International Journal of Hybrid Information Technology
Vol.7, No.2 (2014), pp.347-356
http://dx.doi.org/10.14257/ijhit.2014.7.2.30

Missing Data Imputation Based on Grey System Theory

Guoming Sang, Kai Shi, Zhi Liu and Lijun Gao

Dalian Maritime University

Abstract

This paper proposed a new weighted KNN data filling algorithm based on grey correlation analysis (GBWKNN) by researching the nearest neighbor of missing data filling method. It is aimed at that missing data is not sensitive to noise data and combined with grey system theory and the advantage of the K nearest neighbor algorithm. The experimental results on six UCI data sets showed that its filling accuracy is better than the traditional method of K nearest neighbor and filling algorithm presented by Huang and Lee.

Keywords: *missing data; grey correlation analysis; data filling*

1. Introduction

Data is the basis of data mining .The quality of the data set directly influences the effect of data mining. Any efficient mining algorithm has lost its original advantages without perfect data sets. The existing data preprocessing can handle the data including deletion，noise, inconsistent, repetition, redundancy, omission. It makes the data be suited to a variety of data mining methods. But the traditional data cleaning technology preprocesses the original data simply and roughly .The steps are screening, deletion, addition and transformation. Although this approach can get ideal data, the data always cause the inner changes. Especially for the missing data if we delete the data directly it might lose implicit information. However, we don't preprocesses the data it has no suitable algorithm for data mining. Missing data filling is the best way the researchers thought. So how to handle data missing problem and how to choose filling strategy is important for data preprocessing and data mining

K nearest neighbor (KNN) is a preferred data filling approach at present. It is an evidence-based value estimation method that uses complete record to estimate incomplete record, and confirms the closeness of relationship between data records by using Minkowsk distance [2]. However, Minkowsk distance is quite limited in the applicability of calculating the distance between of data records, while it is more effective in equally distributed dataset. In skewed dataset, optimum result cannot be obtained by using Minkowsk due to unequally distributed data. What's more important is that Minkowsk distance formula works well when handling missing value of continuous attribute, while it does not work very well generally when handling the attribute of discrete type or mixed type. In Reference [1], a new standard used to confirm closeness of relationship (also known as similarity) between data records that is to use the knowledge of Grey Relational Analysis theory to confirm the level of similarity between data records. Grey Relational Analysis theory is superior to Minkowsk distance on handling discrete, continuous and mixed attribute. Therefore, the level of similarity between two data records is confirmed by using Grey Relational Analysis theory to replace Minkowsk distance in KNN.

ISSN: 1738-9968 IJHIT

2. Grey Relational Analysis

Grey Relational Analysis (GRA) is a measuring method of confirming the level of similarity between two data records in Grey System Theory [3]. As for two data records, their grey level will be considered to be larger if they have the same future development; otherwise it is smaller. In conducting GRA, the size of GRE is usually used to determine the relationship between a record with missing data and a record without missing data. As the magnitude order of data attribute of data concentration often has big difference, the data of each attribute in data record would be quantized firstly in order to avoid the bigotry caused by it in filling, so as to make them vary in the range of [0,1]. The calculation formula of data conversion is as following:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (2.1)$$

Hereinto, $\min_A$ and $\max_A$ is the maximum and minimum respectively under attribute A for each case. v' is the mapping of vale V on attribute A in each case to $[new\_min_A, new\_max_A]$ $\text{new\_min}_A, \text{new\_max}_A$,and in the chapter is 0 and 1 respectively.

Definition 1: In dataset $D = \{x_0, x_1, \cdots, x_n\}$, $x_i = \{x_i(1), x_i(2), \cdots, x_i(m)\}$, and $i = 0,1,2,\cdots,n$, m is the number of attributes in each case, so that the grey relationship coefficient of the two cases on attribute A is:

$$GRC(x_0(A), x_i(A)) = \frac{min_{\forall j} min_{\forall k} | x_0(k) - x_j(k) | + \alpha max_{\forall j} max_{\forall k} | x_0(k) - x_j(k) |}{| x_0(A) - x_i(A) | + \alpha max_{\forall j} max_{\forall k} | x_0(k) - x_j(k) |} \quad (2.2)$$

Hereinto, $\alpha \in [0,1]$, (generally $\alpha = 0.5,\ i = j = 1,2,\cdots,n, A = k = 1,2,\cdots,m+1$) and $GRC(x_0(A), x_i(A)) \in [0,1]$represent the level of similarity of cases x_0 and x_i on attribute. When $GRC(x_0(A), x_i(A)) = 1$, it shows that x_0 and x_i have the same attribute value on attribute A; on the contrary, when x_0 and x_i have different values on attrubute A, the value of $GRC(x_0(A), x_i(A))$ tends to 0.

Definition2: In dataset $x_i = \{x_i(1), x_i(2), \cdots, x_i(m)\}, i = 0,1,2,\cdots,n$ and m is the number of attributes in each case, so that the calculation formula for grey similarity of the similarity level between cases x_0 and x_i is determined to be:

$$GRG(x_0, x_i) = \frac{1}{m}\sum_{A=1}^{m} GRC(x_0(A), x_i(A)), i = 1,2,\cdots,n \quad (2.3)$$

The larger the grey similarity between two cases determined in Formula 3, the more similar the two cases. If $\text{GRG}(\text{x}_0,\text{x}_1) > \text{GRG}(\text{x}_0,\text{x}_2)$, it shows that the level of similarity between x_0 and x_1 is smaller than that between x_0 and x_2. $\text{GRG}(\text{x}_0,\text{x}_i)=1$ shows that cases x_0 and x_i are totally irrelevant; $\text{GRG}(\text{x}_0,\text{x}_i)=1$ shows that cases x_0 and x_i are the same.

3. GBWKNN based on Grey System Theory

3.1. GBWKNN Algorithm Description

In dataset $D=\{D_1, D_2, \cdots, D_n\}$ and $D_I=\{D_1, D_2, \cdots, D_r\}$ containing n cases, $r \le n$ is the case set with missing data, while $D_C=\{D_{r+1}, D_{r+2}, \cdots, D_n\}$ is the complete case set. For each case D_i, in which $i=1,2,\cdots,n$, $(m+1)$ attributes are included. $V(i, j)$ represents that it does not contain missing value on i line and j column; $MV(i, j)$ represents that it contains missing value. The filling value here is expressed as $MV_K(i, j)$. The pseudo-code of GBWKNN algorithm is described as following:

Table 1. The pseudo-code of GBWKNN algorithm

Algorithm: GBWKNN

1 input：D,D*I , DC , θ*
2 output： D
3 **for** *DI i ∈ DI* **do**
Calculate missing rate of DI i and sorting DI i ascend
4 **for** *MV(i,j) ∈ DI* **do**
switch *type of j* **do**
Case Discrete: M VK (*i, j*)=*Mode{attribute j in D}*
Case Symbol:M VK (*i, j*)=*Mode{attribute j in D}*
Case Continuous： *M VK* (*i, j*)=*Mean{attribute j in D}*
Repeat
5 **for** *DI i ∈ DI* **do**
instances DK (*i*) = *GetK instance*(*DI i , DI ∪ DC*)
DI i = I mputation(*DI i , DK* (*i*))
6 return *D = DI ∪ DC*
Untill predicted value is invariable or the difference of accuracy rate between the first and the second exceeds θ

3.2. Conditions for End of Algorithm

GBWKNN algorithm is a repeated filling algorithm based on *KNN* algorithm. It has higher accuracy in filling comparing with single filling [4]. In the algorithm, the cases with missing data will firstly be sequenced according to miss rate by filling from the case with the minimum miss rate. For the first time of filling, it will apply the average value of the same attribute for continuous attribute. As to discrete or symbolic attribute value, it will apply the maximum of the same attribute to fill. From the second filling, grey similarity is used in the algorithm to calculate *K* cases of nearest neighbor in each case of missing data, after which new estimation method will be sued to replace the original filling value.

Multiple-repeated filling method is used to replace the missing attribute value in GBWKNN algorithm, but the finiteness of algorithm determines that it must have conditions

for end. The conditions for end of GBWKNN algorithm are given as bellows:

1) As to continuous missing attribute value, the algorithm stops filling when there is algorithm convergence or repeated filling value.

2) As to discrete or symbolic missing attribute value, the algorithm stops filling when difference of accuracy between the first and the second filling is larger than given threshold value Θ which is given by the user.

4. Experiment and Result Analysis

To verify the timeliness of the algorithm in the paper, experiment is carried out on UCI dataset for GBWKNN algorithm. In the experiment, each attribute value is mapped onto [0, 1] in the method from Formula 1 used by each case. The value of K, the number of nearest neighbor, is from 1 to 50.

In the experiment, the verifying method of leave-one-out is used to verify the result of experiment. That is, the predicted value of missing value in each case x_i which contains missing attribute is estimated by all other cases of data concentration apart from case x_i itself. Therefore, in the prediction for each missing attribute value, almost all the cases are used for contribution cases to dedicate the information of the case for estimation of missing value.

4.1. Convergence Analysis

To illustrate the features and strengths of filling algorithm of missing data in the paper, contrast experiment is carried out between KNN filling which has good effect in filling missing data at present and the filling algorithm of missing data proposed by Huang and Lee in 2004. In KNN filling algorithm, the maximum and the minimum based on Euclid spatial distance are taken as the standard whether the relationship between two cases are close. The filling algorithm of missing data proposed by Huang and Lee also uses grey similarity as the standard to evalue the closeness between two cases. The algorithm in the paper takes the significance of missing attribute column in grey similarity into account, but Huang and Lee do not. Meanwhile, the algorithm in the paper is different on estimation method of filling value from theirs. In the experiment, classification accuracy is used as the evaluation standard for discrete attribute missing value; predictive accuracy is used as the evaluation standard for continuous attribute missing value. In handling dataset of experiment, to compare the difference between filling value and true value, all the complete datasets (without any missing data) used are preprocessed through data. The setup of missing attribute value uses the principle of missing at random shown in Formula 4.

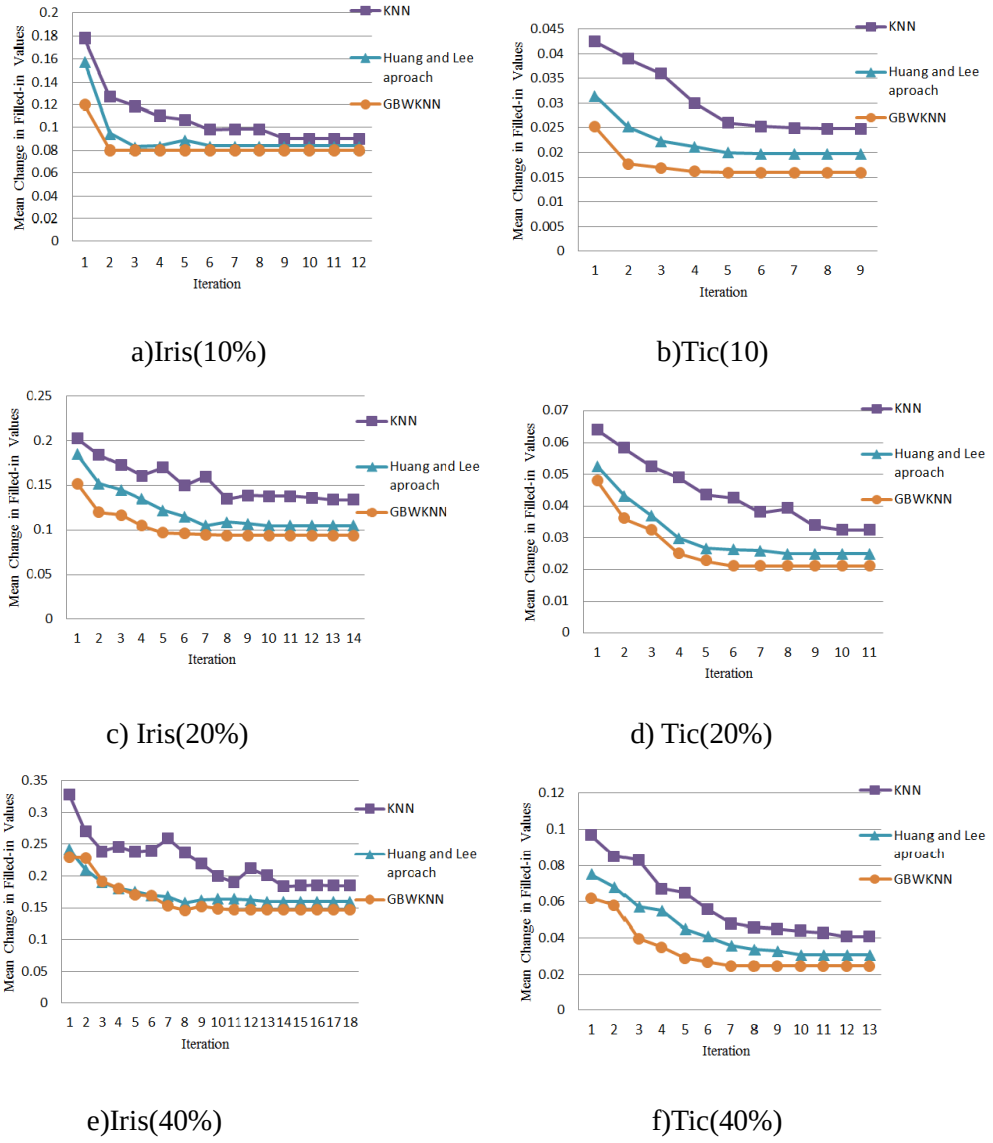
$$p(\varepsilon = 1 | X = x) = \begin{cases} 0.9 - 0.2 | x - 1 | & if \ | x - 1 | \le 4.5 \\ 0.1 \end{cases} \quad (4.1)$$

Hereinto, $\varepsilon = 1$, which indicates that x is a missing value.

The GBWKNN algorithm proposed in the paper is an iterative filling algorithm of EM type [5]. The best predicted value is filled through iterative filling for many times. As iteration means repeat, the precondition for the establishment of algorithm is that it must be finite, that is, the procedure would stop after iteration fore several times. When the mean difference of filling data tends to be zero, the filling algorithm reaches convergence state. The mean

difference of filling data refers to the difference between the average of all filling values in the last iterative filling and the average of all filling values in current iteration. In nonparametric filling model (KNN, nuclear model), the value would generally tend to zero, but not equal to zero, and have the tendency of circulating. It is generally considered that the more the difference between the averages of two iterations tends to zero, the better the convergence effect of the filling is.

In Figure 1, it shows the convergence of the three algorithms under the state of miss rate (10%, 20%, 30%) in different condition attributes (continuous attribute) on the two data (Iris and Tic). The sub-figures a), c) and e) show the changes of mean difference on dataset Iris when the miss rate is 10%, 20% and 40% for the three algorithms; the sub-figures b), d) and f) show that the changes of mean value on dataset Iris when the miss rate is 10%, 20% and 40% for the three algorithms.



a)Iris(10%) b)Tic(10)

c) Iris(20%) d) Tic(20%)

e)Iris(40%) f)Tic(40%)

Figure 1. Experimental results on Iris dataset (left) and Tic dataset (right) for three algorithms

It is clear that GBWKNN algorithm in the paper and the other two comparing algorithms (KNN, Huang and Lee approach) are convergent under different missing attribute types (numeric type) on two different datasets, because the mean difference of filling data tends to be stable after iteration for several times. It shows that it is quite reasonable to fill the estimated value for missing attribute value by using iteration method. It can be seen from the change rule of sub-figures a), c), e) and b), d), f) in Figure 1, the higher the miss rate of a dataset is, the more times of iteration process take place in order to reach convergent state for the three algorithms. Taking dataset Iris for example, when the miss rate is 10%, 20%, 40%, the convergent iterative times of GBWKNN algorithm is 3, 6, and 10 respectively. It is easy to conclude that the more the missing values in a dataset are, the less the useful information can be used in the algorithm of filling missing data. To obtain the optimal filling result, the times of iteration need to be increased correspondingly. In order to verify that the convergent effect of the algorithm proposed in the paper is better, the three algorithms are applied on six UCI datasets, Iris, Tic, Hepatitis, Echocardiogram, Soybean, and Water-treat. Table 2 shows the iterative times of convergence for the three algorithms on the six datasets, and the selection of the best K value. For instance, if the miss rate on dataset Soybean is 6.63%, the iterative times of convergent GBWKNN algorithm are 5 when K value is 3. It is easy to conclude through the table that the iterative times of convergent GBWKNN algorithm are less than that of KNN, Huang and Lee approach algorithms, *i.e.*, its speed of convergence is a little faster than that of the other two algorithms.

Table 2. Iterative times for six datasets after the three algorithms converge

	MR(KNN)	KNN	Huang and Lee approach	GBWKNN
Iris	10% (k = 2)	11	7	6
	20% (k = 2)	13	10	9
	40% (k = 1)	17	13	11
Tic	10% (k = 8)	8	6	6
	20% (k = 5)	10	8	7
	40% (k = 3)	12	10	8
Hepatitis	5.67% (k = 2)	6	5	5
Echocardiogram	7.69% (k = 1)	5	4	3
Soybean	6.63% (k = 3)	8	6	5
Water-treat	2.95% (k = 5)	8	7	6

4.2. Experimental Evaluation Standard of Prediction Accuracy

When the filling algorithm reaches convergent state, whether the missing filling value is closer to true value or it is only the convergence of algorithm, the difference between filling value and true value does not change. As for numeric missing value, Root Mean Square Error (RMSE) is taken as the evaluation standard of prediction accuracy in filling.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_i - e_i')^2} \qquad (4.2)$$

Hereinto, e_i is the initial attribute value, e_i' is the estimated attribute value, and m is the total quantity of missing attribute values. It is clear that the smaller the value of $RMSE$ is, the closer the estimated value to the true value is. On the contrary, the estimated result will be more deflected from the original true value.

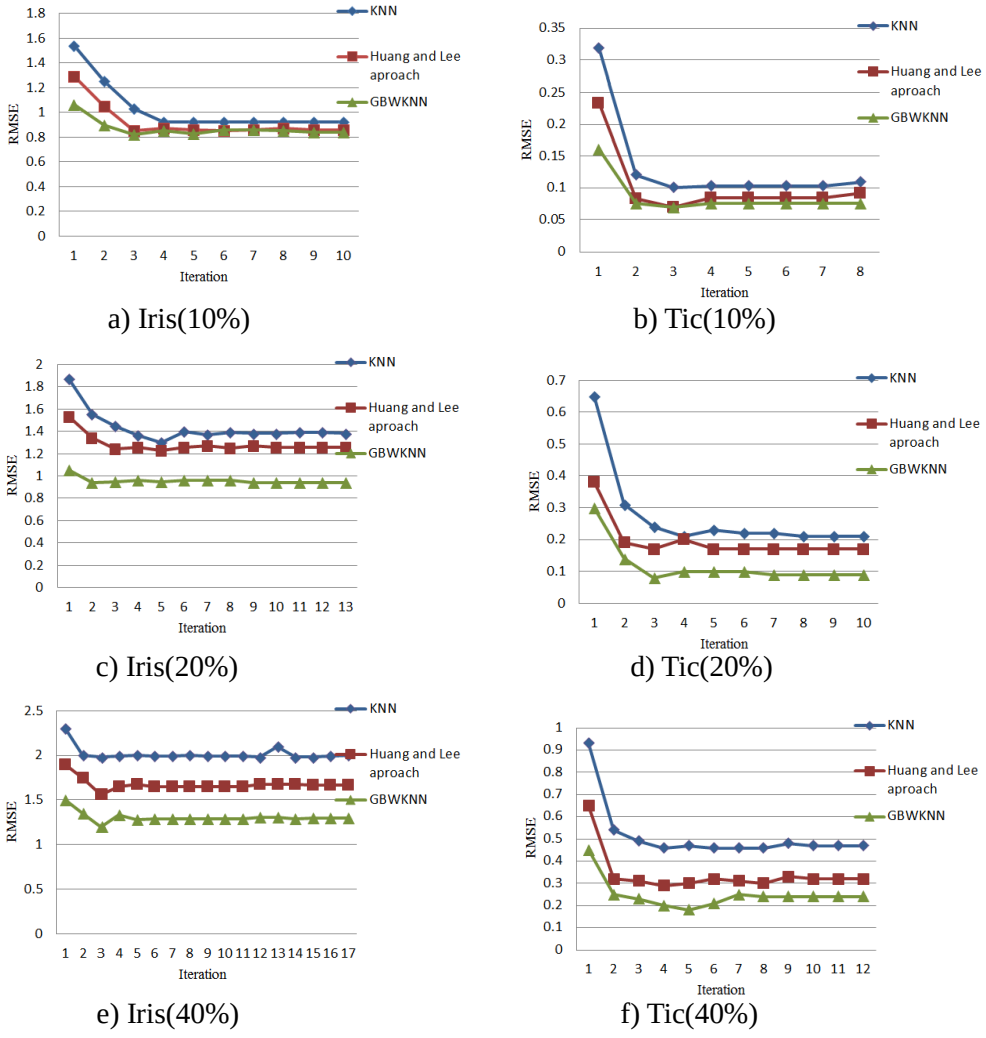


Figure 2. Experimental results on Iris dataset (left) and Tic-tac-toe dataset (right) for three algorithms

Figure 2 shows the changes of RMSE under different miss rate on datasets Iris and Tic for KNN, GBWKNN, Huang and Lee approach. In Figure 2, when the filling algorithm implements convergence, the change rule of RMSE on datasets of different miss rate for the three algorithms is that obvious downtrend exists comparing the RMSE value for the second iteration and the RMSE value for the first iteration, and that the RMSE value for the third

iteration is obviously smaller that the RMSE value for the second iteration. After iteration for several times, the change of RMSE values under all algorithms tends to stable, and even it will not change obviously anymore. However, the GBWKNN algorithm in the paper works better than the other two algorithms. As for the attribute missing value of discrete type, classification accuracy is taken as the standard to evaluate the strength and weakness of algorithm. The higher the classification accuracy is, the better the filling effect is. In Figure 3, it shows the changes of classification accuracy on four missing datasets of discrete type, Hepatitis, Echocardiogram, Soybean, and Water-treat, for the filling algorithm in the paper and other two comparing algorithm. The experimental result indicates that 1) the filling effect of GBWKNN algorithm in the paper is better than the other two algorithms; 2) on different datasets, the classification accuracy after the second filling is usually higher than that after the first filling. Such an experimental result shows that the repetitive filling is more effective than single filling. When the filling algorithm converges, the classification accuracy of filling will not change. In order to keep the accuracy of filling, the repetitive filling would be stopped when the algorithm implements convergence, because it often causes filling cycle to continue filling, and reduces the original accuracy.

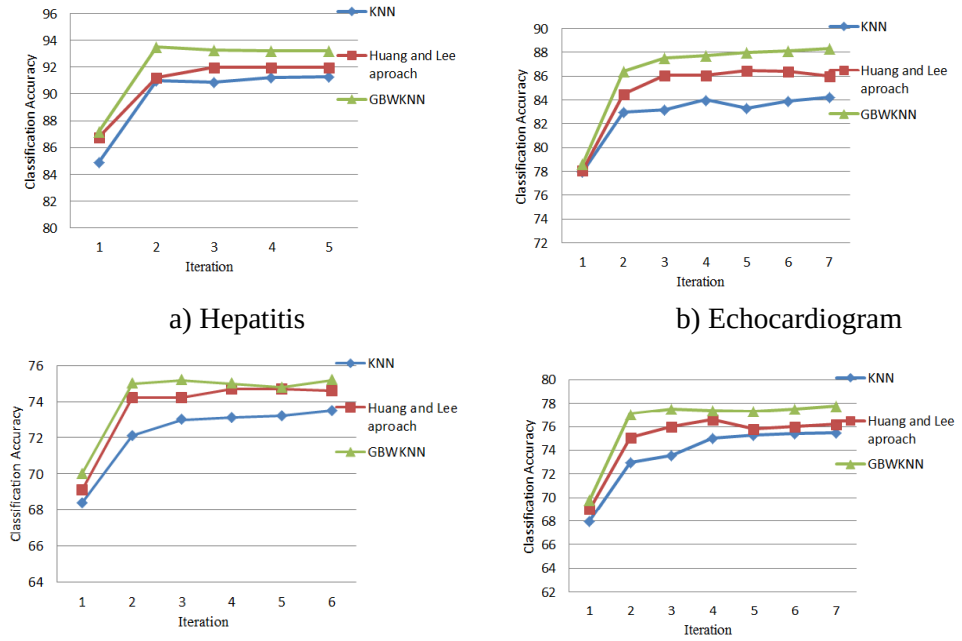


Figure 3. Experimental results on hepatitis, echocardiogram, soybean and water-treat

5. Conclusion

In the paper, a new repetitive missing data filling method of nearest neighbor, GBWKNN, based on grey system is proposed. Comparing to traditional filling method of nearest neighbor, evaluating the similarity level of two cases by using Grey Relational Analysis theory speeds up the convergence of repetitive filling; on the estimation method of missing value, weighted method is used for numeric missing attribute value. The weight of each nearest neighbor case

is calculated firstly, and then the missing attribute value is estimated in the method of weighted average; the maximum class method should be applied to estimate the missing attribute value for missing attribute value of discrete type; in order to make the estimated value closer to the true value, repetitive filling should be conducted to the missing attribute value for several times; the experimental result on the six UCI datasets shows that the filling algorithm in the paper is superior to KNN algorithm and the filling algorithm proposed by Huang and Lee in 2004. On estimation method, it uses RMSE standard to measure the predicted accuracy and the error rate of classification. In that the algorithm is convergent after several steps, and that all the evaluation standards tend to stable after the convergence of algorithm, it is the focus of our future research when the filling can stop so as to make the estimated filling value closer to the true missing value.

Acknowledgements

This work was supported by Science and Technology Planning Project of Dalian City, China (No. 2011E15SF100,2011A17GX073) and the Fundamental Research Funds for the Central Universities (No. 3132013337).

References

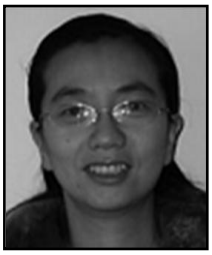
[1] C. C. Huang and H. M. Lee, "A grey-based nearest neighbor approach for missing attribute value prediction", Applied Intelligence, vol. 20, no. 3, **(2004)**, pp. 239-252.
[2] W. Yu, "Study on Text Categorization Based on Decision Tree and K Nearest Neighbors TianJin", Tianjin University.
[3] Z. Manlong, "New Technologies for Imputation and Classification Based on NN Approach", GuangXi Normal University, (**2010**).
[4] L. X. Yi, "Based on Grey- based and KNN algorithm for imputing missing attribute values", Microcomputer Information, vol. 24, no. 5-3, **(2007)**, pp. 246-248.
[5] H. Li, A. Emmanuel, P. Li and M. Wu, "Imputation algorithm of missing values based on EM and Bayesian network", Computer Engineering and Applications, vol. 46, no. 5, **(2010)**, pp. 123-125.

Authors

Guoming Sang

He received the M.Sc. degree in **the** major of Computer Application from Dalian University of Technology, PRC in 1999. He is now an associate professor at the school of Information Science and Technology of Dalian Maritime University. His research interests include wireless sensor networks, artificial intelligent and data mining.

E-mail: sangguoming@dlmu.edu.cn



Zhi Liu

She received the M. Sc. degree from Dalian University of Technology, PRC in 1999. She received the Ph.D. from Dalian Maritime University in 2006. She is now an associate professor of Dalian Maritime University, PRC.

Her research interests include data mining and artificial intelligence.

E-mail: lzsgmsc@126.com

