
ELSEVIER

Pattern Recognition Letters 23 (2002) 1613–1622

Pattern Recognition Letters

www.elsevier.com/locate/patrec

A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets

Xiaolu Huang, Qiuming Zhu *

Digital Imaging and Computer Vision Laboratory, Department of Computer Science, University of Nebraska at Omaha, Omaha, NE 68182-0500, USA

Received 13 July 2001; received in revised form 7 November 2001

Abstract

Missing data handling is an important preparation step for most data discrimination or mining tasks. Inappropriate treatment of missing data may cause large errors or false results. In this paper, we study the effect of a missing data recovery method, namely the pseudo-nearest-neighbor substitution approach, on Gaussian distributed data sets that represent typical cases in data discrimination and data mining applications. The error rate of the proposed recovery method is evaluated by comparing the clustering results of the recovered data sets to the clustering results obtained on the originally complete data sets. The results are also compared with that obtained by applying two other missing data handling methods, the constant default value substitution and the missing data ignorance (non-substitution) methods. The experiment results provided a valuable insight to the improvement of the accuracy for data discrimination and knowledge discovery on large data sets containing missing values.

Keywords: Missing data; Missing data recovery; Data imputation; Data clustering; Gaussian data distribution; Data mining

1. Introduction

The ever-growing data sets stored in large amount of databases and data warehouses are treasure mines with precious information (knowledge) hidden in them. In order to retrieve those information, tools for data mining and knowledge discovery such as on-line analysis process (OLAP), statistical analyzers, and hierarchical clustering are widely used by businesses, government and scientific research institutions (Agrawal et al., 1996).

In most databases and data warehouses, raw data are not ready to be processed by data mining tools because they may contain a lot of irrelevant, inconsistent, or missing data items. Therefore, data discrimination and mining is often a multi-stage process in which people use some formal or informal methods to evaluate the appropriateness of the problems, define processing stages and expected solutions, implement technical approaches and strategies, and produce measurable results. For example, in bio-informatics a typical process for gene expression data discrimination and mining involves roughly the stages of data collection

* Corresponding author. Tel.: +1-402-554-3685; fax: +1-402-554-3400.

E-mail address: zhuq@unomaha.edu (Q. Zhu).

0167-8655/02/$ - see front matter
PII: S0167-8655(02)00125-3

and preparation, cleansing and filtering, clustering and synthesizing, and then the stages of knowledge extraction and representation. Each of these stages has a specific objective and a set of functions to perform.

The data preparation stage aims to getting rid of erroneous data and find the most accurate ways to represent the uncertain information. The absence of certain values for relevant data attributes in data items can seriously affect the accuracy of data mining results. Missing data handling is one of the main issues often dealt with in the data preparation steps. In most cases, missing data should be pre-processed (recovered) so as to allow the whole data set to be processed by a data-mining tool. It has also been known that data preparation and filtering steps take considerable amount of processing time in many data mining projects (Pyle, 1999).

While attributes in most data sets can be distinguished in categories of randomly distributed or non-randomly distributed, the missing data can also be distinguished in these two categories: (1) non-randomly distributed, and (2) randomly distributed. That is, the mechanisms underlying the situations of certain data being missing can be characterized as either random or non-random. But this randomness is by no means related to the randomness of the attribute in the original data set, or at least we do not assume that in this study.

Randomly distributed missing data are the most commonly encountered cases in scientific, economic and business data mining applications (Afifi and Elashoff, 1966). In this paper, we focus on the methods for handling the randomly distributed missing data only. We also focus on the kind of data sets that are in Gaussian random distributions. That is, we study the effects of randomly distributed missing data handling methods on randomly distributed data set. However, it must be pointed out that the randomness of the missing data is unknown in our study and is possibly totally different from the randomness of the original data set.

According to Little and Rubin (1987), the procedures for treating the randomly missing data can be grouped into four categories in general:

(1) *Ignorance-based procedures:* This is a non-recovery method and is the most trivial approach. When some variables are not recorded for some of the data attributes, a simple expedient is to discard the incompletely recorded units entirely and to analyze only the units with complete data. It is generally easy to carry out and may be satisfactory with certain data analysis tasks. However, it can lead to serious biases, especially when missing data are not randomly distributed. Moreover, it is usually very difficult to evaluate the errors caused by the discarded data records (Afifi and Elashoff, 1966). Notice that this method is different from the non-substitution methods. It throws out the entire data point rather than just ignore the missing data values.

(2) *Weighting-based procedures:* This is also a non-substitution (also non-recovery) procedure and is most commonly used in the inferences from sample survey data that contains non-response answers. The weights are designed such that they are inversely proportional to the probability of data presence in selections according to some empirical results. The purpose of the method is to reduce the effect of attributes with large percentage of missing values. The procedure is more applicable to non-randomly distributed missing data (Orchard and Woodbury, 1972).

(3) *Model-based procedures:* This is a missing data recovery method. A missing data replacement is generated by defining a model for the partially missing data and biasing inferences on the likelihood under that model, with parameters estimated by procedures such as maximum likelihood. Advantages of this approach are the flexibility and divergence. One example of the application of this approach is seen in (Krishnamoorthy and Maruthy, 1998), where they proposed three simple exact tests as alternatives to the traditional likelihood ratio test to assess the accuracy of this missing data reconstruction procedures. However, the complexity of these procedures prevented their applications to data mining that deal with very large data sets. It has also been known that the model-based procedures are more suitable to data that maintain certain non-static regularities, such as the time series data sets that are not common to most data mining applications (Little, 1982).

(4) *Imputation-based procedures:* This is the type of missing data substitution methods we discuss in this paper. In this approach, the missing values are filled in by certain means of approximation and the resultant completed data are analyzed by standard statistical analytical methods (Hartley and Hocking, 1971). Commonly used procedures for imputation include: (a) Hot deck imputation, where recorded units in the sample are substituted by a value obtained from the present data set following certain rules, for example the value from the nearest data record (Sande, 1996). (b) Default value imputation, where a constant is used to substitute the missing values, for example all missing values being replaced by value zero or the median of the value range (Afifi and Elashoff, 1966). (c) Statistical imputation, where the missing values are substituted by a statistically inspired value that has a high likelihood for the true occurrence, for example the mean values computed from the set of non-missing data records (Titterington and Sedransk, 1989). (d) Regression imputation, where the missing variables for a unit are estimated by values derived from the known variables according to a given function or some functional forms (Pawlak, 1993). One example of the application of regression imputation based missing data handling approach is Letfus' paper (Letfus, 1999). Problem with the regression imputation is that it raises another critical issue of how to verify the legitimacy of the underlying function assumed for the regression.

Besides the above four procedures, there are also some other missing data handling methods, such as the induction substitution approaches and technically skipping missing data approaches. Strictly speaking, induction substitution approaches also belong to imputation-based procedures (Sande, 1996; Titterington and Sedransk, 1989). However, induction substitution approaches are more individual data object specific among the missing data handling approaches. In this paper, we will focus our study on the above-mentioned imputation-based procedures for missing data recovery.

In a previous study on the effect of missing data, Zhu has derived an analytic form for estimating the error probability of classifications made on the partially available data sets versus that on the complete sets by using the Bhattacharyya bounds (Zhu, 1990). It has been shown that an upper bound of minimum probability of error, under the condition that the data attributes are independently distributed, is established at

$$\frac{1}{\prod_{i=k+1}^{n} \int_{R(x_i)} \sqrt{p(x_i|\omega_1)p(x_i|\omega_2)}\mathrm{d}x_i}, \tag{1.1}$$

where a complete data item is given as $\mathbf{x} = [x_1, x_2, \ldots, x_k, x_{k+1}, \ldots, x_n]$ and a data item with missing attribute values is given as $\mathbf{x}^k = [x_1, x_2, \ldots, x_k]$. The ω_1 and ω_2 are two symbols denoting two distinct classes of the data sets. In the cases that the data attributes are in independent Gaussian distributions, the minimum probability of error of classification with data item $\mathbf{x}^k$ versus $\mathbf{x}$ is bounded by

$$\prod_{i=k+1}^{n} \frac{\sqrt{\sigma_{i1}^2 - \sigma_{i2}^2}}{\sqrt{\sigma_{i1}^2 \sigma_{i2}^2}} \exp\left(\frac{1}{4}\frac{(\mu_{i1} - \mu_{i2})^2}{\sigma_{i1}^2 + \sigma_{i2}^2}\right), \tag{1.2}$$

where (μ_{i1}, μ_{i2}) and $(\sigma_{i1}, \sigma_{i2})$ are the mean and variance values of the ith attribute for the data items in classes ω_1 and ω_2 respectively. We need to mention that the assumption for attribute independence is quite strong in many data discrimination and mining applications. It is common that for data in real applications most of them have correlated features. So the above metric may not be measurable directly and precisely in real applications. As indicated, the metric only gives a theoretically minimum probability of error under the independence assumption. When dealing with real applications, the error rate varies depending on the correlativeness of the data attributes and methods used to handle the missing data, such as the proposed approaches discussed in this paper.

In this paper, we study the practical effects of the data processing errors when some missing data recovery methods are applied to the data sets. Assumptions in our study include: (1) the locations of the missing data in the data set are random with an unknown distribution, (2) the values of the missing data are random with an unknown distribution, (3) the data records are not labeled, i.e., no categorical information about the data items is

given, (4) the missing data are numerically valued, and (5) the data attributes of the data sets are uncorrelated. That is, each data attribute has its own distribution of possible values. However, the values of each attribute may be governed by a Gaussian or non-Gaussian distribution (the exact parameters of these distributions are unknown). In other words, each data attribute is governed by a univariate Gaussian distribution in case that the data set is Gaussian randomly distributed. The assumption (5) can be relaxed if we concentrate on the comparisons of the missing data handling methods, rather than a quantitative measurement of the probability of error of each method precisely. In fact, the correlativeness of data attributes could be used to assist the missing data recovery, especially for the model-based and regression methods. But these are not the main concentration of this paper. Our work is focused on the missing data substitution methods. In the methods we studied, the correlativeness has less effect. The main reason we list the assumption (5) is to indicate that our methods do not make use of the attribute correlations.

The results of our missing data recovery methods are evaluated by comparing the clustering results to that obtained by employing certain other missing data recovery techniques. These methods include that making the use of constant default and statistical imputations, as well as skipping (ignoring) the attributes that have missing values. The method is also evaluated on the basis of the clustering results made on the complete set of the data items (non-missing data sets). Three major parameters are used to generate the testing data sets in the evaluation: (1) missing data rate, ranging from 5% to 40% measured on the data set, (2) number of classes (or clusters) in the data set, and (3) the ranges of Gaussian variances in the experimental Gaussian distributed data sets.

The paper is organized as follows. Section 2 describes the basis of the pseudo-nearest-neighbor substitution method and the procedure. Section 3 presents our experimental results of the proposed method and compares it with three other missing data imputation and recovery/non-recovery methods. Section 4 contains conclusion remarks.

2. Computation of pseudo-nearest-neighbor

To derive the pseudo-nearest-neighbor method for missing data recovery, here we are going to first introduce the concept of *pseudo-similarity* (or dissimilarity measurement) between a data record $\mathbf{x}$ with missing values and a data record without missing values, as well as between two data records with different number of missing values.

Let $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ be a data record in a data set $\{\mathbf{x}\}$. A data record $\mathbf{x}$ with missing values means that some of the elements $x_i \in \mathbf{x}$, $i = 1, \ldots, n$, have no valid attribute values present at the time the vector $\mathbf{x}$ is to be processed as an input to a data mining system. To facilitate the expression and computation, we use a "NULL" symbol to represent the missing value of x_i. That is, when the value of an attribute x_i is missing, we say that it has a value NULL. Note that in random missing cases, any one of the n elements of $\mathbf{x}$ could have a NULL value. Without losing generality and for convenience of expression and computation, we use $\mathbf{x}^k = [x_1, x_2, \ldots, x_k, \mathrm{NULL}_{k+1}, \ldots, \mathrm{NULL}_n]$, $k < n$, to represent a data record $\mathbf{x}$ with $(n-k)$ missing attribute values. That is, we always move the missing elements of $\mathbf{x}$ to its right end and assume that the missing values of $\mathbf{x}^k$ with respect to $\mathbf{x}$ are $[x_{k+1}, \ldots, x_n]$. In brief, we write $\mathbf{x}^k = [x_1, x_2, \ldots, x_k]$, which stands for a vector $\mathbf{x}$ with k non-missing elements. We also say $\mathbf{x}^k$ is an incomplete data record.

Let $\{\mathbf{c}\}$ be a set of categorical centers of data set $\{\mathbf{x}\}$. That is, $\{\mathbf{c}\}$ is a set of complete data record, $\mathbf{c} = [c_1, c_2, \ldots, c_n]$. With the same re-ordering of data elements as $\mathbf{x}^k$ for $\mathbf{c}$, a pseudo-similarity between a data record $\mathbf{x}$ with missing values (here actually the $\mathbf{x}^k$) and a complete data record $\mathbf{c}$ can be defined as

$$S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{c}) = \sum_{i=1}^{k} \Phi(x_i^k, c_i), \tag{2.1}$$

where $\Phi(\cdot)$ is a certain kind of similarity (or distance metric) measurement function. The $S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{c})$ is useful in data clustering. When performing clustering of the data set $\{\mathbf{x}\}$, the $S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{c})$ with respect to each cluster center $\mathbf{c}$ in the collection is compared with each other to determine the belonging of $\mathbf{x}^k$. Note that the categorical center $\mathbf{c}$

always has a complete set of attributes that can be computed on the basis of the presented values of data records or by an initial random selection.

Let $\mathbf{x}^k$ and $\mathbf{x}^l$ be two incomplete data records of data set $\{\mathbf{x}\}$. Again, we can re-arrange the order of the data elements in $\mathbf{x}^k$ and $\mathbf{x}^l$, in such a way respectively, that (1) if an element has its value missing in both $\mathbf{x}^k$ and $\mathbf{x}^l$, then it is placed toward the right end of the vectors, (2) if an element has its value present (i.e., non-missing) in both $\mathbf{x}^k$ and $\mathbf{x}^l$, then it is placed toward the left end of the vector. Let us use a symbol "#" to represent the value that is missing in one of the vectors $\mathbf{x}^k$ and $\mathbf{x}^l$ but not in both, then we can have the $\mathbf{x}^k$ and $\mathbf{x}^l$ be expressed as $\mathbf{x}^k = [x_1, x_2, \ldots, x_d, \#_{d+1}, \ldots, \#_k, \mathrm{NULL}_{k+1}, \ldots, \mathrm{NULL}_n]$ and $\mathbf{x}^l = [x_1, x_2, \ldots, x_d, \#_{d+1}, \ldots, \#_1, \mathrm{NULL}_{l+1}, \ldots, \mathrm{NULL}_n]$, where $d \leqslant \min(k, l)$. Note that it does not matter whether k equals to l or not. The pseudo-similarity between $\mathbf{x}^k$ and $\mathbf{x}^l$ is defined as

$$S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{x}^l) = \sum_{i=1}^{d} \Phi(\mathbf{x}_i^k, \mathbf{x}_i^l) = d \sum_{i=1}^{d} \left(\frac{x_i^k}{\sqrt{\sum_{i=1}^{d} \mathbf{x}_i^k \mathbf{x}_i^k}} \frac{x_i^l}{\sqrt{\sum_{i=1}^{d} \mathbf{x}_i^l \mathbf{x}_i^l}} \right). \tag{2.2}$$

The measurement is actually a weighted correlation value between the two vectors with partially missing element values. It takes count of (1) the number of commonly present elements, and gives more weight on the vectors having more present elements, and (2) the correlation on the present element values. Thus, if two vectors have the same correlation value, then a larger pseudo-similarity $S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{x}^l)$ is given to the vectors having less missing elements. Table 1 shows some examples of the $S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{x}^l)$ measurements.

Nearest neighbor substitution is a typical hot deck imputation method to handle missing data. Let $\mathbf{x}^k = [x_1, x_2, \ldots, x_k, \mathrm{NULL}_{k+1}, \ldots, \mathrm{NULL}_n]$ be the data record with missing values to be recovered. The method first searches for a data record $\mathbf{x}^l$ within the data set $\{\mathbf{x}\}$ such that (1) $\mathbf{x}^l$ has the presence of value $\mathbf{x}_{k+1}$, (2) $\mathbf{x}^l$ has the largest pseudo-similarity value, based on the present data attribute values, (3) the present value $\mathbf{x}_{k+1}$ of $\mathbf{x}^l$ is used to replace the NULL_{k+1} in $\mathbf{x}^k$. Since the pseudo-similarity measurement is used in this evaluation, we call the $\mathbf{x}^k$ and $\mathbf{x}^l$ *pseudo-nearest-neighbors*, and thus the name for the missing data recovery method. It needs to point out that the term "pseudo-nearest-neighbor" was also used by Mojirsheibani (1999) to describe an approach for combining different classifiers in order to construct more effective classification rules. The principle of the technique used there is actually the same as we use here, except that it is used here to identify the most similar data points for missing data recovery.

Table 1
Examples of pseudo-similarity measurement $S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{x}^l)$

$\mathbf{x}^k$	$\mathbf{x}^l$	$S_{\mathrm{p}}(\mathbf{x}^k, \mathbf{x}^l)$
11######	11######	2
10######	11######	1.41
101#####	110#####	1.5
110#####	111#####	2.45
1101####	1110####	2.67
1101####	1111####	3.46
11011###	11110###	3.75
11011###	11111###	4.47

A procedure of the pseudo-nearest-neighbors substitution method for missing data recovery is presented as follows:

Procedure *pseudo-nearest-neighbor method for missing-data recovery*
Pre-condition: a data set $\{\mathbf{x}\}$ with members in format of $\mathbf{x}^k = [x_1, x_2, \ldots, x_k, \mathrm{NULL}_{k+1}, \ldots \mathrm{NULL}_n]$
Post-condition: a data set $\{\mathbf{x}\}$ with members in format of $\mathbf{x}^k = [x_1, x_2, \ldots, x_n]$, i.e., the missing values being substituted by corresponding values of the pseudo-nearest-neighbors.
Computation:

```
For each vector x^k
{  For each NULL valued element x_i^k of x^k
   {  For each x^l ∈ {x} − x^k
      {  If the x_i^l value is non-missing
         Compute S_p(x^k, x^l)
      }
      Find the x^l* that has the largest value of
      S_p(x^k, x^l) among all x^l examined
   }
```

Replace the element $\mathbf{x}_i^k$ of $\mathbf{x}^k$ by the $\mathbf{x}_i^l$ value of $\mathbf{x}^{l*}$

}

3. Experimental results

3.1. Compared methods

We experimented with three imputation methods for recovery of randomly distributed missing data. (1) In hot deck imputation, the randomly distributed missing data at a dimension of a data object is filled with the non-null value from the pseudo-nearest-neighbor of the data set, as described above. The procedure is named "mneighbor". (2) In default value imputation, the randomly distributed missing data at a dimension of a data object is filled with the median value of the whole data set. A procedure is named "mmedian" for the median value substitution. The programs are composed of two steps: in the first part, the median values of the attributes are computed according to the present values in the data records. In the second step, it converts all the missing data in the data set into the median value of the data set. (3) In statistical imputation, the randomly distributed missing data at a dimension of a data object is filled with dimension mean of the whole data set, calculated as such:

$$M_j = \sum_{i=1}^{T} V_{kj}/n \tag{3.1}$$

where T is the total number of data objects in a data set; V_{kj} is the valid data value of the k's data object at dimension j; $n = T - N_j$, where N_j is the total number of data objects that have data missing at js dimension. A procedure is named "mgmean" to carry out this computation.

The performances of above methods are also compared with a missing data ignorance (non-substitution) approach. In this approach, a set of categorical centers $\{\mathbf{c}\}$ for the data set $\{\mathbf{x}\}$ is assumed where, each member of $\{\mathbf{c}\}$ is a complete data record $\mathbf{c} = [c_1, c_2, \ldots, c_n]$ of a data set $\{\mathbf{x}\}$. When performing clustering of the data set, the $S_p(\mathbf{x}^k, \mathbf{c})$ on the presented data values of the $\mathbf{x}$ are computed with respect to $\mathbf{c}$ instead of computing $S_p(\mathbf{x}, \mathbf{c})$, where $\mathbf{x}^k$ is a data point of $\{\mathbf{x}\}$ with $n - k$ missing attribute values. That is, the missing values are skipped (ignored) in computing the similarity of the data point with respect to the cluster centers. A procedure named "mskipping" carries this computation.

3.2. Data sets

We evaluate the missing data recovery schemes with respect to data sets having an underlying Gaussian distribution. The Gaussian distributed data sets are generated using random number generators with the following given parameters: (a) the number of clusters (2–50), (b) the number of data attributes (dimensions) in each cluster (2–500), (c) the number of data objects (points) to be generated for each cluster (150–15 000), and (d) the ranges of the Gaussian mean and Gaussian variance values for each dimension of the cluster. Table 2 gives an example of the selected Gaussian means and Gaussian variances for a data set with 10 clusters and 20 attributes, where capital letters A, B, C, ... are class labels. The data set has the mean range from 10 to 50 and variance range from 0.02 to 10. The test data generation procedure then does the following: (1) Generate a data file that contains the original (no missing values) data sets generated according to the parameters above. Each data object also contains a label to indicate the original cluster the data object belongs to (e.g, A, B, C, ..., etc); and (2) Convert the data sets in the original file to data sets that contains 5%, 10%, 15%, 20%, 25%, 30%, 35%, and up to 40%, respectively, of randomly distributed null values as the data sets of missing values.

3.3. Results

We used a k-means clustering algorithm as a means to qualitatively and quantitatively evaluate the above approaches. We first examine the experimental results on Gaussian distributed data sets with relatively larger mean value ranges so that the clusters in the data sets are relatively separated. That is, the distributions of the data sets of different clusters have only little overlapping regions in the data space. Each test case is done with respect to

Table 2
An example of Gaussian means and Gaussian variances of a data set

Dimension	Cluster A		Cluster B		Cluster C		Cluster D		Cluster E		Cluster F		Cluster G		Cluster H		Cluster I		Cluster J	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
1	19	7.00	37	10.00	17	10.00	10	0.02	35	5.00	40	7.00	10	10.00	42	5.00	42	0.02	16	10.00
2	12	7.00	29	7.00	33	0.02	38	5.00	44	7.00	19	0.02	40	0.02	24	5.00	18	10.00	11	7.00
3	27	5.00	39	10.00	17	7.00	23	10.00	24	10.00	21	7.00	22	0.02	23	0.02	22	10.00	34	0.02
4	42	10.00	23	5.00	16	5.00	35	7.00	49	5.00	34	5.00	16	5.00	22	10.00	48	0.02	25	7.00
5	42	10.00	24	5.00	48	10.00	32	5.00	22	7.00	37	10.00	37	5.00	32	5.00	25	5.00	48	0.02
6	18	0.02	27	5.00	44	0.02	21	7.00	41	10.00	20	7.00	46	5.00	31	5.00	28	7.00	23	7.00
7	21	7.00	36	0.02	45	7.00	18	5.00	14	5.00	43	10.00	17	0.02	40	5.00	35	10.00	25	0.02
8	17	0.02	17	0.02	23	10.00	34	10.00	24	7.00	32	5.00	23	5.00	45	7.00	40	5.00	36	0.02
9	11	7.00	41	10.00	48	10.00	45	0.02	37	7.00	27	10.00	32	5.00	40	10.00	23	7.00	36	7.00
10	14	5.00	41	7.00	46	0.02	39	5.00	23	5.00	29	10.00	34	0.02	42	10.00	37	0.02	15	10.00
11	18	5.00	37	5.00	48	5.00	37	0.02	42	0.02	12	0.02	19	5.00	29	5.00	23	5.00	25	5.00
12	37	7.00	21	5.00	38	7.00	16	5.00	27	5.00	11	5.00	33	10.00	49	5.00	15	10.00	40	7.00
13	29	5.00	30	7.00	40	10.00	47	10.00	25	10.00	10	0.02	48	10.00	17	10.00	38	10.00	43	7.00
14	40	10.00	41	0.02	38	10.00	38	5.00	22	10.00	26	7.00	42	10.00	13	0.02	13	5.00	12	0.02
15	40	7.00	24	10.00	34	7.00	16	7.00	32	10.00	47	10.00	25	5.00	33	10.00	42	0.02	48	10.00
16	46	0.02	45	10.00	12	7.00	13	10.00	31	5.00	24	7.00	20	5.00	42	0.02	38	5.00	28	5.00
17	20	5.00	24	0.02	19	5.00	12	5.00	36	7.00	12	0.02	33	5.00	27	0.02	43	5.00	11	7.00
18	29	0.02	48	5.00	31	0.02	13	5.00	13	0.02	28	7.00	43	10.00	48	7.00	17	0.02	11	10.00
19	40	7.00	31	0.02	21	10.00	46	7.00	35	7.00	39	0.02	43	7.00	21	10.00	26	10.00	24	7.00
20	44	10.00	38	7.00	39	0.02	46	0.02	46	0.02	24	5.00	31	0.02	21	7.00	29	5.00	19	10.00

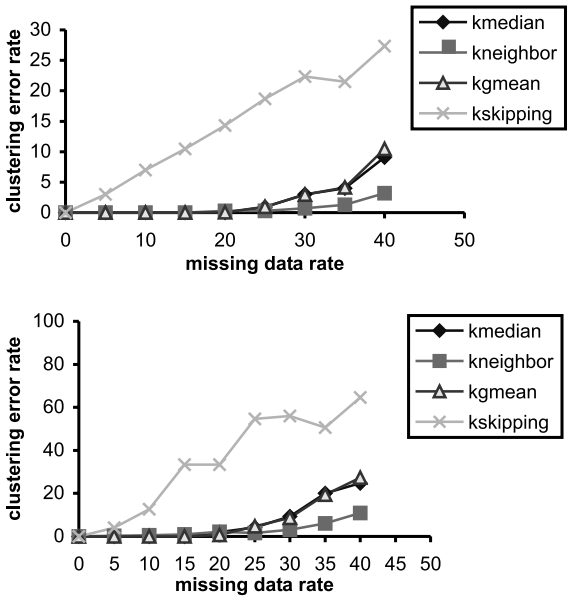

Fig. 1. Experimental results showing clustering errors versus the percentage of missing values for the missing data handling methods on a data set with 10 clusters of fixed mean values but different variances: (a) variance range of 5 and (b) variance range of 10.

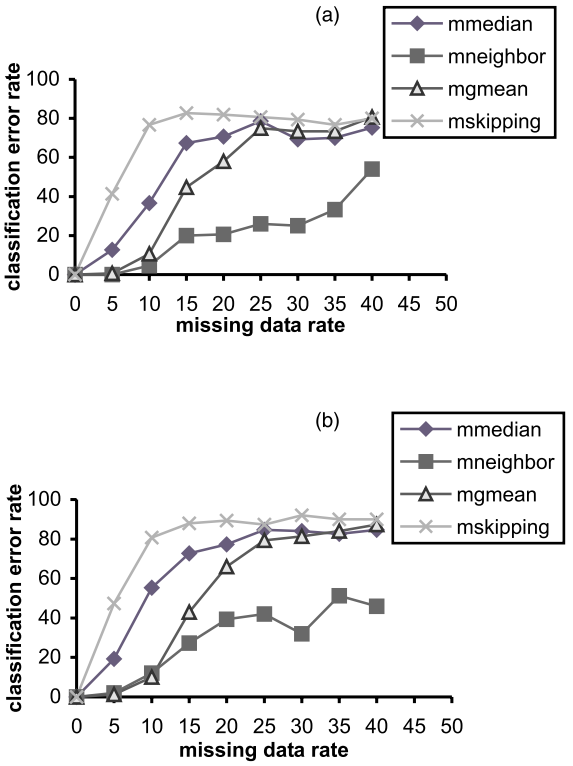

Fig. 2. Experimental results showing clustering errors versus the percentage of missing values for the missing data handling methods on a data set with different cluster numbers and varying mean and variance values: (a) five clusters with variance range of 5 and (b) ten clusters with variance range of 10.

varying missing data rate that ranges from 5% up to 40%. Fig. 1 shows the clustering errors for each of the missing data handling method with respect to the percentage of missing values. The horizontal axis denotes the percentage of data values absent in the records, while the vertical axis denotes the clustering error percentage compared with the original labels of the data set.

We then examine the experimental results on Gaussian distributed data sets with relatively smaller mean value ranges so that the clusters in the data sets are relatively mixed, that is, there are some considerable amount of regions in the data space where data distributions of different clusters have overlapped. Again the tests are done with respect to varying missing data rate that ranges from 5% up to 40%. Fig. 2 shows the clustering errors for each of the missing data handling method with respect to the percentage of missing values.

The experimental results show that the median, neighbor and mean substitution methods all outperformed the skipping methods. Among the three substitution methods, the nearest neighbor substitution has the best performance. The median substitution and mean substitution has almost the same amount of clustering error. This is understandable because of the closeness of these two values in the data sets of Gaussian distributions.

Fig. 3 shows the experimental results of the missing data handling methods with respect to the percentage of missing values on Gaussian distributed data sets of the same mean positions but different variance values. The notation ug002_10b, ug5_10b, and ug10_10b stand for uniform Gaussian distributions of 10 clusters with variance value ranges of 2, 5 and 10, respectively. The missing data rate changes from 5% up to 40% for each of the test data set. Again, it shows that the pseudo-nearest-neighbor method has the best performance than the other methods because the

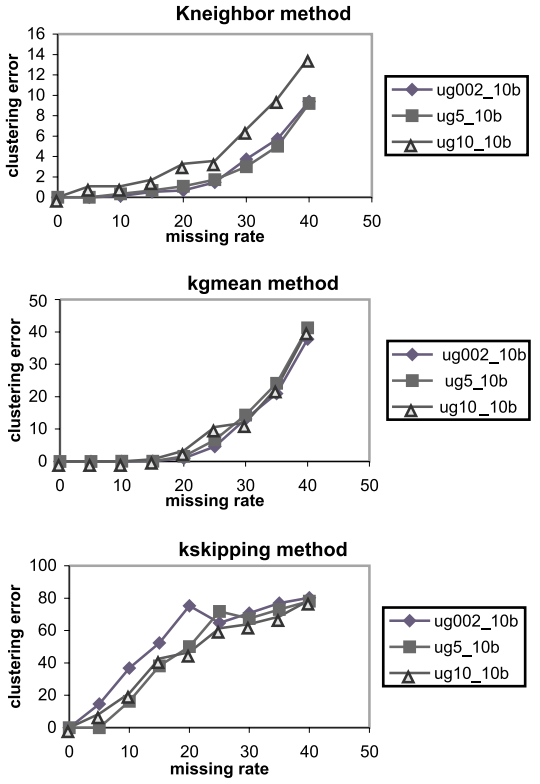

Fig. 3. Experimental results showing clustering errors versus the percentage of missing values for the missing data handling methods on a data set with fixed cluster numbers and mean values but varying variance value ranges.

pseudo-nearest-neighbor method captures the essence of pattern similarities in the original data set.

4. Conclusions

In this paper, we studied a new method, namely the pseudo-nearest-neighbor substitution, for missing data handling in preparation of data sets for data discrimination and mining applications. Performance of the method is compared with other substitution and non-substitution approaches for dealing with data sets containing randomly missing data attribute values. The experimentation results have provided following insights: (1) there is a tendency of increasing classification error rate along the increase of the cluster number k in the data set for all the missing data handling approaches; (2) there is a tendency of increasing classification error rate along the increase of the Gaussian variance ratio for all the missing data handling approaches; (3) The non-substitution (ignorance by skipping the attribute) approach is an inferior missing data handling approach in dealing with Gaussian randomly distributed data sets, and (4) the pseudo-nearest-neighbor approach provides the best results to Gaussian random data sets among the substitution and non-substitution methods evaluated in our experiments. The application of these results to data mining and knowledge discovery could help the selection of missing data handling method during the data preparation step for different data structures and enable a more reliable and efficient decision making under uncertainties and incompleteness of data collections presented.

Acknowledgements

The work described in this paper was partly supported by AFOSR grant no. F49620-99-1-0211. The authors thank the anonymous reviewers for their valuable and very detailed comments that helped improve the presentation of this paper.

References

Afifi, A., Elashoff, R., 1966. Missing observations in multivariate statistics I: Review of the literature. J. Am. Stat. Assoc. 61, 595–604.

Agrawal, R., Metha, M., Shafer, J., Srikant, R., Arning, A., Bollkinger, T. 1996. The Quest Data Mining System, Proceedings of 1996 International Conference on Data Mining and Knowledge Discovery (KDD'96), Portland, Oregon, pp. 244–249.

Hartley, H., Hocking, R., 1971. The analysis of incomplete data. Biometrics 27, 783–808.

Krishnamoorthy, K., Maruthy, K., 1998. Some simple test procedures for normal mean vector with incomplete data. Ann. Institutional Stat. Math. 50 (3), 531–542.

Letfus, V., 1999. Daily relative sunspot numbers 1749–1848: reconstruction of missing observations. Solar Phys. 184, 201–211.

Little, R., 1982. Models for nonresponse in sample surveys. J. Am. Stat. Assoc. 77, 237–250.

Little, R., Rubin, R., 1987. In: Statistical Analysis with Missing Data. Wiley, New York, pp. 24–29.

Mojirsheibani, M. 1999. A Pseudo-Nearest-Neighbor Combined Classifier. Proceedings of the 31st Symposium on the

Interface: Models, Predictions, and Computing. pp. 195–197.

Orchard, T., Woodbury, M., 1972. A missing information principle: Theory and applications. Proceedings of the 6th Berkeley Symposium on Mathematics, Statistics, and Probability 1, 697–715.

Pawlak, M., 1993. Kernel classification rules from missing data. IEEE Trans. Information Theory 39 (3), 979–988.

Pyle, D., 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, CA.

Sande, I. 1996. Hot deck imputation procedures. In: Madow, W.G., Olkin, I. (Eds.), Incomplete Data in Sample Surveys: Symposium on Incomplete Data Proceedings, vol. III. pp. 225–248.

Titterington, D., Sedransk, J., 1989. Imputation of missing values using density estimation. Stat. Probab. Lett. 8, 411–418.

Zhu, Q., 1990. On the minimum probability of error of classification with incomplete patterns. Pattern Recognition 23 (11), 1281–1290.