A grey-based nearest neighbor approach for predicting missing attribute values

Chi-Chun Huang[1] and Hahn-Ming Lee[2]
[1]Department of Electronic Engineering
[2]Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei 106, Taiwan

Abstract

In this paper, we propose a grey-based nearest neighbor approach to predict missing attribute values in an accurate manner. First, the nearest neighbors of an instance with missing attribute values are found through grey relational analysis. Accordingly, the known attribute values derived from these nearest neighbors are chosen to infer those missing. The Iris flower dataset was used to demonstrate the performance of the proposed approach. Experimental results show that our method performs better than both multiple imputation and mean substitution.

***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 methods are usually designed to handle perfect data. However, real-world classification tasks often involve incomplete data, i.e., data contain some missing attribute values (or blanks). In fact, incomplete information can be caused by error, equipment failure, change of plans, etc [6]. Owing to the difficulty with missing attribute values, most learning algorithms are not well adapted to some application domains (e.g., Web applications containing a lot of 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. Different issues related to missing attribute values can then be briefly described as follows. First, missing attribute values usually appear in the training set. This seems to imply that, in the training phase, a reliable method for dealing with those missing is frequently necessary. Another major concern is how to classify a new, unseen instance that has an incomplete feature vector [13]. Furthermore, in order to resolve the usefulness of data containing blanks and reduce the classification errors in the learning system, the system developer has to concentrate on estimating missing attribute values as accurately as possible.

In general, incomplete data greatly affect the performance of classification algorithms. That is, a robust and effective approach for handling incomplete data in classification tasks is very important. Both Friedman [9] and Quinlan [12] adopted a common strategy, ignoring blanks, to tackle problems with unknown attribute values during training. Nevertheless, this method is not applicable when numerous training instances contain blanks and may yield inferior performance [13]. An alternative way is throwing away instances with missing attribute values in the training phase, but it probably results in the loss of valuable information. In the machine learning literature, several techniques for estimating missing attribute values have been proposed, including Expectation-Maximization (EM) principle [4], decision tree induction [12], Bayesian approach [2], and multiple imputation [10,14]. Most of these methods are quite complicated and time consuming, even though they have been used to deal with different incomplete-data problems.

In this paper, we propose a grey-based nearest neighbor method to predict missing attribute values in an easy and accurate manner. The nearest neighbor concept [3,8] and grey relational analysis [5] play principal roles in method development. Given a set of instances, the difference between an instance and its nearest neighbor is certainly minimal. Thus, it is reasonable to assume that an instance containing blanks and its nearest neighbor would have the same (or nearly the same) attribute values. Here, the known attribute values, derived from the nearest neighbors of an instance with missing attribute values, are chosen to infer those missing. Generally, similarity functions such as Euclidean distance are used to determine the 'nearness' (or relationship) between two instances. However, Euclidean-like distances are mainly suitable for domains with numeric attributes. In order to overcome this shortcoming, the above nearest neighbors are found through grey relational analysis, which is appropriate for both symbolic and numeric attributes and provides whole relational orders (wholeness [16]) for the entire relational space. The Iris flower dataset was used to demonstrate the performance of the proposed method. Experimental results show that our approach reveals its superiority.

The rest of this paper is organized as follows. We review the nearest neighbor concept and grey relational analysis in Sections 2 and 3, respectively. In Section 4 we propose a grey-based nearest neighbor algorithm for predicting missing attribute values. In Section 5 an example is given to illustrate the proposed predicting approach. In Section 6 experiments on the Iris flower dataset are reported. Finally, we conclude in Section 7.

2. The nearest neighbor concept

In this section, the nearest neighbor concept we adopt for predicting missing attribute values is reviewed.

In learning from examples, proper decisions (e.g., classification, prediction) for a new instance i can be made by using information extracted from a set of training instances, in particular from the nearest instances of i. For example, we might estimate a new employee's salary by using that of another employee who has similar educations, work experiences, etc. Generally the 'nearness' between two instances is determined by some similarity functions, e.g., Euclidean metric. Based on the concept of 'nearest neighbor', many learning algorithms have been investigated, such as instance-based learning [1] and memory-based reasoning [15].

Since its inception in 1957 [8], the *nearest neighbor (NN) rule* [3] built from the above concept has been successfully applied to a wide variety of application domains. This simple principle can be stated as follows. Given a set of training instances, an unseen instance is classified according to the training instance which is the nearest. An extended version, called *majority voting* or *k-NN rule*, classifies the unseen instance in the majority classification of its k nearest neighbors.

The *NN* rule has many advantages over other classification methods. For example, it is fairly straightforward to understand and easy to implement. In addition, Cover and Hart [3] have shown 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 probability of error.

3. Grey relational analysis

In 1984, Deng [5] proposed a measurement method, called *grey relational analysis* (GRA), to determine the relationships among a referential observation and the compared observations by calculating the *grey relational coefficient* (GRC) and the *grey relational grade* (GRG). Assume that we have 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 $x_e = (x_e(1), x_e(2), \ldots, x_e(n))$. The grey relational coefficient can then be obtained as follows.

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

where $\zeta \in [0,1]$ (Usually, let $\zeta = 0.5$), $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, m$, $k = 1, 2, \ldots, n$ and $p = 1, 2, \ldots, n$.

From above, the grey relational grade is expressed as follows.

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

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

Obviously, the *GRG* takes values ranging from 0 to 1. The significant effect of grey relational analysis can be described as follows.

If $GRG(x_0, x_1)$ is larger than $GRG(x_0, x_2)$, for example, 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 meets four principal axioms [16], including

1) Normality
   $0 < GRG(x_0, x_i) \le 1, \forall i$
2) Dual Symmetry
   If there are only two observations (i.e., x_0 and x_1) in the relational space, then
   $GRG(x_0, x_1) = GRG(x_1, x_0)$
3) Wholeness
   If there are three or more observations in the relational space, then
   $GRG(x_0, x_i) \overset{often}{\neq} GRG(x_i, x_0), \forall i$
4) Approachability
   $GRG(x_0, x_i)$ decreases along with $|x_0(p) - x_i(p)|$ increasing.

Based on these axioms, grey relational analysis has some benefits. For example, it provides a normalized measuring function (Normality) to analyze the relational structure. Also, it yields whole relational orders (wholeness) for the entire relational space and is appropriate for both symbolic and numeric attributes.

Before calculating the grey relational coefficient and the grey relational grade, one of the following methods should be used for data preprocessing [11]:

1) Upper-bound effectiveness measuring (i.e. large-the-better)

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

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

2) Lower-bound effectiveness measuring (i.e. small-the-better)

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

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

3) Moderate effectiveness measuring (i.e. normal-the-better)

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

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

Usually, upper-bound effectiveness measurement and lower-bound effectiveness measurement would achieve similar effects. As for moderate effectiveness measurement, the system developer has to specify a new value. In this paper, upper-bound effectiveness measurement was adopted for data preprocessing.

As mentioned in Section 2, the 'nearness' between two instances can be determined by some appropriate similarity functions. In this paper, the nearest neighbors of an instance with missing attribute values are found by using grey relational analysis, instead of calculating the Euclidean distance, which is mainly suitable for domains with numeric attributes. Consequently, the valid attribute values derived from these nearest neighbors are used to infer those missing. In the next section, we will discuss this idea in more detail.

4. A grey-based nearest neighbor approach

Given a set of instances, the difference between an instance and its nearest neighbor is certainly minimal. Thus, it is reasonable to assume that an instance

containing blanks and its nearest neighbor would have the same (or nearly the same) attribute values. In other words, the value of missing attribute of instance *i* could be accurately estimated by finding the known attribute value of the nearest instance of *i*. However, in order to avoid sacrificing valuable information, more nearest neighbors (*k-NN*) should also be taken into consideration during the estimation period.

Next we detail a grey-based nearest neighbor algorithm for predicting unknown attribute values. Restated, the nearest neighbors of an instance, which are chosen to infer missing attribute values, are found through grey relational analysis.

Assume that we have 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 we may assume that the values of numeric attributes $r, r+1, \ldots, r+h-1$ of x_0 (i.e., $x_0(r), x_0(r+1), \ldots, x_0(r+h-1)$) are unknown, where $1 \leqq r \leqq r+h-1 \leqq n$. The proposed predicting algorithm can then be stated below.

Step1. Calculate the grey relational coefficient (GRC) and the grey relational grade (GRG) between x_0 and x_i, for i = 1, 2, …, m. Notice that all attributes are available here except attributes r, $r+1, \ldots, r+h-1$.

Step2. Find k nearest instances of x_0 based on the magnitude of $GRG(x_0, x_i)$, where i = 1, 2, …, m and $k \leqq m$.

Step3. Derive k values associated with attribute d ($r \leqq d \leqq r+h-1$), respectively, from the above k nearest instances, i.e., k attribute values, say $v_{d1}, v_{d2}, \ldots, v_{dk}$, can be obtained.

Step4. Predict the value of 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,

$$x_0(d) = p_{di},$$

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

By using the majority voting method with tiebreak rule [3], the proposed algorithm is also suitable for application domains in which the missing attributes are symbolic. Thus, the proposed approach yields a so-called *k-NN* method (*k* estimations are generated) to cope with imperfect-data problems.

Let m denote the number of compared instances and n denote 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 among the referential instance and other compared instances, which in general is bounded above by $m \times \log m$.

5. An example

In this section, an example is given to illustrate the proposed predicting approach. Assume that we have a small set $\{x_0, x_1, x_2, \ldots, x_7\}$ of eight instances, as shown in Table 1. Each instance x_e is represented by five attributes (A, B, C, D, E) and has already been preprocessed. Each attribute has an associated value ranging from 0 to 1.

Table 1
Set of eight instances

Instance	Attributes				
	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

If the value of attribute A associated with instance x_0 in Table 1 (i.e., 0.92) is missing, then the proposed predicting procedure can be performed as below.

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

Here, we have

$$\min_{\forall j} \min_{\forall k} \left| x_0(k) - x_j(k) \right| = 0.01 \quad \text{and}$$

$$\max_{\forall j} \max_{\forall k} \left| x_0(k) - x_j(k) \right| = 0.94,$$

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

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

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

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

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

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

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

Based on the following expression

$$GRG(x_0, x_6) > GRG(x_0, x_5) > GRG(x_0, x_4) > GRG(x_0, x_2) > GRG(x_0, x_1) > GRG(x_0, x_7) > GRG(x_0, x_3),$$

four nearest neighbors (*NN*s) of instance x_0, for example, could be found. As a result, 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, we derive four attribute values, 0.96, 1, 0.85, and 0.86, respectively from instances x_6, x_5, x_4, and x_2.

Eventually, we choose four estimated values (average values),

0.96,

(0.96+1)/2 = 0.98,

(0.96+1+0.85)/3 = 0.9367, and

(0.96+1+0.85+0.86)/4 = 0.9175

to predict the value of the missing attribute of instance x_0 (i.e., 0.92).

As a result, the prediction errors are, respectively,

0.96-0.92 = 0.04,

0.98-0.92 = 0.06,

0.9367-0.92 = 0.0167, and

0.9175-0.92 = -0.0025.

6. Experimental results

To demonstrate the effectiveness of the proposed predicting approach, we evaluated it on Fisher's Iris dataset [7], which contains 150 instances. All instances are divided equally 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 had already been preprocessed by upper-bound effectiveness measurement (see Section 3) and each attribute took values ranging from 0 to 1. In addition, we assumed that the number of nearest neighbors, *k* chosen in Step 2 varied from 1 to 50.

For each experiment, a method called *leave-one-out cross-validation* was adopted. That is, the value of missing attribute of instance *i* was predicted by all of the instances except instance *i* itself. Therefore, for every missing value prediction, nearly all of the instances were selected as the compared instances. In each run, the prediction accuracy was measured by using the Root Mean Square Error (RMSE), which is expressed as follows.

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

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

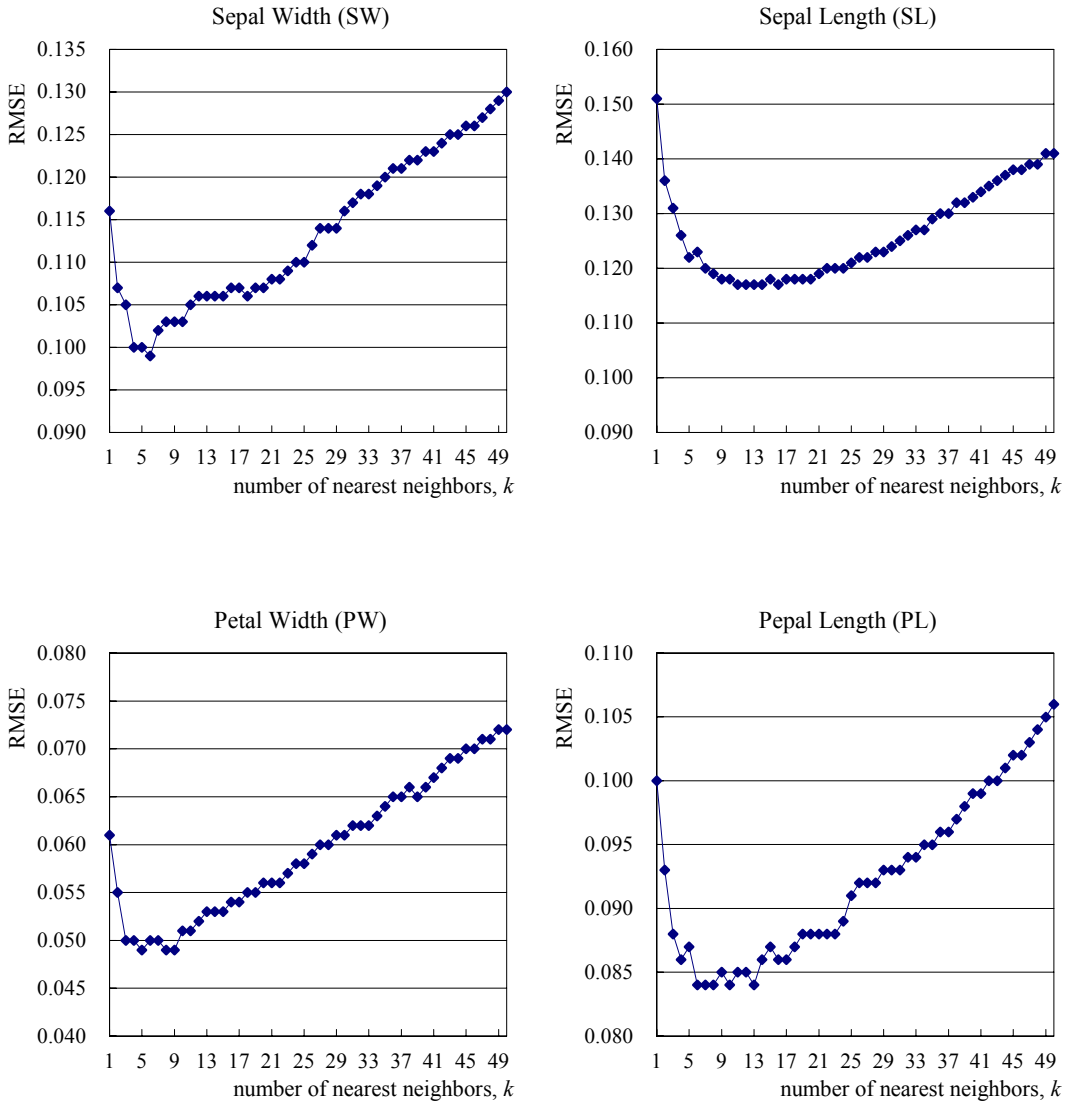

Fig. 1 Experimental results on the Iris dataset with four attributes

Fig. 1 showed the experimental results for all four attributes. The best choice of k (number of nearest neighbors) for attribute SW, SL, PW, and PL was respectively 6, 13, 5, and 10. Although the 1-*NN* method was not quite ideal, it still yielded acceptable results.

Table 2 compared the accuracy of the proposed predicting method with that of *multiple imputation* [10] and that of *mean substitution*. 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 (i.e., to create predictions for the *distributions* of each missing value [10]). In this approach, it should be assumed that the data are missing at random. As for mean substitution, the missing attribute value is directly substituted by mean of known values. It is easily seen that our approach leads to superior performance compared to both multiple imputation and mean substitution.

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

Method	Accuracy (RMSE)			
	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

7. Conclusions

In this paper, we propose a grey-based nearest neighbor approach to deal with incomplete-data problems. The nearest neighbors of an instance with missing attribute values are found by using grey relational analysis. Consequently, the valid attribute values derived from these nearest neighbors are used to predict those unknown. Experimental results have shown that our proposed approach yields superior performance.

References

[1] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37-66, 1991.

[2] W. L. Buntine and A. S.Weigend, "Bayesian backpropagation," *Complex Systems*, vol. 5, pp. 603-643, 1991.

[3] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.

[4] 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.

[5] J. Deng, "The theory and method of socioeconomic grey systems," *Social Sciences in China*, vol. 6, pp. 47-60, 1984. (in Chinese)

[6] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 10, pp. 617-621, 1979.

[7] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics* Part 2, vol. 7, pp. 179-188, 1936.

[8] 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.

[9] J. H. Friedman, "A recursive partitioning decision rule for nonparametric classification," *IEEE Transactions on Computers*, pp. 404-408, 1977.

[10] 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.

[11] 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.

[12] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

[13] J. R. Quinlan, "Unknown attribute values in induction," in *Proceedings of the Sixth International Machine Learning Workshop*, San Mateo, CA: Morgan Kaufmann, pp. 164-168, 1989.

[14] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, Wiley, New York, 1987.

[15] C. Stanfill and D. Waltz, "Towards memory-based reasoning," *Communications of the ACM*, vol. 29, no. 12, pp. 1213-1228, 1986.

[16] 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