A New Effective Information Decomposition Approach for Missing Data Recovery

Shigang Liu[1], and Honghua Dai[1]
[1]School of Information Technology, Deakin University 221, Burwood Highway, VIC 3125, Australia

Abstract - *It is well recognized that missing data could cause severe problem in data mining. Due to its importance lots of work has been done in the past. Several algorithms [5-8] are proposed for missing data recovery. This paper presents a new 1-dimensional linear information decomposition (1-DLID) approach which is easier for use in missing data recovery. In this article, we study one particular problem, in which 1-dimensional data set is given and certain percentage of data are missing without any other additional information. Then the proposed 1-DLID method is used for creating the complete data set from both the generated data set and real-world data set. Comparatively, our experiments showed that the proposed method is reliable and can be used for the recovery of data set with missing values. The advantages of the proposed method are: 1) Will not change the distribution of the data set. 2) Easy to use for 1-dimensional dataset. 3) Have a higher accuracy, especially there is 10%~30% data missing. 4) No need to provide the historical data set.*

Keywords: Information decomposition, Missing data recovery, 1-dimensional data

1 Introduction

In recent decades, missing data recovery have been broadly studied and applied in various domains in order to solve many complicated and important real-world problems, such as pattern recognition, natural language processing, medical diagnosis, and so on[1,2], in the hope of improving performance. Meanwhile missing values recovery imputation is an existing yet challenging problem in both machine leaning and data mining [3]. On the other hand, Missing values in real-world data cause severe problem for the learning and knowledge discovery. In most cases, missing data problem is caused by data logging procedure and systems. Let's take a manufacturing line for example, it is impossible to record all the line variables of all the products at any time. That is to say, variables which are recorded are only certain kind of products, which can be regarded as incomplete measurement data values. In addition, the topic of missing data has attracted considerable attention in the last decade, as evidenced by several recent trends. First, many graduating PhDs in statistics and computer science are now claiming "missing data" as an area of research. Second, it has become difficult to publish empirical work in sociology without discussion of how missing data was handled. Thirdly, several methods for handling missing data has sprouted-up over the last few years, which will be discussed later. Missing data is important to consider, because they may lead to substantial biases in analyses [4] and in sometimes result in incorrect decision making. On the other hand, missing data could be harmless except reducing statistical power.

The approach discussed in this paper is good at processing data set with data missing at random (which is called MAR) and helpful for analyzing the incomplete data set. Before our approach is presented, we would like to discuss the identification scenarios for missing values pointed out by Little and Rubin (1987) [5]. Based on the values of attributes and the missingness of attributes, the categories of missing data include missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). If data are MAR or MCAR, they can also be referred to as "ignorable" data while those MNAR are "non-ignorable" [6]. There are various methods that have been proposed to deal with missing data with each of these methods premised on a specific missing data mechanism [7-8], which will be discussed in the following section.

This paper is structured as follows: Section 2 is a brief overview of previous methods. Section 3 describes terminologies. That is to say we give some general knowledge used in the algorithm. Section 4 is devoted to the introduction of our 1-dimensional linear information decomposition approach. Section 5 presents experiment results explanation based on generated data set. The real-world values experiments are organized in section6. Section 7 concludes this paper and section 8 describes our future work.

2 Current techniques and existing problems

There are many methods that used in missing data recovery [7-8]. However every method has its own problems. For example, filling manually is very time consuming when the missing data set is very large, it is impossible to make good use of this approach; MI algorithm is flexible and time expensive. Another drawback is that this method is adequate for statistics better than data mining. Moreover, it is widely used in multivariate normal data. Last but not least is that some distribution for the stochasticity must be assumed, which can be problematic as well [7] etc.. In this paper, regarding to 1-Dimensional data set without any historical dataset

provided, we summarize the most commonly used methods as follows:

1) Listwise deletion[7-8]. By far most times researchers would like to simply omit those instances with missing attribute-values and run the analyses only on the complete instances.

The major problems of this method are that when parametric model based on the attribute-values are not MAR this approach does not work well. Moreover, this method may lead to a large amount of data being thrown away, miss some important information.

2) Filling Manually[7-8]. This method based on the experience of the experts and used in some of statistical area with small missing data set.

The major problem of this method is that it is time consuming particularly when the missing data set is very large. It is impossible to make good use of this approach.

3) Mean/Mode Imputation[7-8]. It means Replacing missing values with the sample mean. In fact, this method is simple and save time when the missing data is numerical rather than non-numerical.

The major problems of this approach are it will make the distribution more peaked around the mean and assumes all the missing data should be MCAR.

4) The Expectation Maximization(EM) algorithm. The EM algorithm is an elaborate technique for incomplete data or data set with missing values. The EM algorithm[9-10] is an approach used for finding the maximum-likelihood estimate of the parameters based on the assumption of the distribution for a given incomplete data set. Usually the EM algorithm is used for the following two situations, first there are indeed missing values, because of limitation of observation process. The second situation is when optimizing the likelihood of a function, it is analytically intractable while the likelihood function could be simplified by assuming the existence values for additional, however, missing or latent parameters. There are two steps in EM algorithm, E-step (expectation step) to compute the expectation of the expected value of the complete data log-likelihood with respect to the unknown data given the observed data and current parameter estimates. The second step (M-step) is to maximize the expectation which was computed in the E-step. That is to say, we find out the new parameter $\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^i)$. These two steps are repeated until $|\theta^{(i+1)} - \theta^i| \le error$, the error is the values we fixed before the two steps are repeated.

For this paper, during the experiments the initial values of the missing data are produced by computer in random. Then we use it to finish the E-step, after that we can maximize the likelihood function and get the θ, then with the help of θ we can renew the missing data, and then go to M-step.

Problems: Firstly, it is time consuming towards multi-dimensional dataset. Secondly, the algorithm doesn't produce standard errors for the parameters. Thirdly, it may converge to a local maximum of the observed data likelihood function, and this depends on starting values.

Overall, every methods have its own problems, that is to say it is hard to find an algorithm that suitable for all kind of problems. For example, it is said handling missing data by eliminating cases with missing data ("listwise deletion" or "complete case analysis") will lead to the predicted results away from the reality when the remaining data cannot be representative of the whole data set. Moreover, the Expectation Maximization (EM) algorithm is also one method that is used for data mining, however, it can be regarded as an auxiliary method such as bootstrapping when obtaining standard errors.

The major contribution of this paper is to propose a new method which is 1-dimensional linear information decomposition (1-DLID) approach used for missing data recovery. The 1-DLID method is useful in two aspects, one is that it can be used for the recovery of missing data; on the other hand, it can create or generate data for the incomplete data. Compared with other algorithms, 1-DLID has its own advantages. For example, unlike EM algorithm, 1-DLID approach do not need to set the latent variables even more, we do not have to know any kind of probability distribution. Therefore, 1-DLID approach is easy to use and can be used in any kind of one dimension numerical data set with missing values.

3 Basic terminologies used in missing data recovery

3.1 Information Distribution [11]

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a sample observed from an experiment, and $U = \{u_1, u_2, \cdots, u_m\}$ be the discrete universe of X.

A mapping from $X \times U$ to [0,1],

$\mu : X \times U \to [0,1]$,

$(x,u) \to \mu(x,u)$

is called an information distribution of X on U, if $\mu(x,u)$ has the following properties:

1)Reflexive. $\forall x \in X$, if $\exists u \in U$, such that $x = u$, then $\mu(x,u) = 1$.

2)Decreasing. For $x \in X$, $\forall u', u'' \in U$,

if $\|u' - x\| \le \|u'' - x\|$, then $\mu(x,u') \ge \mu(x,u'')$.

3)Conserved. That is to say $\sum_{j=1}^{m} \mu(x_i, u_j) = 1$,

$i = 1, 2, \cdots, n$.

3.2 1-Dimension Linear Information Distribution[11]

Let $X = \{x_i \mid i = 1, 2, \cdots, n\}$ be a given sample, R is the universe of discourse of X, and $U = \{u_1, u_2, \cdots, u_m\}$ is the discrete universe of X, where $u_j - u_{j-1} \equiv h, j = 2, 3, \cdots, m$. For $x_i \in X$, and $u_j \in U$, the following formula is called 1-dimensional linear-information distribution:

$$\mu(x_i,u_j)=\begin{cases}1-|x_i-u_j|/h & if\ |x_i-u_j|\le h\\ 0 & if\ |x_i-u_j|>h\end{cases} \quad (1)$$

Where h is called step length and μ is called linear distribution.

Obviously, μ satisfies all properties of an information distribution function.

For example, let $X=\{5,6\}$. What we want to do is to calculate their relative frequency between 3.4 and 8.2 in order to get the soft-histogram [11].That is $X\subset[3.4,8.6]$. Assume we would like to have three intervals between 3.4 and 8.2. That is $h=\frac{8.2-3.4}{3}=1.6,$ therefore we can get three intervals $[3.4,5)\cup[5,6.6)\cup[6.6,8.2).$

We chose u_i as the center of each intervals. Accordingly $u_1=(3.4+5)/2=4.2$, $u_2=(5+6.6)/2=5.8$ and $u_3=7.4$, Thus $U=\{4.2,5.8,7.4\}$. Given $x_1=5,x_2=6$, we can get:

$\mu(x_1,u_1)=1-|5-4.2|/1.6=0.5$, $\mu(x_1,u_2)=1-|5-5.8|/1.6=0.5$, $\mu(x_1,u_3)=0$, $\mu(x_2,u_1)=0$, $\mu(x_2,u_2)=1-|6-5.8|/1.6=0.875$ $\mu(x_2,u_2)=1-|6-7.4|/1.6=0.125$, $\mu(x_2,u_3)=0$.

3.3 1-Dimension Linear Information decomposition

Let $X=\{x_i\mid i=1,2,\cdots,n\}$ be a given sample, R is the universe of discourse of X, $A=[a,b]$, where $a=min\{x_i\mid i=1,2,\cdots,n\}$ and $b=max\{x_i\mid i=1,2,\cdots,n\}$. t is the settled number of the intervals that $A=[a,b]$ is being divided, usually t is the number of missing values. That is the step length $h=(b-a)/t$, $A=\bigcup_{j=1}^{t}A_j$ and $A_j=[a+(j-1)*h,a+j*h]$. $U=\{u_1,u_2,\cdots,u_t\}$ is the discrete universe of R where $u_j-u_{j-1}\equiv h, j=2,3,\cdots,t$ and $u_j=(a+(j-1)*h+a+j*h)/2$, that is to say u_j is the center of A_j. For $\mu(x_i,u_j)$ is obtained from formula (1), $x_i\in X$, and $u_j\in U$, m_{ij} obtained from formula (2) is called 1-dimensional linear information decomposition from x_i to A_j.

$$m_{ij}=\mu(x_i,u_j)*x_i \quad (2)$$

Where h is called step length and μ is called linear distribution.

For example, $X=\{3.4,5,6,8.2\},t=3$, then we get $x_1=3.4,x_2=5,x_3=6,x_4=8.2$, $u_1=4.2$, $u_2=5.8,u_3=7.4$ $h=1.6$ and $A_1=[3.4,5),A_2=[5,6.6),A_3=[6.6,8.2).$

Therefore, 1-dimensional linear information decomposition from x_2 to A_1 is:

$m_{21}=\mu(x_2,u_1)*x_2=(1-|5-4.2|/1.6)*5=2.5$

Similarly, we can get:

$m_{22}=\mu(x_2,u_2)*x_2=(1-|5-5.8|/1.6)*5=2.5$

$m_{31}=\mu(x_3,u_1)*x_3=0$

$m_{32}=\mu(x_3,u_2)*x_3=(1-|6-5.8|/1.6)*6=5.25$ etc..

4 1-DLID approach for missing data recovery

In the following discussion, the detailed steps about how 1-dimensional linear information decomposition method used in missing data recovery will be introduced. First of all, we would like to note that this paper focus on numerical missing data.

Let $X=\{x_i\mid i=1,2,\cdots,n\}$ be a missing data (incomplete data) set; the number of missing values is t, the missed values denoted as $\{m_k\mid k=1,2,\cdots,t\}$.

Let $a=min\{x_i\mid i=1,2,\cdots,n\}; b=max\{x_i\mid i=1,2,\cdots,n\}$. Then we get an interval $[a,b]$. Bear that if we let $c=a\pm0.5$ or $c=a\pm1$ and $d=b\pm0.5$ or $d=b\pm1$, we can get another interval that is $[c,d]$ and will help us get another recovery missing data, thus we can choose the average values of all the missing data.

Let $h=\frac{b-a}{t}$ $A_i=[a+(i-1)*h,a+i*h),i=1,2,\cdots,t.$

And we get $(a+(i-1)*h+a+i*h)/2$, then we find out the number of $x_i\in A_i$ and we get $\{x_i\}=A_i\cap X$.

To clarify, we denote $Y_i=\{y_l\mid l=1,2,\cdots,s\}=\{x_i\}=A_i\cap X$, then we get $\sum_{l=1}^{s}y_l/s$, $i=1,2,\cdots,t.$

In the 1-dimensional linear information decomposition approach, we choose the linear distribution as:

$$\mu(y_j,u_i)=\begin{cases}1-|y_j-u_i|/h & if\ |y_j-u_i|\le h\\ 0 & if\ |y_j-u_i|>h\end{cases}$$

Then we can calculate the following values:

$f_{A_i}(\tilde{y}_i,u_i)$, $f_{A_i}(\tilde{y}_{i+1},u_i)$ and $f_{A_i}(\tilde{y}_{i-1},u_i)$, finally we get the i^{th} missing data value, which is $m_i=(f_{A_i}(\tilde{y}_i,u_i)+f_{A_i}(\tilde{y}_{i+1},u_i)+f_{A_i}(\tilde{y}_{i-1},u_i))/3$. If one of them is 0, for example, $f_{A_i}(\tilde{y}_{i+1},u_i)=0$, we get $m_i=(f_{A_i}(\tilde{y}_i,u_i)+f_{A_i}(\tilde{y}_{i-1},u_i))/2$, once two of them are 0, for example, $f_{A_i}(\tilde{y}_{i+1},u_i)=f_{A_i}(\tilde{y}_{i-1},u_i)=0$, then $m_i=f_{A_i}(\tilde{y}_i,u_i)$.

The following steps are used to generate the missing values for the data set with missing values:

1. Given the incomplete data set X and the number of missing data values.
2. Compute A_i and u_i.

3. Compute $\tilde{y}_i, i = 1, 2, \cdots, t.$ Where t is the number of missing values.

4. For each i compute $f_{A_i}(\tilde{y}_i, u_i)$, $f_{A_i}(\tilde{y}_{i+1}, u_i)$ and $f_{A_i}(\tilde{y}_{i-1}, u_i)$.

5. Compute m_i , if $f_{A_i}(\tilde{y}_i, u_i) = f_{A_i}(\tilde{y}_{i+1}, u_i)$ $= f_{A_i}(\tilde{y}_{i-1}, u_i) = 0$ then $m_i = 0$, then $m_i = mean(\mathrm{X})$. Otherwise, $m_i = (f_{A_i}(\tilde{y}_i, u_i) + f_{A_i}(\tilde{y}_{i+1}, u_i) + f_{A_i}(\tilde{y}_{i-1}, u_i))/3$.

If one of them is 0, for example, $f_{A_i}(\tilde{y}_{i+1}, u_i) = 0$, we get $m_i = (f_{A_i}(\tilde{y}_i, u_i) + f_{A_i}(\tilde{y}_{i-1}, u_i))/2$, once two of them are 0, for example, $f_{A_i}(\tilde{y}_{i+1}, u_i) = f_{A_i}(\tilde{y}_{i-1}, u_i) = 0$, then $m_i = f_{A_i}(\tilde{y}_i, u_i)$.

The 1-dimensional linear information decomposition approach is easy to use; because it doesn't need any restrictions as long as we know the data set with missing values or incomplete data set and the number of missing values. In the following sections, the experiments and explanation will be discussed.

5 Experiments and results analyze with generated data set

5.1 Experimental Data

In order to make sure whether the proposed algorithm works well or not, we generate data from the Gaussian distribution and Gamma distribution. That is to say the generated data set $\{x_i\} \sim N(\mu, \sigma^2)$ or $\{x_{ij}\} \sim N(\mu, \sigma^2)$ and $\{x_i\} \sim \Gamma(\alpha, \beta)$ or $\{x_{ij}\} \sim \Gamma(\alpha, \beta)$ which can be regard as a matrix, either $1 \times n$ or $p \times n$. Then we get rid of some of the data values randomly with the help of computer (matlab). That is to say, we create the data set with missing values and ready for using in the experiments. With every data set with missing values we used increasing levels of 'missingness': 5%, 10%, 20%, 30% and 50%.

The reason why we choose Gaussian distribution and Gamma distribution is that because Gaussian distribution is widely used in research area, which can be regarded the data values is distribute averagely beside the means of the data set. However, not every data set that with missing values can follow Gaussian distribution in daily life. In order to show that 1-dimensional linear information decomposition method is good at processing any kind of data set rather than Gaussian distribution. We used data from Gamma distribution for experiment, because the property of data sets is totally different from each other. We use the following data sets:

Table 1: Information about the datasets used in this paper

Dataset Number	Instances	Data from $X \sim N(10, 5^2)$ Missing values (%)	Data from $X \sim \Gamma(10, 5)$ Missing values (%)
1	100	5%	
2	100	10%	
3	100	20%	
4	100	30%	
5	100	50%	
6	1000	5%	
7	1000	10%	
8	1000	20%	
9	1000	30%	
10	1000	50%	
11	1000		5%
12	1000		10%
13	1000		20%
14	1000		30%
15	1000		50%

In the following table the results arrived on a Window 8 laptop equipped with Core i7-2600 CPU at 3.40 GHz and 8.00 GB RAM is presented. And the matlab 7.0 is use for evaluation.

5.2 Experimental Strategy

Because 1-DLID method only require the condition of the incomplete data set and the number of missing values without any more information such as probability distribution or the incomplete data should meet the need of Bayesian estimation, mean/mode imputation method and listwise deletion method can be used in the data sets. However, in order to show 1-DLID approach works well and can achieve good results most of times. We would like to do experiment with one of most popular used algorithm which is EM algorithm. Overall, the experiments are based on four approaches that is 1-DLID method, Mean imputation method[6-9], listwise deletion method[6-9] and EM algorithm[6-9].

To ensure this is not the case, we performed the following for each experimental run:

1). Generate a data set with matlab 7.0 and save it into a file. We would like to generate a $1 \times n$ data set. And we denote this data set as F_t.

2). We get rid a certain percentage of the data from F_t, and we get the missing data set X , which we mentioned before.

3). Then we come to the proposed steps in section 3.

5.3 Evaluation Criteria

Before the results were presented, we would like to give a brief explanation of the errors of all the parameters. First of all the predicted parameters are calculated by the complete data, which is the total of data set with missing values and the recovered data. Then we compare the predicted parameters and the original ones and give the following definition:

μ error is defined as $|\tilde{\mu} - 10|/10$, where $\tilde{\mu}$ is the predicted parameter.

σ error is defined as $|\tilde{\sigma} - 5|/5$, where $\tilde{\sigma}$ is the predicted parameter.

α error is defined as $|\tilde{\alpha}-10|/10$, where $\tilde{\alpha}$ is the predicted parameter.

β error is defined as $|\tilde{\beta}-5|/5$, where $\tilde{\beta}$ is the predicted parameter.

5.4 Results

After choosing different intervals $[a,b]$ or $[c,d]$ etc., choosing of a good interval is very important, not only it helps to achieve a good results but also save time. Most times the intervals are chosen as $a=min(X)\pm 0.5$, $b=max(X)\pm 0.5$. However, sometimes are chosen as $a=min(X)\pm 1$, $b=max(X)\pm 1$ or others. Because it is difficult to choose a perfect interval, we will discuss in our future papers. In case we can get better results, we chosen three different intervals and use the average results, which can be regarded better than only choose one interval. After the experiments, we got the following results:

Table 2: Comparison of the results of dataset $X \sim N(10,5^2)$

	μ					σ				
Dataset number	Original	Proposed method	Deletion method	Mean method	EM algorithm	Original	Proposed method	Deletion method	Mean method	EM algorithm
1	10	9.9718	9.9292	9.9292	9.8325	5	5.0851	5.1300	4.9987	5.1056
2	10	10.1883	10.2293	10.2293	10.3433	5	5.1695	5.2450	4.9730	5.1308
3	10	10.1675	10.2267	10.2267	10.2139	5	5.0831	5.1697	4.8706	5.1101
4	10	10.0391	9.9273	9.9273	9.8525	5	4.9853	4.7080	4.7080	5.1760
5	10	10.5420	10.5867	10.5867	10.9646	5	5.0015	5.0551	4.6068	5.3577
6	10	9.8100	9.7690	9.7690	9.8215	5	4.7211	4.6920	4.5731	4.9697
7	10	9.8499	9.8372	9.8372	10.2400	5	4.7624	4.7619	4.5173	5.5702
8	10	9.6138	9.7825	9.7825	10.1881	5	4.7439	4.6884	4.1929	5.2286
9	10	9.4849	9.7904	9.7904	10.6379	5	4.3620	4.6613	3.8991	5.8313
10	10	9.3260	9.8733	9.8733	11.7081	5	4.1955	4.7951	3.3889	6.8376

Table 3: Comparison of the results of the dataset $X \sim \Gamma(10,5)$

	α					β				
Dataset number	Original	Proposed method	Deletion method	Mean method	EM algorithm	Original	Proposed method	Deletion method	Mean method	EM algorithm
11	10	9.4530	9.6000	10.0968	9.6469	5	5.4397	5.3265	5.0644	5.2971
12	10	9.9587	9.8964	10.9781	9.7294	5	5.1642	5.1517	4.6619	5.2484
13	10	10.0140	9.9374	13.3813	9.9384	5	5.1065	5.1647	4.1452	5.1584
14	10	10.1487	9.3651	13.3096	9.5682	5	5.0231	5.5433	3.9005	5.3870
15	10	9.7572	9.4903	18.8185	10.3844	5	4.9736	5.3379	2.6919	4.8208

In order to make the results clear to understand, we have presented them in the following pictures:

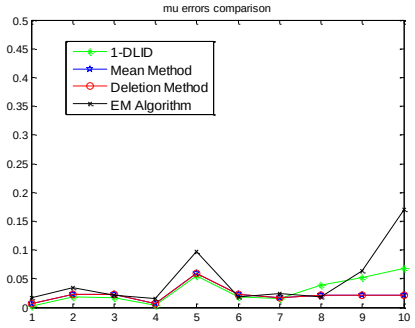

Fig. 1: μ error comparison of each method from data 1 to data 10 of $X \sim N(10,5^2)$

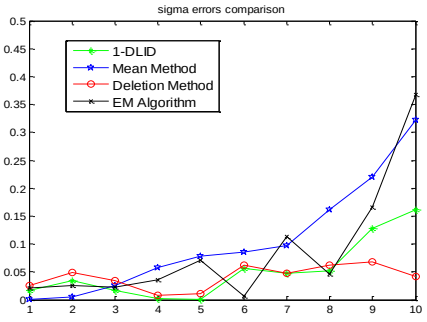

Fig. 2: σ error comparison of each method from data 1 to data 10 of $X \sim N(10,5^2)$

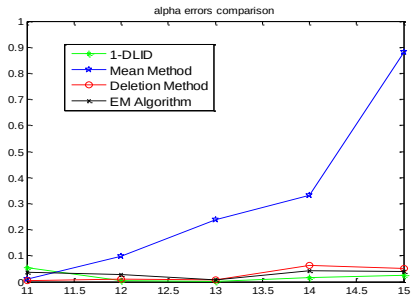

Fig. 3: α error comparison from data 10 to data 15 of $X \sim \Gamma(10,5)$

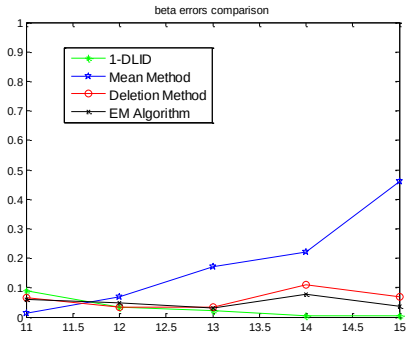

Fig. 4: β error comparison of each method from data 10 to data 15 of $X \sim \Gamma(10,5)$

As can be seen from figure 1, 1-DLID approach works well especially for dataset from 1 to 7, even when there are 50% of data lost, the μ error is still within 10% which is smaller than EM method. For the reason that delete method and mean imputation method achieved better results for μ from dataset 8 to dataset 10, we think the normal distribution dataset made them works well towards this. However, the results predicted by 1-DLID method are acceptable. Moreover, the σ errors showed in figure 2 also proved that 1-DLID approach perform better than the other three methods, especially from dataset 1 to dataset 8. While it is illustrated that the 1-DLID method also achieves a better results compared with EM method and mean imputation method. Again because normal distribution data set that delete method works well even though 50% of data values lost. And

this is can be seen from $X \sim \Gamma(10,5)$ datasets. Figure 3 told us that though 1-DLID method may not achieve a good result towards α when there are 5% data lost, it indeed works very well on any other datasets except dataset 1. We have to say, this error is acceptable, because it is only about 5.43%. Similarly, the trend in figure 4 seems the same as figure 3 towards the errors of β. Figure 4 described that 1-DLID approach presents a much better results which the errors decreased from 8.79% to 0.53% gradually while the errors of other methods are bigger than 1-DLID method, particularly, take the mean imputation method for example, its error reached nearly 50% when there are 50% data lost, which is unacceptable.

6 Implementation and evaluation with real-world data set

To evaluate the proposed method, a suitable and standard data set is needed. In this paper, the data set from Wisconsin Diagnostic Breast Cancer (WDBC) was chosen for our experiments. The original data set was provided by Dr.Willian H. Wolberg, W. Nick Street and Olvi L. Mangasarian of University of Wisconsin. We chose this kind of dataset because:

1. It has sufficiently large number of attributes and records, which is not only make sense for this paper, but also helpful for our future experiments, which will based on large numbers of attributes.
2. Except the class attributes, all the other data are numerical data, which suit for the proposed approach.
3. The dataset is from the UCI website, which is reliable and can be downloaded.

Information about the dataset:
Number of the instances: 198
Number of attributes:34 input real-valued features(ID, outcome, 32 real-valued input features). Based on our method is good at processing one-dimensional data set, we randomly chose one attributes for our experiments.
Experiment attribute feature: field 4, which is the Mean Radius of the cell nucleus.
Experiment data Missing attribute values: None
Missing attribute values of the whole data: Lymph node status is missing in 4 cases.

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 4 is Mean Radius, field 14 is Radius SE, and field 24 is Worst Radius.

The chosen dataset is selected for evaluating four missing data imputation approaches because it is suitable for doing experiment on computer for the Delete method (listwise deletion), Mean imputation method, EM imputation and 1-DLID approach. While Filling Manually and Hot Decking Imputation is too time consuming and is not good at processing large number data in data mining. Multiple Imputation works well in high-dimensional data set, and we will do experiment based on MI method in our future experiment once we explored our method works in multi-dimensional data set.

Missing data were deleted randomly by the computer, and then recovered with the help of the four method separately and then compared the index of cluster: Rank index and Silhouette index.

The following tables and figures show the performance of the proposed method and other approaches. Precisely, the performance for the reconstructing of the WDBC dataset is based on the performance of the classification measures. That is to say, the higher rate or the better classification performance means better imputation of the missing values.

Table 4: Results of Rand Index

Missing values (%)	Clustering Accuracy			
	Listwise deletion	Mean imputation	EM imputation	Proposed method
10	64.34%	66.99%	65.26%	63.61%
20	62.58%	68.2%	61.56%	85.07%
30	58.11%	72.76%	56.39%	82.56%
40	61.15%	73.46%	81.57%	85.93%
50	66.23%	80.94%	68.2%	92.21%

The following figure 5 illustrates the performance of the methods (in terms of Rand Index) towards different percentage of missing values. It can be seen from the chart that the proposed 1-DLID method shows a better results than the other three methods.

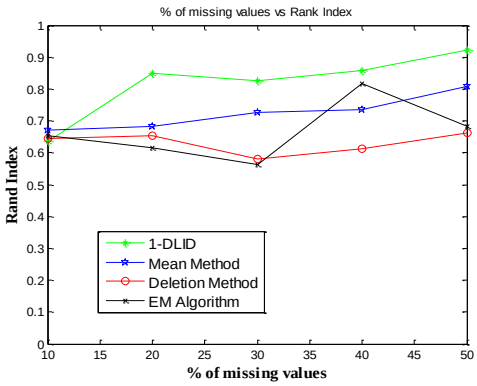

Fig. 5: Percentage of Missing Values vs. Rank Index

The following table, the results in terms of Silhouette index. The higher rate or better classification performance means the better imputation of missing values.

Table 5: Results of Silhouette index

Missing values (%)	Clustering Accuracy			
	Listwise deletion	Mean imputation	EM imputation	Proposed method
10	74.58%	76.75%	74.96%	75.39%
20	75.05%	79.01%	76.8%	89.75%
30	73.8%	82.51%	75.31%	87.17%
40	70.85%	80.76%	85.74%	90.21%
50	75.85%	87.1%	77.8%	94.76%

The following figure 6 illustrates the performance of the methods (in terms of Silhouette index) towards different percentage of missing values. It can be seen from the chart that the proposed 1-DLID method shows a better results than the other three methods.

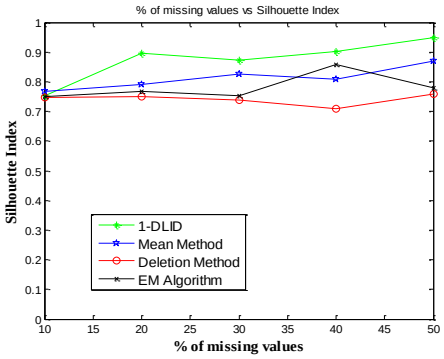

Fig. 6: Percentage of Missing Values vs. Silhouette index

Just as we discussed in the first section, delete method (or listwise deletion) and mean imputation method usually cause bias when the data is not normal distribution. For example, several noise data or leverage data can change the mean of the data set, and delete some data usually causes some important information got lost. For EM algorithm, it may converge to a local maximum of the observed data likelihood function, and this depending on starting values. However, 1-DLID approach doesn't have such problems. It can make good use of the existing data values. One problem is that how to choose a good interval for the 1-DLID method is very important, because it can reflect the recovery data set. We will do more research and talk this in one particular paper.

7 Conclusions

This paper presented an 1-DLID approach for missing data recovery. In the experiments, data are generated based on Gaussian distribution and Gamma distribution while the missing data is created by computer, that is to say the missing data were chosen randomly from computer and removed them from the related complete data set to get the test data sets we need. Regarding to the 1-dimensional data set, we compare our method with deletion method (or listwise deletion), mean imputation method and EM algorithm and compare our results with the other approaches. The experimental results showed that our approach has a precise results, especially when missing values between 10% and 30%. More importantly, the proposed method is easy to use. If needed researchers can use the 1-DLID algorithm several times based on different interval, and then choose the average values of each data, the results would be improved and more reliable. From the generated experiments, we can see that the proposed approach does not change the distribution of the data set. And from the real-world data set, a higher accuracy is achieved compared with the other methods.

8 Future Works

This paper has presented an 1-DLID approach which is used for data recovery for the analysis of incomplete data set. Like most method, 1-DLID method can create data for the data set with missing values while 1-DLID method do not need to provide the estimation distribution, which is easier to use. However, this method works well in 1-dimensional data set. Our future work is to do more exploration and make it work in 2-dimensional data set and then multidimensional data set and compare with MI algorithm. What is more, how to choose a proper interval that used for data recovery is of vital important, and it becomes one of the problems that we should overcome in the future. Lastly, we would like to develop it into a sophisticated software which will contribute to the society and help people recover the missing data.

9 References

[1] A. Ragab, S. Yacout, Ouali and S. Mohamed, Intelligent Data Mining For Automatic Face Recognition, Turkish Online Journal of Science & Technology, Apr. 2013, Vol. 3, Issue 2, pp. 92-96.

[2] K. Lokanayaki and A. Malathi, Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis, International Journal of Computer Applications, Sep. 2013, Vol. 77, pp. 26-29.

[3] R. Somasundaram and R. Nedunchezhian, " Evaluation on Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications, Vol 21, No. 10, May 2011, pp. 14-19.

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2006.

[5] R. Little and D. Rubin. Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, 1987.

[6] E. Foster and G. Fang: Alternative methods for handling attrition: an illustration using data from the Fast Track evaluation, Eval Rev 2004, Vol.28, pp. 434–464.

[7] S. Lynch. Missing data (Soc 504), Princeton University Sociology 504 Class Notes, 2003.

[8] P. Allison: Missing data. Thousand Oaks, CA: Sage; 2000.

[9] A.P.Dempster, N.M. Laird, and D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. Ser. B., 39, 1977.

[10] C.F.J. Wu. On the convergence properties of the EM algorithm, The Annals of Statistics 11, pp. 95–103, 1983.

[11] C.F. Huang, Demonstration of benefit of information distribution for probability estimation, Signal Processing 80.6, pp.1037-1048, 2000.