First, certain types of data, for example demographics, tend to be more accurate than other types, such as laboratory results. Second, the analysis of missing data may require large amounts of complete, reliable data in order to make accurate estimations [14]. Different estimation methods based on the least square principle and multiple linear regressions for gene expression data analysis are compared in [15]. Similarly, comprehensive comparison of estimation methods for bioinformatics applications: a method based on Singular Value Decomposition, weighted K-NN, and a row average model is found in [37].

III. DATASET

The Pima Indian diabetes dataset [13], includes 768 complete instances described by 8 features (labeled as number of times pregnant, glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function and age). The class distribution is class value 1 interpreted as "tested positive for diabetes" in 268 numbers of instances and class value 0 in 500 numbers of instances. There is no missing data present in the training dataset.

IV. OBJECTIVE / PROBLEM DEFINITION

The objective of the proposed approach, GCFIT_MISS_IMPUTE is to impute the missing values in the Type II diabetes Databases and to evaluate its performance by estimating average imputation error. The average imputation error is the measure which represents degree of inconsistency between the observed and imputed values. The approach is experimented on PIMA Indian Type II Diabetes Dataset, which originally do not have any missing data. All the 8 attributes are considered for the experiments as the decision attribute is derived using these 8 attributes. Datasets with different percentage of missing data (from 5% to 85%) were generated using random labeling feature. For each percentage of missing data, 20 random simulations are to be conducted.

The framework comprises of four modules:

1. The records which have similar missing-attributes are grouped in to blocks and formed as testing dataset.
2. The patient records from the Type II Diabetes database are considered as being vectors of values with nine attributes and they are represented in the same way chromosomes are. The distance between two chromosomes refers to the first eight attributes only; as the eighth attribute is a class label of the dataset.
3. The similarity between the chromosomes is measured by the fitness function. A fitness function is to define the similarity measure between two chromosomes which refers to the patient records in the dataset. Distance between the chromosomes can be used as the similarity measure.

   $$similarity(a,b) = \sum_{i=1}^{n} d(a_i, b_i)$$

   where a and b are the two chromosomes and n represents the number of attributes.

   A fitness function can be of any function derived using the genes of the chromosomes. As the genes are represented by numeric values the fitness function is defined using distance between the chromosomes. Having a missing-attribute-record as chromosome c$= (c_1, c_2, c_3 \ \ldots \ c_n)$ and a patient record from the training dataset $p = (\ p_1, p_2, p_3 \ \ldots \ p_n)$ the distance between c and p is computed by $d(c, p)$

   $$d(c,p) = \sum_{i=1}^{n} \frac{|c_i - p_i|}{b_i - a_i}$$

   where a_i and b_i represent the lower and upper bounds of the i-th attribute. The suitable fitness function can be designed in order to achieve a more efficiency.
4. In selection process, the best-believed chromosome(s) are matched by using fitness value as a criterion. For every missing-attribute-record in the training dataset the similar patient record(s) is selected using the fitness function. The missing-attribute value is then imputed with the value from the in the complete chromosome which is evaluated as the similar one. If more than one similar chromosomes are identified then the missing-attribute value is imputed with the mean value of the respective attribute(s) of the similar chromosomes.

The results are validated using the average imputation error E, of the missing attribute(s).

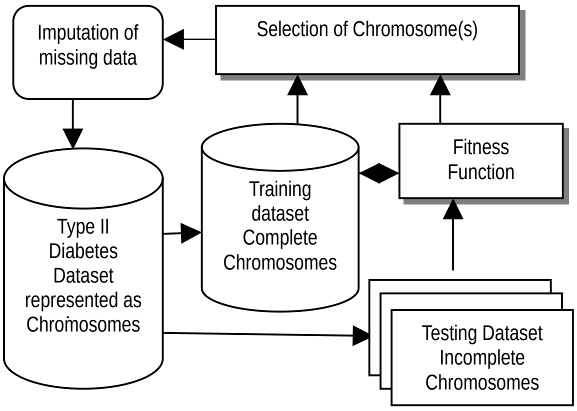
Imputation of missing data
Selection of Chromosome(s)
Fitness Function
Type II Diabetes Dataset represented as Chromosomes
Training dataset Complete Chromosomes
Testing Dataset Incomplete Chromosomes

Figure-1 The proposed Framework GCFIT_MIS_IMPUTE

A. Algorithm

```
GCFIT_MIS_IMPUTE( )
// D: Dataset
  M: Number of records
  N: Number of attributes
  b:Block(s)- testing dataset
  Attr:Attributes
  C:Chromosome
  R:Record //
  D_Train = D;
  D_Test = Ø;
 Do Form_blocks(D)
{
    For i = 1 to $2^n$-2 do
    { // Formulate $b_i$ with set of records R with $i^{th}$ attribute missing //
        $b_i$ = Ø;
      For j = 1 to m do
      {
          If Value($Attr_i$(R)) = '?' then
          $b_i = b_i$ U {R};
      }
      D_Train = D_Train - $b_i$;
      D_Test = D_Test U $b_i$;
      For i = 1 to n
      {
            SELECT_CHRO(C) = FITNESS (D_Train, b, C) ;
            IMPUTE (C, $b_i$));
      }
}
```

B. Formulation of blocks

The PIMA database is scanned to group the similar missing-attributes records into a block. The maximum number of blocks that will be generated is *(2^n - 2)* where *n* is the total number of attributes in the dataset.

Blocks –

$\mathbf{B_1}= (X,a_2,a_3,a_4,a_5,a_6,a_7,a_8),$	$\mathbf{B_2}=(a_1,X,a_3,a_4,a_5,a_6,a_7,a_8),$
$(X,a_2,a_3,a_4,a_5,a_6,a_7,a_8),$	$(a_1,X,a_3,a_4,a_5,a_6,a_7,a_8),$
$(X,a_2,a_3,a_4,a_5,a_6,a_7,a_8),$	$(a_1,X,a_3,a_4,a_5,a_6,a_7,a_8),$
…………………………	…………………………
…………………………	…………………………
…………………………	…………………………
$(X, a_2,a_3,a_4,a_5,a_6,a_7,a_8),$	$(a_1, X,a_3,a_4,a_5,a_6,a_7,a_8)$
$\mathbf{B_3} = (a_1,a_2,X,a_4,a_5,a_6,a_7,a_8),$	$\mathbf{B_4}=(a_1,X,X,a_4,a_5,a_6,a_7,a_8),$
$(a_1,a_2,X,a_4,a_5,a_6,a_7,a_8),$	$(a_1,X,X,a_4,a_5,a_6,a_7,a_8),$
$(a_1,a_2,X,a_4,a_5,a_6,a_7,a_8),$	$(a_1,X,X,a_4,a_5,a_6,a_7,a_8),$
…………………………	…………………………
…………………………	…………………………
…………………………	…………………………
$(a_1, a_2,X,a_4,a_5,a_6,a_7,a_8),$	$(a_1, X,X,a_4,a_5,a_6,a_7,a_8)$

Figure-2 Blocks of missing data

Figure-2 shows the formulation of the blocks and the missing attribute(s) is/are indicated as X. These blocks collectively form the testing dataset. The records in the blocks are passed as parameter in the algorithm to select the best chromosomes using fitness function.

C. Fitness evaluation of chromosomes

A fitness function is a particular type of objective function that quantifies the optimality of a solution (that is, a chromosome) in a genetic algorithm so that that particular chromosome may be ranked against all the other chromosomes. The fitness function should be able to distinguish between the individual chromosomes and should be able to recognize similarities too. Fitness function introduces a criterion for the selection of chromosomes. A fitness function, the evaluation function of chromosomes, is necessary to detect the similar chromosomes.

In this case, the fitness function can be defined to return a value representing the similarity with other patient records which are treated as chromosomes in the dataset. The chromosomes in the current population are evaluated by using measurement of fitness, called fitness function. Based on the fitness value, these chromosomes will be selected at a particular time as the best chromosomes which match the missing-attribute-record(s). The fitter chromosomes have a higher probability of being selected.

E. Selection of best Chromosomes and Imputation of missing data

Finally, best chromosomes are selected for every missing record in every block. The missing-attribute-value(s) in each record of the testing dataset will be the respective value of the attribute in the chromosome which is selected as best matched one, if there are more than one best chromosomes are selected then the mean value of the respective attribute(s) and the complete dataset is generated without any missing data.

V. EXPERIMENTAL RESULTS

The proposed framework is experimented using MATLAB. For each percentage of missing data 20 random simulations were conducted. The proposed algorithm is implemented and the missing data are imputed. The accuracy and the consistency of the estimated imputed values are validated using Average Imputation Error. The average imputation error (AIE) [14] is computed as given below.

$$AIE = \left(\sum_{k=1}^{m} \left(\left(\sum_{i=1}^{n} \left(| Oij - Iij | / (\max j - \min j) \right) \right) / n \right) \right) / m$$

Where n is the number of imputed values, m is the number of random simulations for each missing value, O_{ij} is the original value of the attribute j, I_{ij} is the imputed value, Max_j is the maximum value of the of the attribute j, Min_j is the is the minimum value of the of the attribute j, j is the corresponding attribute to which O_i and I_i belong.

VI. PERFORMANCE ANALYSIS

The Average Imputation Error(AIE) for all the eight attributes for various percentage of missing data is shown in Table-I in Annexure-I. The average imputation error varies significantly from attribute to attribute due to the nature of the distribution of the respective attributes of the training dataset. The overall average error value varies from a minimum of 0.133713 to a maximum of 0.1425 with the range of 0.008787 and the variation is also minimal for different of missing data.

The average imputation error varies significantly from attribute to attribute due to the nature of the distribution of the respective attributes of the dataset. The overall average error value varies from a minimum of 0.133713 to a maximum of 0.1425 with the range of 0.008787 and the variation is also minimal for different of missing data.

The impact of the size of missing data on computation of Average Imputation Error(AIE) is shown Figure-3 in Annexure-I. It is observed from the Figure-3 that there is a steady increase in Average Imputation Error(AIE) when there is an increase in the size of missing data, but it also shows stability then and there in its performance. The stability in the performance is due to the strength of the fitness function in selecting the best chromosomes from the training dataset which matches the missing record. If the proposed framework GCFIT_MIS_IMPUTE, selects only one best chromosome then there an increase in the Average Imputation Error. But, if more than one best chromosomes are selected by the framework, the mean value of the respective missing attribute of the selected chromosomes is used as the most probable imputed value in the missing record. The imputation process fully depends on the performance of the fitness function in selecting the best chromosome(s). Wherever there is an increase in the Average Imputation Error(AIE) it indicates that the imputation of missing data will be less consistent with the observed values.

The performance of the proposed approach GCFIT_MIS_IMPUTE is compared with other imputation methods namely 10-NN, NORM, EMImpute Columns, LSImpute_Rows, Mean Imputation and is shown in Table-II in the Annexure-I. One-third of the dataset is considered for performance analysis. The performance up to 35% of missing data is taken up for comparison. The performance of the framework, GCFIT_MIS_IMPUTE is compared with the other missing data imputation methods and it is shown in Table-II and Figure-4 in Annexure-I.

It is observed from the table-II and Figure-4 in the Annexure that NORM method produces highest error rate with least accurate estimation results and EMImpute_columns produces lower error rate. The proposed framework GCFIT_MIS_IMPUTE performs next to NORM with the average imputation error ranging from 13.4 to 13.8. Though, the Mean_Imputation, 10-NN, EMImpute and LSImpute_Rows methods show better performance in terms of average imputation error than GCFIT_MIS_IMPUTE, the error value increases with the size of the missing data. But, It is observed from the Table-II and Figure-4 that GCFIT_MIS_IMPUTE is able to achieve stable error rate for span during the evaluation. The Average Imputation Error(AIE) value increases when there is an increase in the size of missing data, which is due to non-selection of complete chromosomes for the process of imputation. As the selection of best chromosomes for the missing record is fully depends on the effectiveness of the fitness function the performance of the imputation process is depends on the effective design of the fitness function.

VII. CONCLUSION

Data mining in medical applications have several distinguishing features [38]. In medical data mining, though the data sets are small but every feature is highly sensitive and impacts significantly in decision making while diagnosing the disease. The missing data occurs in the clinical database due to a number of situations: experiments can be costly due to the involvement of the personnel and use of expensive instrumentation, and due to the potential discomfort of the patients involved. The quality of data may be further affected by several sources of uncertainty, like those from measurement errors or missing data or from errors in coding the information buried in textual reports. In medical data mining where accuracy is crucial in making decisions in the pathological analysis efficient methods to prevent inaccuracies or incomplete data is required.

The framework, GCFIT_MIS_IMPUTE imputes the missing data by selecting the best chromosomes from the training dataset using the fitness function. Genetic chromo dynamic framework ensures the applicability of the method to any dataset irrespective of the number of attributes as it treats the records as chromosomes. The framework exhibits relatively better performance to NORM as it produces stable and less Average Imputation Error(AIE). But, the performance of the framework is solely depends on the effectiveness of the fitness function. Hence, the future work may be extended to design the fitness function which suits for the application with effectiveness in selecting the best chromosomes for the imputation process.

ANNEXURE - I

TABLE-I

AVERAGE IMPUTATION ERROR FOR THE PIMA INDIAN TYPE II DIABETES DATASET

% of Missing Data	Number of times pregnant	Glucose tolerance test	Diastolic blood pressure	Triceps skin fold thickness	2 -hour serum insulin	Body mass index	Diabetes pedigree function	Age	Average
5%	0.164257	0.14065	0.1331	0.1223	0.0914	0.10793	0.14348	0.170362	0.134185
10%	0.159502	0.143775	0.13102	0.12348	0.08968	0.10986	0.14084	0.171547	0.133713
15%	0.158797	0.14248	0.135515	0.12403	0.09348	0.111775	0.14149	0.17078	0.134793
20%	0.15713	0.14275	0.13201	0.1236	0.10138	0.110105	0.14114	0.17078	0.134862
25%	0.16029	0.14461	0.133487	0.124304	0.10291	0.10768	0.14099	0.172641	0.135864
30%	0.16779	0.14677	0.13563	0.1244	0.10446	0.107215	0.14153	0.172641	0.137555
35%	0.16836	0.14802	0.13601	0.1242	0.102115	0.10651	0.14435	0.173849	0.137927
40%	0.1699	0.15127	0.13407	0.12519	0.10038	0.10596	0.14443	0.177282	0.13856
45%	0.1733	0.14892	0.136445	0.12293	0.106065	0.10565	0.14378	0.175775	0.139108
50%	0.17651	0.14973	0.13574	0.1248	0.10464	0.106215	0.143135	0.17755	0.13979
55%	0.1804	0.149755	0.13376	0.123	0.100319	0.10661	0.146955	0.177655	0.139807
60%	0.17964	0.15274	0.130215	0.12169	0.103165	0.107265	0.14343	0.17934	0.139686
65%	0.18368	0.15044	0.13005	0.12205	0.10315	0.109365	0.14159	0.17878	0.139888
70%	0.186266	0.152753	0.13429	0.12792	0.10314	0.110105	0.143845	0.17951	0.142229
75%	0.18757	0.153731	0.13699	0.12642	0.10317	0.11031	0.140675	0.181135	0.1425
80%	0.19241	0.15081	0.134345	0.12605	0.103205	0.110215	0.13933	0.18012	0.142061

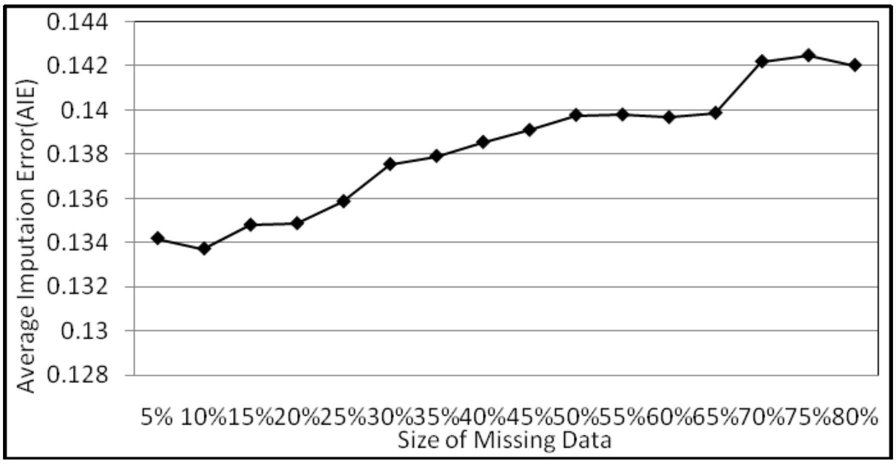

Figure-3 Performance of GCFIT_MIS_IMPUTE

TABLE-II

COMPARATIVE PERFORMACE OF VARIOUS MISSING DATA IMPUTATION METHODS

Method	Percentage of Missing Data						
	5%	10%	15%	20%	25%	30%	35%
10-NN	10.2±10.0	10.4±10.2	10.4±10.0	10.7±10.1	10.9±10.2	11.3±10.3	11.7±10.5
NORM	16.0±13.4	15.9±13.6	15.7±13.5	16.1±13.6	16.2±13.6	16.3±13.7	16.4±13.9
EMImpute_Columns	8.5±23.8	8.5±23.6	8.6±23.4	8.5±23.3	8.5±23.1	8.7±22.9	8.1±22.3
LSImpute_Rows	10.9±23.9	11.6±23.9	11.6±24.0	11.8±23.9	11.6±23.9	11.8±23.9	12.2±23.3
Mean Imputation	12.5±10.5	12.3±10.5	12.4±10.4	12.4±10.5	12.4±10.5	12.4±10.4	12.3±10.3
GCFIT_MIS_IMPUTE	13.4±24.7	13.4±24.9	13.5±25.0	13.5±25.7	13.6±25.7	13.8±25.9	13.8±25.9

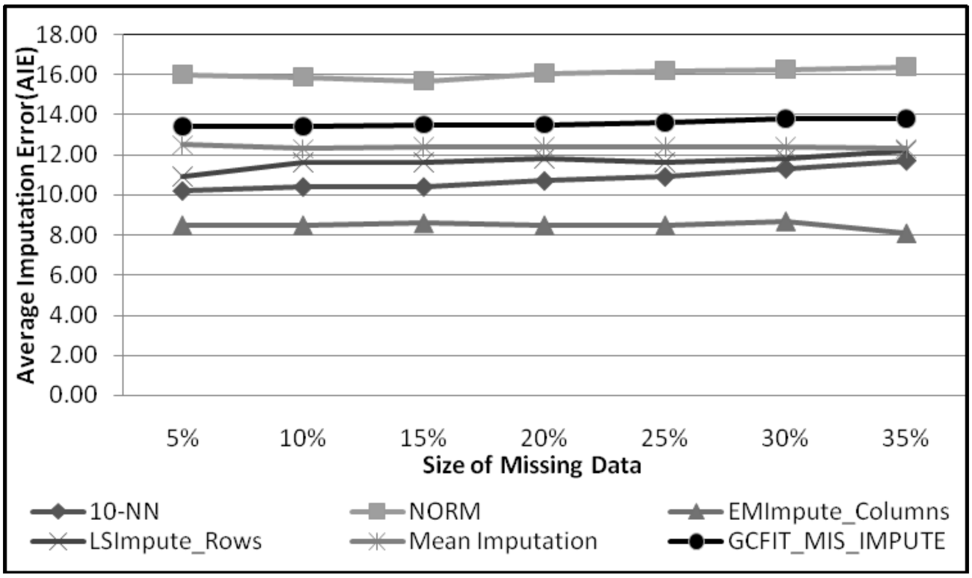

Figure-4 Performance of different mission data imputation methods

REFERENCES

[1] Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3—23.

[2] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med 001;22:215—31.

[3] Ceglowski, A., Churilov, L. and Wassertheil, J. (2005), “Knowledge discovery through mining emergency department data”, proceedings of the 38th Hawaii International Conference on System Sciences.

[4] Isken, M. and Rajagopalan, B. (2002), “Data mining to support simulation modeling of patient flow in hospitals”, Journal of Medical Systems, Vol. 26, pp. 179-97.

[5] Ridley, S., Jones, S., Shahini, A., Brampton, W., Nielsen, M. and Rowan, K. (1998), “Classification trees: a possible method for ISO-resource grouping in intensive care”, Anaesthesia, Vol. 53, pp. 833-40.

[6] E.W. K. CynthiaM. Farquhar and J. R. Slutsky. *Clinicians’ attitudes to clinical practice guidelines. Medical Journal of Australia,* Aust 2002; 177: 5 02-506.

[7] Krzysztof, C. and Kurgan, L. (2002), “Trends in data mining and knowledge discovery”, in Pal, N.R., Jain, L.C. and Teoderesku, N. (Eds), Knowledge Discovery in Advanced Information Systems, Springer, Berlin.

[8] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998), Discovering Data Mining: From Concepts to Implementation, Prentice-Hall, Upper Saddle River, NJ.

[9] Liu, H. and Motoda, H. (1998), Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic, Boston, MA.

[10] Han, J. and Kamber, M. (2000), Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, San Matel, CA.

[11] Kantardzic, M. (2003), Data Mining Concepts, Models, Methods and Algorithms, Wiley-IEEE Computer Society Press, New York, NY.

[12] Marvin L.Brown and John F.Kros Data mining and the impact of missing data, Industrial Management & Data systems 103/8 [2003] 611-621.

[13] C.I.Blake, C.J.Merz, UCI Repository of machine learning databases, website: http://www.ics.uci.edu/~mlearn/mlrepository.html , 1998.

[14] Mariso Giardina, Yongyang Huo, Francisco Azuaje, Paul McCullagh, Roy Harper, “A Missing Data Estimation Analysis in Type II Diabetes Databases”, Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS’05) 1063-7125/05 2005 IEEE.

[15] H. Trond, D. Bjarte , J. Inge, “LSimpute: accurate estimation of missing values in microarray data with least squares methods”, Nucleic Acids Research, 32(3), Feb 2004.

[16] J.L. Schafer, NORM: multiple Imputations of incomplete multivariate data under a normal model, version 2.03, software for Windows 95, 98, NT, Website: http;//www.stat.psu.edu/~jls/misoftwa.html, 1999

[17] M. Pfaff, K. Weller, D. Woetzel, R. Guthke, K. Schroeder, G. Stein, R. Pohlmeier, and J. Vienken, “Prediction of cardiovascular risk in hemodialysis patients by data mining”, Methods

of information in medicine, 2004, vol. 43, pp.106-13.

[18] P. McEwan, J.E. Wiliam, J.D. Griffiths, A. Bagust, J.R. Peters, and P. Hopkinson, "Evaluating the performance of the framingham risk equations in a population with diabetes", Diabetic Medicine, vol. 4, 2004, pp.318-326.

[19] K. Gu, C. Cowie, and M. Harris. "Mortality in adults with and without diabetes in a national cohort of the US population, 1971-1993",DiabetesCare,vol.21,1998,pp.1138-1145.

[20] J.W. Grzymala-Busse, M. Hu, A comparison of several approaches to missing attribute values in data mining, in: Proceedings of the 2nd International Conference on Rough sets and Current Trends in Computing 2000 (RSCTC'2000),Banff, Canada, 2000, pp. 340–347.

[21] G. Batista, M. Monard, An analysis of four missing data treatment methods for supervised learning, Applied Artificial Intelligence 17 (5/6) (2003) 519–533.

[22] K. Chan, T.W. Lee, T.J. Sejnowski, Variational Bayesian learning of ICA with missing data, Neural Comput. 15 (8) (2003) 1991–2011.

[23] Moth'd Belal, Al-Daoud, "A New Algorithm for Cluster Initialization", Transactions on Engineering, Computing and Technology v4 February 2005 ISSN 1305-5313

[24] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman and Hall, 1984.

[25] J.H. Friedman, R. Kohavi, Y. Yun, Lazy decision trees, in: Proceedings of 13th AAAI and 8th IAAI, 1996, pp. 717–724.

[26] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, Inc., USA, 1986.

[27] P. Clark, T. Niblett, The CN2 induction algorithm, Machine Learning 3 (4) (1989) 261–283.

[28] M. Hu, S.M. Salvucci, M.P. Cohen, Evaluation of some popular imputation algorithms, in: The Survey Research Methods Section of the ASA, 1998, pp. 308–313

[29] J.-F. Beaumont, On regression imputation in the presence of non-ignorable non-response, in: Proceedings of the Survey Research 570 Methods Section, ASA, 2000, pp. 580–585.

[30] L. Coppola, M. Di Zio, O. Luzi, A. Ponti, M. Scanu, Bayesian networks for imputation in official statistics: a case study, in: 575 DataClean Conference, 2000, pp. 30–31.

[31] K. Lakshminarayan, S.A. Harp, T. Samad, Imputation of missing data in industrial databases, Applied Intelligence 11 (3) (1999) 259– 601 275.

[32] OT Abdala, M Saeed, "Estimation of Missing Values in Clinical Laboratory Measurements of ICU Patients Using a Weighted K-Nearest Neighbors Algorithm", Computers in Cardiology 2004; 31:693-696**,** IEEE 2004

[33] B.L. Ford, Incomplete data in sample surveys, An Overview of Hot-deck Procedures, Academic Press, 1983.

[34]] A. Farhangfar, L. Kurgan, W. Pedrycz, A novel framework for imputation of missing values in databases, IEEE Transactions on Systems, Man, and Cybernetics.

[35] Barnard J and Meng X (1999) " Applications of multiple imputation in medical studies:from AIDES to NHANES", Statistical Methods in Medical Research, Vol. 8,pp 17-36.

[36] Van Buren.S, Boshuizen.H and Knook.D (1999) "Multiple imputation of missing blood pressure covariates in survival analysis", Statistics in Medicine, Vol.18, pp 681-94

[37] O. Troyanskaya, M. Cantor, G.Sherlock, P.Bronw, T. Hastie, R. Tibshirani, D. Botstein, R. Altman, "Missing value estimation methods for DNA Microarrays", Bioinformatics, 17(6), 2001, pp.520-525.

[38] K.J. Cios, G.W. Moore, Uniqueness of medical data mining, Artif. Intell. Med. 26 (2002) 1–24.