Applied Intelligence 20, 239–252, 2004

A Grey-Based Nearest Neighbor Approach for Missing Attribute Value Prediction

CHI-CHUN HUANG
Department of Electronic Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan

HAHN-MING LEE
Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan

Abstract. This paper proposes a grey-based nearest neighbor approach to predict accurately missing attribute values. First, grey relational analysis is employed to determine the nearest neighbors of an instance with missing attribute values. Accordingly, the known attribute values derived from these nearest neighbors are used to infer those missing values. Two datasets were used to demonstrate the performance of the proposed method. Experimental results show that our method outperforms both multiple imputation and mean substitution. Moreover, the proposed method was evaluated using five classification problems with incomplete data. Experimental results indicate that the accuracy of classification is maintained or even increased when the proposed method is applied for missing attribute value prediction.

Keywords: missing attribute values, grey-based nearest neighbor approach, grey relational analysis, the nearest neighbor concept

1. Introduction

Various learning algorithms have been developed for pattern classification. These approaches are typically designed to deal with perfect data. However, real-world classification tasks commonly involve incomplete data, such that the data contain some missing attribute values (or blanks). In fact, incomplete information can be caused by error, equipment failure, change of plans, and so on [11]. Most learning algorithms are not well adapted to some application domains due to the difficulty with missing attribute values (for example, Web applications that contain several blanks).

In supervised learning, a learning system is given a training set of labeled instances, where each instance consists of a feature vector and an output value. However, missing attribute values usually appear in the training set, implying that, in the training phase, a reliable method for dealing with those missing values is necessary. Another important concern is how to classify a new, unseen instance with an incomplete feature vector [26]. Furthermore, the system developer has to concentrate on estimating missing attribute values as accurately as possible to resolve the usefulness of data that contain blanks and thus has opportunities to reduce the classification errors of the learning system.

In general, incomplete data greatly affect the performance of classification algorithms. Hence, a robust and effective approach to handling incomplete data in classification tasks is very important. Both Friedman [16] and Quinlan [25] adopted a common strategy, ignoring blanks, to address problems with unknown attribute values during training. Nevertheless, this method is not applicable when many training instances contain blanks, since the method may yield inferior performance [26]. An alternative method is to throw away instances with missing attribute values in the training phase, but such a method probably results in a loss of valuable information. In [2], an evolving model based on a system of fuzzy rules is proposed to tolerate missing values—predicting the most possible values for the missing attributes. Literature on machine learning includes several techniques for estimating missing attribute values, including the Expectation-Maximization (EM) principle [7], decision tree induction [25], the Bayesian approach [4], and multiple imputation [21, 29]. Most of these methods are quite complicated and time consuming, even though they have already been successfully applied to various incomplete data problems.

This paper proposes a grey-based nearest neighbor method to predict missing attribute values in an easy and accurate manner. The nearest neighbor concept [6, 13] and grey relational analysis [8–10] play the most important roles in the proposed method. According to the nearest neighbor rule, the similarity between an instance and its nearest neighbor—determined from the difference between instances—is certainly maximal (that is, the difference between instances is minimal). Thus, the assumption that an instance and its nearest neighbor would have the same (or nearly the same) attribute values is reasonable. Hence, the known attribute values, derived from the nearest neighbors of an instance with missing attribute values, are used to infer those missing values.

Here, the proposed method can be viewed as an *instance-based learning approach* [1]. To successfully build such an instance-based learning approach, we believe that the relational structure of all instances (i.e., the relationships between instances) in a specific domain should be determined precisely. Generally, similarity functions such as Euclidean distance can be used to determine the 'nearness' (or relationship) between two instances. However, Euclidean-like distances are mainly suitable for some application domains (such as domains with numeric attributes). In the proposed approach, grey relational analysis, which is more appropriate to determine the 'nearness' (or relationship) between two instances than Euclidean-like distances [18], is used to describe the relational structure of all instances. Here, grey relational analysis gives whole relational orders (wholeness [35]) over the entire relational space. Besides, the prediction errors in the proposed method can be bounded due to the nearest neighbor rule [6]. By using the proposed method, missing attribute values can be estimated with high accuracy. Accordingly, an imperfect dataset can be handled as a complete dataset for classification tasks. Meanwhile, the usefulness of data that contain blanks can be resolved. In addition, the proposed method is suitable for both symbolic and numeric attributes.

Two datasets were used to show the performance of the proposed method. Experimental results show that the approach presented here is superior to multiple imputation and mean substitution. Furthermore, the proposed method was evaluated by considering five classification problems with incomplete data. Experimental results indicate that the accuracy of classification can be maintained or even increased if the proposed method is applied beforehand.

The rest of this paper is organized as follows. Sections 2 and 3 review the nearest neighbor concept and grey relational analysis, respectively. Section 4 proposes a grey-based nearest neighbor algorithm to predict missing attribute values. Section 5 gives an example to illustrate the proposed method. Section 6 describes experiments on the Iris flower dataset and five classification tasks. Finally, Section 7 gives our conclusions.

2. The Nearest Neighbor Concept

This section reviews the nearest neighbor concept adopted to predict missing attribute values. In learning from examples, proper decisions (for example, classification, prediction) for a new instance i can be made using information extracted from a set of training instances, in particular from the instances nearest to i. For example, a new employee's salary may be estimated from that of another employee with similar education, work experience, and so on. The 'nearness' between two instances is generally determined by some similarity functions, for example, Euclidean metric. Several learning algorithms, based on the concept of 'nearest neighbor', have been investigated, such as instance-based learning [1] and memory-based reasoning [31].

Since its inception in 1957 [13], the *nearest neighbor (NN) rule* [6] built from the above concept has been successfully applied to a broad range of application domains. This simple principle can be stated as follows. Given a set of training instances, an unseen instance is classified according to its nearest training instance. An extended version, called *majority voting* or the *k-NN rule*, classifies the unseen instance into the majority class of its k nearest neighbors.

The *NN* rule offers many advantages over alternative classification methods. For example, it is fairly straightforward to understand and easy to implement. Furthermore, Cover and Hart [6] have demonstrated that, for any number of classifications, the probability of error of the *NN* rule is bounded between R* and 2R*, where R* denotes the Bayes' error (i.e., optimal error).

3. Grey Relational Analysis

As a measurement method, *grey relational analysis* (GRA) [8–10] is proposed to determine the relationships among a referential observation and compared observations by calculating the *grey relational coefficient* (GRC) and the *grey relational grade* (GRG). Consider a set of observations $\{x_0, x_1, x_2, \ldots, x_m\}$, where x_0 is the referential observation and $x_1, x_2, \ldots, x_m$ are the compared observations. Each observation x_e has n attributes and is denoted as $x_e = (x_e(1), x_e(2), \ldots, x_e(n))$ (Generally, all numeric attributes should be preprocessed and then have associated values between zero and one. The preprocessing methods are discussed later.) The grey relational coefficient is then

$$\mathrm{GRC}(x_0(p), x_i(p)) = \frac{\min_{\forall j}\min_{\forall k}|x_0(k) - x_j(k)| + \zeta\max_{\forall j}\max_{\forall k}|x_0(k) - x_j(k)|}{|x_0(p) - x_i(p)| + \zeta\max_{\forall j}\max_{\forall k}|x_0(k) - x_j(k)|}, \quad (1)$$

where $\zeta \in [0, 1]$ (and, normally, let $\zeta = .5$), $i = j = 1, 2, \ldots, m$ and, $k = p = 1, 2, \ldots, n$.

In Eq. (1), $\mathrm{GRC}(x_0(p), x_i(p))$, which takes a value between zero and one, can be viewed as the similarity between $x_0(p)$ and $x_i(p)$. If $\mathrm{GRC}(x_0(p), x_1(p))$ exceeds $\mathrm{GRC}(x_0(p), x_2(p))$ then the similarity between $x_0(p)$ and $x_1(p)$ is larger than that between $x_0(p)$ and $x_2(p)$; otherwise the former is smaller than the latter. Moreover, if x_0 and x_i have the same values for numeric attribute p, $\mathrm{GRC}(x_0(p), x_i(p))$ will equal one (i.e., the similarity between $x_0(p)$ and $x_i(p)$ is maximal). By contrast, if x_0 and x_i have very different values for numeric attribute p, $\mathrm{GRC}(x_0(p), x_i(p))$ will approximate zero. In Section 4.1, we will detail how the proposed approach is extended to deal with symbolic attributes.

And the grey relational grade is expressed as follows.

$$\mathrm{GRG}(x_0, x_i) = \frac{1}{n}\sum_{k=1}^{n}\mathrm{GRC}(x_0(k), x_i(k)), \quad (2)$$

where $i = 1, 2, \ldots, m$.

Clearly, the GRG takes a value between zero and one. The significant effect of grey relational analysis is as follows.

If $\mathrm{GRG}(x_0, x_1)$ exceeds $\mathrm{GRG}(x_0, x_2)$ then the difference between x_0 and x_1 is smaller than that between x_0 and x_2; otherwise the former is larger than the latter.

Despite its simplicity, grey relational analysis obeys four principal axioms [35]:

(1) *Normality*: $\mathrm{GRG}(x_0, x_i)$ takes a value between zero and one.
(2) *Dual symmetry*: If only two observations (x_0 and x_1) are made in the relational space, then, $\mathrm{GRG}(x_0, x_1) = \mathrm{GRG}(x_1, x_0)$
(3) *Wholeness*: If three or more observations are made in the relational space, then GRG $(x_0, x_i)_{\text{seldom equals}}$ $\mathrm{GRG}(x_i, x_0)$, $\forall i$
(4) *Approachability*: $\mathrm{GRG}(x_0, x_i)$ decreases as the difference between $x_0(p)$ and $x_i(p)$ increases (other values in Eqs. (1) and (2) are held constant).

Based on these axioms, grey relational analysis offers some advantages. For example, it gives a normalized measuring function (Normality)—a proper method for measuring the similarities or differences among observations—to analyze the relational structure. And grey relational analysis gives whole relational orders (wholeness [35]) over the entire relational space. In this paper, the relationships among instances, used for predicting missing attribute values, are determined according to the magnitude of GRG ranging from zero to one.

Before the grey relational coefficient and the grey relational grade are calculated, one of the following methods should be used for preprocessing numeric attributes [23].

(1) Upper-bound effectiveness measurement (larger is better)

$$x'_p(j) = \frac{x_p(j) - \min_{\forall i} x_i(j)}{\max_{\forall i} x_i(j) - \min_{\forall i} x_i(j)}, \quad (3)$$

where $x_i(j)$ is the value of attribute j associated with instance x_i, $x'_p(j)$ is the output value of attribute j associated with instance x_p obtained following the preprocessing phase; m is the number of instances; n is the number of attributes; $i = p = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, n$.

(2) Lower-bound effectiveness measurement (smaller is better)

$$x'_p(j) = \frac{\max_{\forall i} x_i(j) - x_p(j)}{\max_{\forall i} x_i(j) - \min_{\forall i} x_i(j)}, \qquad (4)$$

where $x_i(j)$ is the value of attribute j associated with instance x_i, $x'_p(j)$ is the output value of attribute j associated with instance x_p obtained following the preprocessing phase; m is the number of instances; n is the number of attributes; $i = p = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, n$.

(3) Moderate effectiveness measurement

$$x'_p(j) = \frac{|x_p(j) - x_{\text{specified}}|}{\max_{\forall i} x_i(j) - \min_{\forall i} x_i(j)}, \qquad (5)$$

where $x_i(j)$ is the value of attribute j associated with instance x_i, $x_{\text{specified}}$ is the value specified by the system developer; $x'_p(j)$ is the output value of attribute j associated with instance x_p obtained following the preprocessing phase; m is the number of instances; n is the number of attributes; $i = p = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, n$.

Normally, upper-bound and lower-bound effectiveness measurements have similar effects on data preprocessing in the proposed approach. This work adopts the upper-bound effectiveness measurement for data preprocessing such that all numeric attributes can be transferred into values between zero and one (the lower-bound effectiveness measurement, which has similar effects on data preprocessing, can also be adopted). As for moderate effectiveness measurement, the system developer must specify (predefine) a new value. The moderate measurement is not appropriate in the proposed approach.

As stated in Section 2, the 'nearness' between two instances can be determined by appropriate similarity functions. In this work, grey relational analysis, which is more appropriate than Euclidean-like distances [18], is employed to determine the nearest neighbors of an instance with missing attribute values. Accordingly, the valid attribute values derived from these nearest neighbors are used to infer those missing values. The next section will further detail this idea.

4. A Grey-Based Nearest Neighbor Approach

This Section details a grey-based nearest neighbor algorithm for missing attribute value prediction. Accordingly, complexity analysis is presented.

4.1. A Grey-Based Nearest Neighbor Algorithm

According to the nearest neighbor rule, the similarity between an instance and its nearest neighbor—calculated from the difference between instances—is certainly maximal (that is, the difference between instances is minimal). Thus, assuming that an instance and its nearest neighbor have the same (or nearly the same) attribute values is reasonable. Restated, the value of the missing attribute of instance i can be accurately estimated by finding the known attribute value of the instance nearest to i. However, more nearest neighbors (*k-NN*) should also be considered during the estimation period to prevent sacrificing valuable information (i.e., to increase the confidence). Next, a grey-based nearest neighbor algorithm for predicting unknown attribute values is detailed.

Consider a set T of $m + 1$ instances, denoted by $T = \{x_0, x_1, x_2, \ldots, x_m\}$, where x_0 is an instance with h missing attribute values and $x_1, x_2, \ldots, x_m$ are all other known instances. Each instance x_e has n attributes and is denoted as $x_e = (x_e(1), x_e(2), \ldots, x_e(n))$. Without loss of generality, the values of numeric attributes (Notably, method for *predicting symbolic attributes* is described later) $r, r + 1, \ldots, r + h - 1$ of x_0 (i.e., $x_0(r), x_0(r + 1), \ldots, x_0(r + h - 1)$) are assumed to be unknown, where $1 < r \leqq r + h - 1 \leqq n$. The proposed predictive algorithm is then as follows.

Step 1. Calculate the grey relational coefficient (GRC) between x_0 and x_i, for $i = 1, 2, \ldots, m$ as follows.

If attribute p of each instance x_e is numeric, the value of $\text{GRC}(x_0(p), x_i(p))$ is calculated by Eq. (1).

If attribute p of each instance x_e is symbolic, the value of $\text{GRC}(x_0(p), x_i(p))$ is calculated as

$$\text{GRC}(x_0(p), x_i(p)) = 1, \quad \text{if } x_0(p) \text{ and } x_i(p) \text{ are the same.}$$

$$\text{GRC}(x_0(p), x_i(p)) = 0, \quad \text{if } x_0(p) \text{ and } x_i(p) \text{ are different.}$$

Accordingly, calculate the grey relational grade (GRG) between x_0 and x_i, for $i = 1, 2, \ldots, m$ by

Eq. (2). Notably, all attributes are available here except $r, r+1, \ldots, r+h-1$ for x_0. Besides, the class labels of instances are not used for missing attribute value prediction.

Step 2. Find the k nearest instances of x_0 based on the magnitude of GRG(x_0, x_i).

Step 3. Derive k values associated with attribute $d(r \leqq d \leqq r+h-1)$, from the above k nearest instances; that is, obtain k attribute values, say $v_{d1}, v_{d2}, \ldots, v_{dk}$.

Step 4. Predict the value of the missing attribute d of x_0 (i.e., $x_0(d)$) based on k estimated values, $p_{d1}, p_{d2}, \ldots, p_{dk}$. That is,

$$x_0(d) = p_{di}, \tag{6}$$

where $p_{di} = \frac{1}{i}\sum_{s=1}^{i} v_{ds}, \forall i \leq k$.

As stated in Section 3, GRC($x_0(p), x_i(p)$), which takes a value between zero and one, can be viewed as the similarity between $x_0(p)$ and $x_i(p)$. Here, we detail how the proposed approach is extended to deal with symbolic attributes. Similar to the methods used in [1] and [18], if x_0 and x_i have the same values for symbolic attribute p, GRC($x_0(p), x_i(p)$) will be set to one (i.e., the similarity between $x_0(p)$ and $x_i(p)$ is maximal). By contrast, if x_0 and x_i have different values for symbolic attribute p, GRC($x_0(p), x_i(p)$) will be set to zero (i.e., the similarity between $x_0(p)$ and $x_i(p)$ is minimal). Thus, the proposed approach can be applied for domains with numeric and symbolic attributes.

Using the majority voting method with the tiebreak rule [6], the proposed algorithm is also applicable for application domains in which the missing attributes are symbolic (i.e., for predicting symbolic attributes). That is, the missing attribute value of an instance is classified in the majority value of the corresponding attribute of its nearest neighbors, and a scheme is used to break a tie. Thus, the proposed approach yields a so-called *k-NN* method (k predictions are generated) to handle imperfect data problems.

As stated in Section 1, the proposed method can be viewed as an instance-based learning approach. In the case of irregular input distributions (i.e., instances close to each other probably having very different values for the same attribute), the proposed predictive method may yield inferior results, as many instance-based learning approaches. According to the nearest neighbor rule [6], the prediction errors in the proposed predictive approach can be bounded between R and 2R, where R denotes the Bayes' error (i.e., optimal error). With irregular input distributions, the prediction error, regarding to the optimal error, will inevitably be increased. In other words, predicting missing attribute values will become more difficult because of irregular input distributions.

As a result, missing attribute values can be estimated by the above *k-NN* method. Then, the usefulness of data that contain blanks is resolved. Meanwhile, an imperfect dataset can be handled as a complete dataset for classification tasks.

4.2. Complexity Analysis

Let m denote the number of compared instances and n be the number of attributes. The time complexity of calculating the GRC and the GRG is $O(mn)$. Furthermore, the total processing time also includes sorting all the grey relational grades between the referential instance and other compared instances, which in general is bounded above by $m \times \log m$.

5. Illustrative Example

This section gives an example to demonstrate the proposed predictive method. Consider a small set $\{x_0, x_1, x_2, \ldots, x_7\}$ of eight instances, as listed in Table 1. Each instance x_e is represented by five numeric attributes (A, B, C, D, E) and has already been preprocessed. Each attribute has an associated value between zero and one.

If the value of attribute A associated with instance x_0 in Table 1 (0.92) is unknown, then the proposed predictive procedure can be followed as below.

Table 1. Set of eight instances.

	Attributes				
Instances	A	B	C	D	E
x_0	0.92	0.94	0.25	0.07	0.84
x_1	0	0.17	0.81	1	0.15
x_2	0.86	1	0	0.23	1
x_3	0.23	0.21	1	0.99	0
x_4	0.85	0.82	0.21	0	0.93
x_5	1	0.88	0.14	0.14	0.87
x_6	0.96	0.95	0.09	0.13	0.85
x_7	0.18	0	0.91	0.98	0.09

First, the grey relational coefficient (GRC) and the grey relational grade (GRG) between x_0 and x_i, for $i = 1, 2, \ldots, 7$, are calculated as follows.

$$\min_{\forall j} \min_{\forall k} |x_0(k) - x_j(k)| = 0.01 \quad \text{and}$$
$$\max_{\forall j} \max_{\forall k} |x_0(k) - x_j(k)| = 0.94,$$

where $j = 1, 2, \ldots, 7$ and $k = 1, 2, \ldots, 4$.

Accordingly, the expression of the grey relational coefficient (GRC) is,

$$\begin{aligned} &\mathrm{GRC}(x_0(p), x_i(p)) \\ &= \frac{\min_{\forall j} \min_{\forall k} |x_0(k) - x_j(k)| + 0.5 \max_{\forall j} \max_{\forall k} |x_0(k) - x_j(k)|}{|x_0(p) - x_i(p)| + 0.5 \max_{\forall j} \max_{\forall k} |x_0(k) - x_j(k)|} \\ &= \frac{0.01 + 0.5 \times 0.94}{|x_0(p) - x_i(p)| + 0.5 \times 0.94}, \end{aligned}$$

where $i = 1, 2, \ldots, 7$, $j = 1, 2, \ldots, 7$, $k = 1, 2, \ldots, 4$ and $p = 1, 2, \ldots, 4$.

And the expression of grey relational grade (GRG) is,

$$\mathrm{GRG}(x_0, x_i) = \frac{1}{4} \sum_{k=1}^{4} \mathrm{GRC}(x_0(k), x_i(k)),$$

where $i = 1, 2, \ldots, 7$.

Accordingly, $\mathrm{GRG}(x_0, x_1) = 0.4024$, $\mathrm{GRG}(x_0, x_2) = 0.7740$, $\mathrm{GRG}(x_0, x_3) = 0.3763$, $\mathrm{GRG}(x_0, x_4) = 0.8752$, $\mathrm{GRG}(x_0, x_5) = 0.8955$, $\mathrm{GRG}(x_0, x_6) = 0.9169$, and $\mathrm{GRG}(x_0, x_7) = 0.3766$, respectively.

According to the following expression,

$$\begin{aligned} &\mathrm{GRG}(x_0, x_6) > \mathrm{GRG}(x_0, x_5) > \mathrm{GRG}(x_0, x_4) \\ &\quad > \mathrm{GRG}(x_0, x_2) > \mathrm{GRG}(x_0, x_1) > \mathrm{GRG}(x_0, x_7) \\ &\quad > \mathrm{GRG}(x_0, x_3), \end{aligned}$$

the four nearest neighbors (NNs) of instance x_0, for example, can be determined. Consequently, instances x_6, x_5, x_4, and x_2 are, respectively, the 1-*NN*, 2-*NN*, 3-*NN*, and 4-*NN* of instance x_0. Here, four attribute values, 0.96, 1, 0.85, and 0.86, are respectively derived from instances x_6, x_5, x_4, and x_2.

Eventually, four estimated values (mean values),

$$\begin{aligned} &0.96, \\ &(0.96 + 1)/2 = 0.98, \\ &(0.96 + 1 + 0.85)/3 = 0.9367, \quad \text{and} \\ &(0.96 + 1 + 0.85 + 0.86)/4 = 0.9175 \end{aligned}$$

can be used to predict the value of the missing attribute of instance $x_0(0.92)$, and their prediction errors are,

$$\begin{aligned} |0.96 - 0.92| &= 0.04, \\ |0.98 - 0.92| &= 0.06, \\ |0.9367 - 0.92| &= 0.0167, \ \text{and} \\ |0.9175 - 0.92| &= 0.0025, \ \text{respectively.} \end{aligned}$$

6. Experimental Results

This section describes experiments on two datasets and five classification tasks to demonstrate the effectiveness of the proposed approach.

6.1. Experimental Evaluation of Prediction Accuracy

First, the proposed predictive approach was evaluated on Fisher's Iris dataset [12], which contains 150 instances, to demonstrate the approach's effectiveness. All instances are equally divided into three classes: *Setosa*, *Versicolor*, and *Virginica*. Each instance is described by four attributes: *Sepal Width* (SW), *Sepal Length* (SL), *Petal Width* (PW), and *Petal Length* (PL). In the experiments, each instance was preprocessed by the upper-bound effectiveness measurement (see Section 3) and each attribute took values between zero and one. Furthermore, the number of nearest neighbors, k, chosen in Step 2 was assumed to vary from 1 to 50.

A method called *leave-one-out cross-validation* [34] was adopted in each experiment. That is, the value of the missing attribute of instance i is predicted by all of the instances except instance i itself. Thus, for every missing value prediction, nearly all of the instances are selected as compared instances. In each run, the accuracy of prediction was measured using the Root Mean Square Error (RMSE), as follows.

$$\mathrm{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (e_i - \tilde{e}_i)^2}, \tag{7}$$

where e_i is the original attribute value; $\tilde{e}_i$ is the estimated attribute value, and m is the total number of predictions.

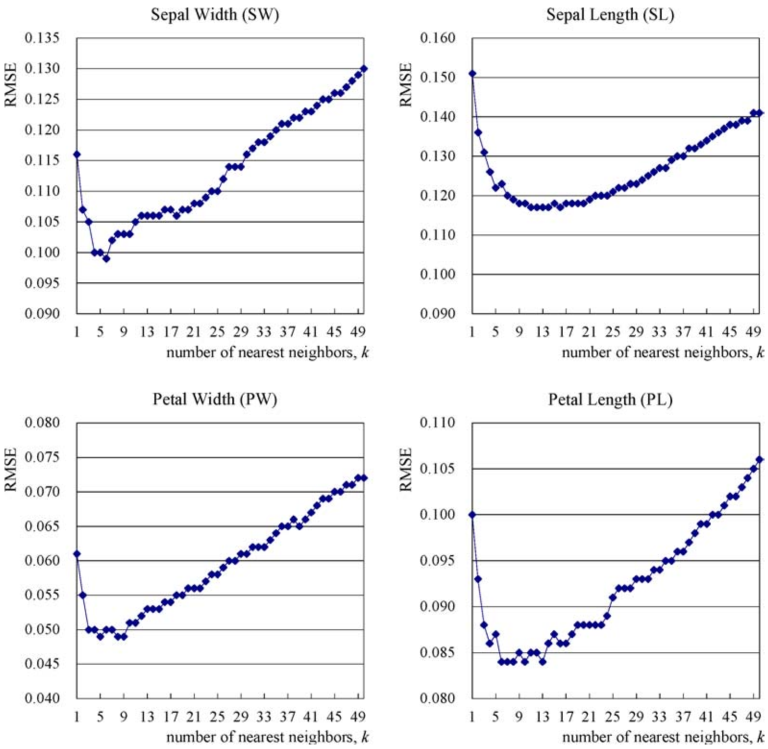


Figure 1. Experimental results on the Iris dataset with four attributes.

Figure 1 presents the experimental results for all four attributes of the Iris dataset. The best choices of k (number of nearest neighbors) for attributes SW, SL, PW, and PL are 6, 13, 5, and 10, respectively (these values of k can be determined by the cross-validation technique). Although the 1-*NN* method is not quite ideal, it still yields acceptable results.

Also, the proposed predictive approach was evaluated on the Liver-disorders dataset [3], which contains 345 instances with six attributes, to demonstrate the approach's effectiveness. Figure 2 presents the experimental results for all six attributes of the Liver-disorders dataset. The best choices of k (number of nearest neighbors) for attributes *mcv*, *alkphos*, *sgpt*, *sgot*, *gammagt* and *drinks* are 41, 50, 4, 4, 27 and 17, respectively.

Tables 2 and 3 compare the accuracy of the proposed predictive method with that of *multiple imputation* [21] and that of *mean substitution* on Fisher's Iris dataset and the Liver-disorders dataset, respectively. In multiple imputation, a statistical model (imputation-posterior and EM algorithm) is required to compute five (default) imputations (estimated values) for each missing value in a dataset (that is, to create predictions for the *distributions* of each missing value [21]). This approach assumes that data are missing at random. In mean substitution, the missing attribute value is directly substituted for the average known values. Meanwhile, significant tests (paired t-tests [33]), as shown in Tables 4 and 5, are also done for the comparisons. The prediction error differences were evaluated using paired t-tests with significance level set to 99% (**) and 95% (*). The approach presented here can be easily seen to outperform both multiple imputation and mean substitution.

Table 2. A comparison with multiple imputation and mean substitution for the Iris domain.

	Accuracy (RMSE)			
Methods	SW	SL	PW	PL
Our approach (Minimum)	0.0994	0.1167	0.0491	0.0837
Our approach (Average)	0.1137	0.1264	0.0595	0.0924
Our approach (Maximum)	0.1301	0.1508	0.0723	0.1064
Multiple imputation (Minimum)	0.1193	0.1649	0.0742	0.1027
Multiple imputation (Average)	0.1261	0.1765	0.0795	0.1141
Multiple imputation (Maximum)	0.1322	0.1858	0.0901	0.1211
Mean substitution	0.2308	0.1813	0.3001	0.3190

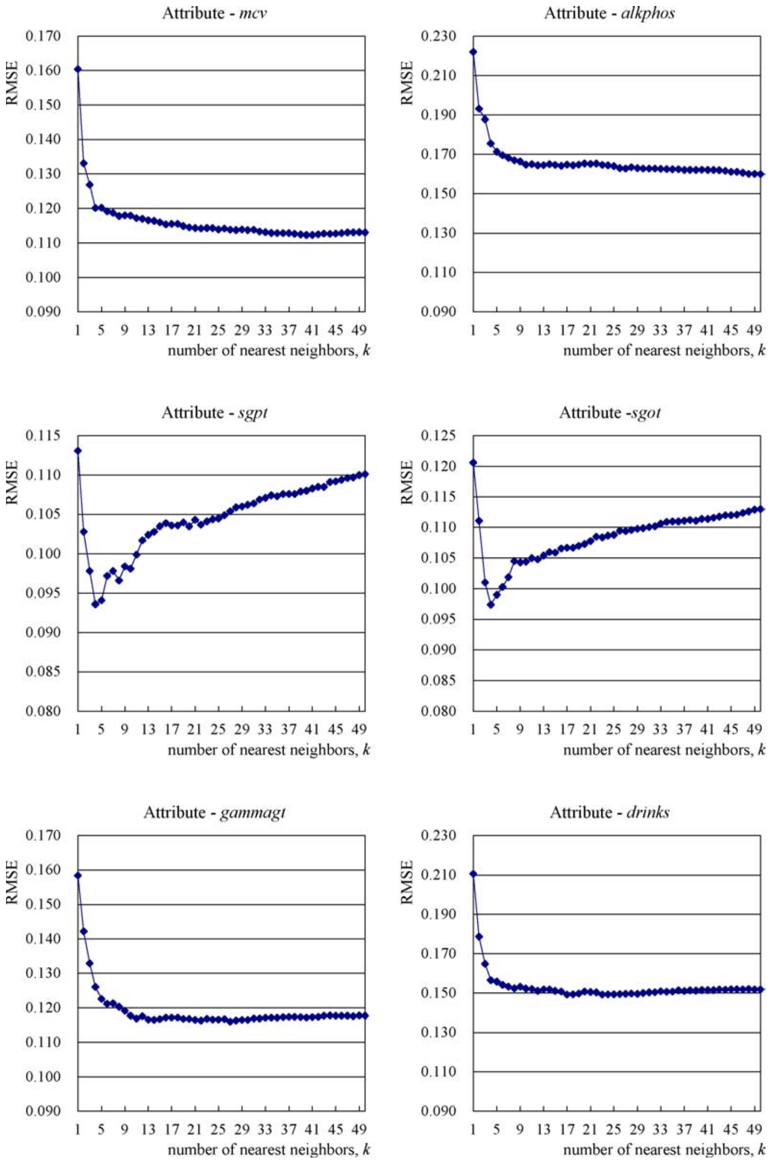


Figure 2. Experimental results on the Liver-disorders dataset with six attributes.

6.2. Experimental Evaluation of Classification Accuracy

After predicting (estimating) missing attribute values with high accuracy, an imperfect dataset can be handled as a complete dataset in classification tasks. Restated, in a given dataset, instances with missing attribute values, which would be repaired after the estimation period, can be treated as complete training instances in the training phase or as complete testing instances in the testing phase.

The proposed method was evaluated using five application domains with incomplete data, which are described below—to demonstrate that the grey-based predictive approach can handle incomplete data problems in classification tasks and maintain or even increase the accuracy of classification.

6.2.1. The Hepatitis Diagnosis Problem. The hepatitis diagnosis problem, which was taken from Carnegie-Mellon University [5], contains 155 instances, where each instance consists of 19 attributes and a class label.

Table 3. A comparison with multiple imputation and mean substitution for the liver-disorders domain.

	Accuracy (RMSE)					
Methods	*mcv*	*alkphos*	*sgpt*	*sgot*	*gammagt*	*drinks*
Our approach (Minimum)	0.1123	0.1599	0.0936	0.0974	0.1160	0.1493
Our approach (Average)	0.1162	0.1661	0.1047	0.1085	0.1193	0.1533
Our approach (Maximum)	0.1604	0.2220	0.1131	0.1206	0.1584	0.2106
Multiple imputation (Minimum)	0.1201	0.1713	0.1028	0.1111	0.1226	0.1558
Multiple imputation (Average)	0.1321	0.1900	0.1091	0.1140	0.1364	0.1733
Multiple imputation (Maximum)	0.1734	0.2439	0.1204	0.1225	0.1780	0.2335
Mean substitution	0.1778	0.2423	0.1963	0.1985	0.2042	0.2535

Table 4. Significant tests for the Iris domain. The prediction error differences were evaluated using paired t-tests with significance level set to 99% (**) and 95% (*).

	SW	SL	PW	PL
Compare our approach with multiple imputation	**	**	**	**
Compare our approach with mean substitution	**	**	**	**

5.67% (167/2,945) of all attribute values (155*19 = 2, 945) were unknown. This is a two-class classification problem with numeric and symbolic attributes.

6.2.2. The Bridges Domain. The bridges domain [28] contains 108 instances, where each instance consists of 13 attributes and a class label. 5.56% (78/1,404) of all attribute values (108*13 = 1, 404) were unknown. In the experiments, attribute ‘TYPE’ was used as the class label. This is a multiple-class classification problem with numeric and symbolic attributes.

6.2.3. The Echocardiogram Domain. The echocardiogram domain [30] contains 132 instances, where each instance consists of 13 attributes and a class label. 7.69% (132/1,716) of all attribute values (132*13 = 1, 716) were unknown. The goal of this classification project is to determine whether the patient was alive at one year. This is a two-class classification problem with numeric and symbolic attributes.

6.2.4. The Water-Treatment Domain. The water-treatment domain [3] contains 527 instances, where each instance consists of 38 attributes and a class label. 2.95% (591/20,026) of all attribute values (527*38 = 20, 026) were unknown. This is a multiple-class classification problem with numeric attributes.

6.2.5. The Soybean Domain. The soybean domain [24] contains 307 instances, where each instance consists of 35 attributes and a class label. 6.63% (712/10,745) of all attribute values (307*35 = 10, 745) were unknown. This is a multiple-class classification problem with symbolic attributes.

The experiments were divided into two parts for each application domain. (a) The proposed predictive method was not applied for missing attribute prediction. (b) The proposed predictive method was applied for missing attribute prediction. Consequently, for each application domain underlying different classification algorithms, a comparison can be done for demonstrating if the accuracy of classification is maintained or even increased by applying the proposed method for missing attribute prediction beforehand. As stated earlier, the class labels of instances are not used for missing attribute value prediction. There might be another attribute that is highly correlated with the class, functioning as a de facto proxy for the class. This situation may occur in the proposed approach, as well as many instance-based learning methods. However, in

Table 5. Significant tests for the liver-disorders domain. The prediction error differences were evaluated using paired t-tests with significance level set to 99% (**) and 95% (*).

	mcv	*alkphos*	*sgpt*	*sgot*	*gammagt*	*drinks*
Compare our approach with multiple imputation	**	*	**	**	**	*
Compare our approach with mean substitution	**	**	**	**	**	**

the proposed approach, all attributes should be considered to determine the nearest neighbors of instances (the class labels are not used here). In other words, the proposed approach concentrates on the differences among instances regardless of whether another attribute is highly correlated with the class. Accordingly, the accuracy of classification corresponding to a particular classification algorithm was measured. Various classification algorithms were investigated for comparison, including Decision Stump [34], Decision Table [22], HyperPipes [34], IB1 [1], KNN [1], Decision Tree [27], Kernel Density [34], Kstar [20], Logistic [34], NaiveBayes [19], 1R* [17], Voted Perception [15] and ADTree [14]. Notably, when the proposed predictive method was not applied, missing attribute values of each application domain would be handled by the original method of each classification algorithm. For example, Decision Stump [34] deals with missing values by extending a third branch from its one-level decision tree (That is, by treating a missing attribute as a separate value).

Of these classification algorithms, Logistic and ADTree are generally suited only to two-class classification problems, whereas others are suited to both two-class and multiple-class classification problems.

For each experiment, *ten-fold cross-validation* [32] was performed ten times to determine the accuracy of classification. That is, the entire dataset in a classification task was randomly split into ten parts, where each part was treated as the testing instances and all others were treated as the training instances. Thus, the average accuracy of classification was determined from the ten sub-experiments. Meanwhile, the best choice of k (number of nearest neighbors), as stated in the previous section, was determined by the cross-validation technique.

Tables 6 to 10, one for each of the five application domains, compare the accuracy of classification that underlie various classification algorithms (significant tests—paired t-tests [33]—are also done for the comparisons). For example, when the decision tree was employed to solve the hepatitis diagnosis domain (in Table 6), the accuracy of classification was increased from 78.71% (the proposed predictive method was not applied) to 83.23% (the proposed predictive method was applied).

Furthermore, we also conducted experiments in which all instances with missing attribute values of a specific domain were used as testing instances and others were used as training instances. These experimental results are shown in Tables 11 to 15. The accuracy of

Table 6. A comparison (ten-fold cross-validation for the entire dataset) when the proposed predictive method was applied or not underlie different classification algorithms for the Hepatitis Diagnosis domain. An asterisk (*) indicates a significant difference at the 95% confidence level, using a paired t-test.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	74.19	77.42	63.23	80.00	82.58	78.71	78.71	79.35	83.87	81.29	80.00	83.87	79.35
Our method was applied	**81.29***	**83.87***	**65.16**	80.00	**84.52***	**83.23***	80.00	**83.87***	**85.16**	**82.58**	80.00	**85.81***	**84.52***

Table 7. A comparison (ten-fold cross-validation for the entire dataset) when the proposed predictive method was applied or not underlie different classification algorithms for the Bridges domain. An asterisk (*) indicates a significant difference at the 95% confidence level, using a paired t-test.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	57.14	60.95	50.48	60.00	60.95	71.43	60.00	63.81	N/A	68.57	55.24	N/A	N/A
Our method was applied	57.14	60.95	50.48	**62.86***	**64.76***	**74.29***	**63.81***	63.81	N/A	**72.38***	**58.10***	N/A	N/A

Table 8. A comparison (ten-fold cross-validation for the entire dataset) when the proposed predictive method was applied or not underlie different classification algorithms for the Echocardiogram domain. An asterisk (*) indicates a significant difference at the 95% confidence level, using a paired *t*-test.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	95.95	95.95	87.84	87.84	91.89	95.95	93.24	91.89	94.59	95.95	95.95	91.89	97.30
Our method was applied	**97.30***	**97.30***	**89.19**	**94.59***	**93.24**	**97.30***	**95.95***	91.89	**100.00***	95.95	**97.30***	91.89	97.30

Table 9. A comparison (ten-fold cross-validation for the entire dataset) when the proposed predictive method was applied or not underlie different classification algorithms for the Water-Treatment domain. An asterisk (*) indicates a significant difference at the 95% confidence level, using a paired *t*-test.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	55.03	66.41	61.86	68.88	69.45	70.40	74.00	70.40	N/A	75.52	59.58	N/A	N/A
Our method was applied	**62.24***	**70.21***	**62.24**	**73.62***	**74.38***	**71.73**	**74.19**	**70.59**	N/A	**75.90**	**60.91**	N/A	N/A

Table 10. A comparison (ten-fold cross-validation for the entire dataset) when the proposed predictive method was applied or not underlie different classification algorithms for the Soybean domain. An asterisk (*) indicates a significant difference at the 95% confidence level, using a paired *t*-test.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	26.06	81.76	89.58	90.88	75.90	85.02	91.86	91.33	N/A	90.55	39.41	N/A	N/A
Our method was applied	26.06	**83.39***	**89.90**	90.88	**76.87**	**88.93***	91.86	**92.51***	N/A	90.55	39.41	N/A	N/A

Table 11. A comparison (all instances with missing attribute values were used as testing instances and others were used as training instances) when the proposed predictive method was applied or not underlie different classification algorithms for the Hepatitis Diagnosis domain.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	78.67	77.33	58.67	80.00	80.00	77.33	80.00	78.67	70.67	84.00	24.00	74.67	65.33
Our method was applied	**89.33**	77.33	**61.33**	**82.67**	**82.67**	**78.67**	**82.67**	**81.33**	**76.00**	84.00	**77.33**	74.67	**80.00**

Table 12. A comparison (all instances with missing attribute values were used as testing instances and others were used as training instances) when the proposed predictive method was applied or not underlie different classification algorithms for the bridges domain.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	80.00	68.57	57.14	62.86	62.86	54.29	62.84	71.43	N/A	71.43	62.86	N/A	N/A
Our method was applied	80.00	68.57	**60.00**	**77.14**	**77.14**	**57.14**	**77.14**	**77.14**	N/A	**77.14**	62.86	N/A	N/A

Table 13. A comparison (all instances with missing attribute values were used as testing instances and others were used as training instances) when the proposed predictive method was applied or not underlie different classification algorithms for the Echocardiogram domain.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	78.57	78.57	88.57	77.14	77.14	71.43	78.57	77.14	72.86	78.57	78.57	77.14	60
Our method was applied	**85.71**	**85.71**	**91.43**	**80.00**	**80.00**	**81.43**	**81.43**	**82.86**	**78.57**	**81.43**	**85.71**	**84.29**	**81.43**

Table 14. A comparison (all instances with missing attribute values were used as testing instances and others were used as training instances) when the proposed predictive method was applied or not underlie different classification algorithms for the Water-Treatment domain.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	61.22	63.27	62.59	55.78	55.78	67.35	71.43	72.11	N/A	73.47	59.86	N/A	N/A
Our method was applied	**63.27**	**71.43**	**63.27**	**70.75**	**70.75**	67.35	**72.79**	72.11	N/A	73.47	**61.90**	N/A	N/A

Table 15. A comparison (all instances with missing attribute values were used as testing instances and others were used as training instances) when the proposed predictive method was applied or not underlie different classification algorithms for the Soybean domain.

	Accuracy of classification (%) underlie a particular classification algorithm												
Methods	Decision Stump	Decision Table	Hyper Pipes	IB1	KNN	Decision Tree	Kernel Density	Kstar	Logistic	Naive Bayes	1R*	Voted Perception	ADTree
Our method was not applied	9.76	53.69	9.76	58.54	56.10	9.76	58.54	53.66	N/A	51.22	2.44	N/A	N/A
Our method was applied	9.76	**58.54**	**56.10**	58.54	**58.54**	**58.54**	58.54	**58.54**	N/A	**58.54**	**58.54**	N/A	N/A

classification in classification tasks can be easily seen to be maintained or increased, if the proposed predictive method for handling missing attribute values was followed beforehand.

In summary, missing attribute values can be easily and accurately estimated using the proposed predictive method. Consequently, the usefulness of data that contain blanks is resolved. Meanwhile, an imperfect dataset can be handled as a complete dataset and the classification errors in classification tasks can be reduced.

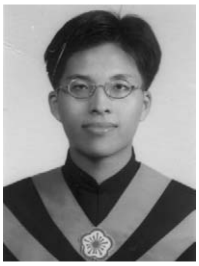
7. Conclusions

This paper has proposed a grey-based nearest neighbor approach to handle incomplete data problems. Here, grey relational analysis is employed to determine the nearest neighbors of an instance with missing attribute values. Thus, the valid attribute values derived from these nearest neighbors are used to predict those missing values. Experimental results indicated that the proposed approach outperforms other alternative methods. Furthermore, the proposed method was evaluated using five classification problems with incomplete data. Experimental results revealed that the accuracy of classification can be maintained or even increased by applying the proposed method beforehand.

References

1. D.W. Aha, D. Kibler, and M.K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.
2. M.R. Berthold and K.-P. Huber, "Missing values and learning of fuzzy Rules," *Int. J. Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 6, no. 2, 1998.
3. C.L. Blake and C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.
4. W.L. Buntine and A.S. Weigend, "Bayesian backpropagation," *Complex Systems*, vol. 5, pp. 603–643, 1991.
5. B. Cestnik, I. Kononenko, and I. Bratko, "Assistant 86: A knowledge-elicitation tool for sophisticated users," in *Progress in Machine Learning*, edited by I. Bratko and N. Lavrac, Sigma Press: Wilmslow, 1987, pp. 31–45.
6. T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
7. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, series B, vol. 39, pp. 1–38, 1977.
8. J. Deng, "The theory and method of socioeconomic grey systems," *Social Sciences in China*, vol. 6, pp. 47–60, 1984 (in Chinese).
9. J. Deng, "Introduction to grey system theory," *The Journal of Grey System*, vol. 1, pp. 1–24, 1989.
10. J. Deng, "Grey information space," *The Journal of Grey System*, vol. 1, pp. 103–117, 1989.
11. J.K. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 10, pp. 617–621, 1979.
12. R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics* Part 2, vol. 7, pp. 179–188, 1936.
13. E. Fix and J.L. Hodges, "Discriminatory analysis: Nonparametric discrimination: consistency properties," Technical Report Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
14. Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *Proc. of the 16th International Conference on Machine Learning*, Bled, Slovenia, 1999, pp. 124–133.
15. Y. Freund and R.E. Schapire, "Large margin classification using the perceptron algorithm," in *Proc. 11th Annual Conf. on Comput. Learning Theory*, ACM Press: New York, NY, 1998, pp. 209–217.
16. J.H. Friedman, "A recursive partitioning decision rule for nonparametric classification," *IEEE Transactions on Computers*, pp. 404–408, 1977.
17. R.C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.
18. C.C. Huang and H.M. Lee, "An instance-based learning approach based on grey relational structure," in *Proc. of the UK Workshop on Computational Intelligence (UKCI-02)*, Birmingham, Sept., 2002.
19. G.H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proc. of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 338–345.
20. G.C. John and E.T. Leonard, "K*: An instance-based learner using an entropic distance measure," in *Proc. of the 12th International Conference on Machine Learning*, 1995, pp. 108–114.
21. G. King, J. Honaker, A. Joseph, and K. Scheve, "Analyzing incomplete political science data: An alternative algorithm for multiple imputation," *American Political Science Review*, vol. 95, no. 1, pp. 49–69, 2001.
22. R. Kohavi, "The power of decision tables," in *European Conference on Machine Learning*, 1995.
23. C.T. Lin and S.Y. Yang, "Selection of home mortgage loans using grey relational analysis," *The Journal of Grey System*, vol. 4, pp. 359–368, 1999.
24. R.S. Michalski and R.L. Chilausky, "Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis," *International Journal of Policy Analysis and Information Systems*, vol. 4, no. 2, 1980.
25. J.R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.
26. J.R. Quinlan, "Unknown attribute values in induction," in *Proc. of the Sixth International Machine Learning Workshop*, Morgan Kaufmann: San Mateo, CA, 1989, pp. 164–168.
27. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers: San Mateo, CA, 1993.
28. Y. Reich, "Converging to 'ideal' design knowledge by learning," in *Proc. of the First International Workshop on Formal Methods in Engineering Design*, 1990, pp. 330–349.
29. D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley: New York, 1987.
30. S. Salzberg, "Exemplar-based learning: Theory and implementation," Technical Report TR-10-88, Harvard University, Center for Research in Computing Technology, 1988.
31. C. Stanfill and D. Waltz, "Towards memory-based reasoning," *Communications of the ACM*, vol. 29, no. 12, pp. 1213–1228, 1986.

32. M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, vol. B, 36, pp. 111–147, 1974.
33. C.J. Watson, P. Billingsley, D.J. Croft and D.V. Huntsberger, *Statistics for Management and Economics*, 5th edition, Allyn and Bacon, Boston, 1993.
34. I. Witten and E. Frank, *Data Mining—Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann: San Francisco, CA, 2000.
35. J.H. Wu, M.L. You, and K.L. Wen, "A modified grey relational analysis," *The Journal of Grey System*, vol. 3, pp. 287–292, 1999.



Chi-Chun Huang received the M.S. degree from the Department of Information Management at National Central University, Taipei, Taiwan. He is currently pursuing a Ph.D. degree in Electronic Engineering at National Taiwan University of Science and Technology, Taipei, Taiwan. His research includes intelligent Internet systems, grey theory, neural networks and pattern recognition.

Hahn-Ming Lee is currently Professor and Chairman in the Department of Computer Science and Information Engineering at National Taiwan University of Science and Technology, Taipei, Taiwan. He received the B.S. degree and Ph.D. degree from the Department of Computer Science and Information Engineering at National Taiwan University in 1984 and 1991, respectively. His research interests include, intelligent Internet systems, fuzzy computing, neural networks and machine learning. He is a member of IEEE, TAAI, CFSA and IICM.