The nearest neighbor algorithm of filling missing data based on cluster analysis

Zhang Chi
Department of mathematics and Computer ,
Wuhan Polytechnic University
Electronic banking department,department of hubei branch,
Agricultural Bank of China
Wuhan,China
e-mail: xtfo119@163.com

Fong Hong-cai
Department of mathematics and Computer ,
Wuhan Polytechnic University
Wuhan,China
e-mail: fenghc@whpu.edu.cn

Jin Kai
Department of mathematics and Computer ,
Wuhan Polytechnic University
Wuhan,China
e-mail: sir.jinkai@gmail.com

Yang Ting
Department of mathematics and Computer ,
Wuhan Polytechnic University
Wuhan,China
e-mail: 731785434@qq.com

***Abstract*—Missing data universally exists in various research fields and it results in bad computational performance and effcet. In order to improve the accuracy of filling in the missing data, a filling missing data algorithm of the nearest neighbor based on the cluster analysis is proposed by this paper. After clustering data analysis,the algorithm assigns weights according to the categories and improves calculation formula and filling value calculation based on the MGNN (Mahalanobis-Gray and Nearest Neighbor algorithm) algorithm.The experimental results show that the filling accuracy of the method is higher than traditional KNN algorithm and MGNN algorithm.**

Keywords-Grey mcorrelation; Mahalanobis distance; Cluster analysis; Nearest neighbor alorithm; Maximum

Ⅰ. INTRODUCTION

In data statistical analysis, statistical decision statement is one of the important reference basis for making decision.However, due to the differences of obtaining data and understanding the data structure and other factors, the phenomenon of the missing data appears in report often happens in practice.This kind of phenomenon is known as data loss, it gives decision some influence. To solve the problem, researchers at home and abroad put forward a lot of filling in missing data methods, these methods can be roughly divided into two kinds, one is the statistics method,and the other is the data mining method..The methods used with parameter are EM method, linear regression method, multiple filling method and so on. These methods provide a good filling effect, but they will produce a lot of data deviation and they affect the quality of filling in missing data if the structure of data was understand not fully or people use a wrong data model[1]. Data mining methods includes rough set method, neural network, decision tree, bayesian network,nearest neighbor method and so on, and the nearest neighbor method is the most widely used as filling effect is best [2].The principle of the nearest neighbor algorithm is the closer the distance,the closer the relationship between examples.And the core of the nearest neighbor algorithm is calculating the distance between the two examples[5]. The traditional nearest neighbor algorithm is based on calculating the correlation of examples by Euclidean distance which applicability is limited [9] - [11],but a method of using markov distance instead of Euclidean distance makes the application range extended proposed by paper [4]. If using markov distance as distance judgment standard only, unknown information will be treated as known information, so paper [5] presentes according to markov distance and grey relational analysis to calculate the distance of two examples. However,the formula to calculate the distance and filling the missing value by mean or median in paper [5] can make the calculation accuracy negative effects.Therefore this article improves the formula to calculate the distance on the base of paper [5], and fills in the missing value based on distributing weighting through the clustering analysis.It improves the accuracy and enhances the filling effect.

Ⅱ.THE BASIC IDEA OF THE ALGORITHM

A.Filling the missing data based on K neighbor algorithm

K neighbor method is the expansion of the nearest neighbor method, the basic rule is finding out the K nearest neighbors from the testing sample in all N samples. We use Euclidean distance to calculate the distance between the samples in practice commonly, such as the distance between the two cases x_0 and x_i in the p dimensional space can be expressed as :

$$Dist(x_{0,}x_i) = (|x_{01} - x_{i1}|^2 + |x_{02} - x_{i2}|^2 + \cdots |x_{0p} - x_{ip}|^2)^{\frac{1}{2}} \qquad (1)$$

The algorithm of filling the missing data Based on the Euclidean distance can be described as:

(1)For each example that contains missing data ,calculate it with all other the distance between the case through type (1);

(2)Sort the distance in sequence and select K minimum distances.

(3)If the datas to the case are discrete, then take the value in which the most kinds of the K distances as filling value; If the datas to the case are continuity, then take the median or average of the value in which the K distances as filling value.

B.The distance to MGNN algorithm

MGNN (Mahalanobis - Gray and Nearest Neighbor) algorithm is the Nearest Neighbor algorithm through calculating the marsh distance and grey relational analysis. Paper[5-7] prove that grey relational analysis and markov distance both can be used to calculate the similarity between the two attributes, but used in different areas. The effect using Grey correlation analysis to calculate the two examples of the similarity in the case of the density related not obviousis is very good.On the other side, if the density of the cases related obviousis, we often use Euclidean distance or markov distance to determine the similarity among the cases. For Euclidean distance is the special case of markov distance and markov distance is in more extensive use, choosing markov distance as the supplement of grey relational analysis is more effective.And the distance of the two examples between x_0 and x_i can be calculated by type (2)

$$Dist(x_0,x_i)=\lambda GRG(x_0,x_i)+(1-\lambda)Mahal(x_0,x_i);$$
$$i=1,2,\cdots,n \qquad (2)$$

In the above formula (2), n is positive integer,and $\lambda\in[0,1]$ is parameters to adjust, which can be set a certain value by the awareness degree according to the correlation of the data. If the correlation of the data is known completely,then $\lambda=0$ .Otherwise $\lambda=1$.

In the process of the Grey Relational analysis,we usually use GRG(Grey Relational Grade) to describe the degree of the relationship between two cases .The GRG of x_0 and x_i is defined as:

$$GRG(x_0,x_i)=\frac{1}{n}\sum GRC(x_0,x_i), i=1,2,3,\cdots n \qquad (3)$$

The GRC (Grey Relational Coefficient, GRC) called the relationship Coefficient, defined as follows:

$$GRC(x_0(p),x_j(p))=\frac{\min_{\forall j}\min_{\forall k}|x_0(k)-x_j(k)|+\rho\max_{\forall j}\max_{\forall k}|x_0(k)-x_j(k)|}{|x_0(p)-x_j(p)|+\rho\max_{\forall j}\max_{\forall k}|x_0(k)-x_j(k)|} \qquad (4)$$

$\rho\in[0,0.5]$ is distinguishing coefficient in Formula (4), and the role is to eliminate the influence of GRC distortion due to value **Δmax** isexcessive[4]; $j=1,2,\cdots,n$, $k=p=1,2,3,\cdots,m$. n is the number of all sequences to be compared,and m is the number of elements of the n sequences. $|x_0(p)-x_j(p)|$ is the absolute of the difference of the p_th elements in two compare sequences.The larger relation coefficient , the more close relationship between the two events,Otherwise,the relationship between the two events is not close.When $|x_0(p)-x_j(p)|$ is taken minimum value, $GRC(x_0(p),x_j(p)=1$; when $|x_0(p)-x_j(p)|$ is taken the maximum, $GRC(x_0(p),x_j(p)$ is minimum value.So $GRC(x_0(p),x_j(p)\in(0,1]$.

Grey Relational Grade of x_0 and x_i in Formula (3) is the arithmetic mean value of Relational Coefficient. $GRG(x_0(p),x_i(p))\in(0,1)$, if $GRG(x_0(p),x_1(p))>GRG(x_0(p),x_2(p))$, the similarity of $x_0(p)$ and $x_1(p)$ is greater than the similarity of $x_0(p)$ with $x_2(p)$ [6].

$Mahal(x_0,x_i)$ is the markov distance of x_0 and x_i . It is covariance distance,and it is also a kind of effective method to calculate the similarity of two unknown samples. The markov distance of the sample i with the sample j can be expressed as follow:

$$d_{ij}=\sqrt{(x_i-x_j)^T\sum{}^{-1}(x_i-x_j)} \qquad (5)$$

In the above formula (5), T is transpose, and $\sum$ is covariance matrix of sample. We can get markov distance when the inverse matrix of $\sum$ exists.If the inverse matrix of $\sum$ does not exist, use Euclidean distance to describe two samples of similarity.

C.Clustering analysis

Clustering analysis is a kind of exploratory analysis.It automatically classify according to the characteristics of data itself, and it doesn't need a classification standard before the process of classification. The result is a sense tend to resemble each other in the same class examples,the sense not in the similar cases tend to not similar.Clustering analysis is necessary in the huge data. System clustering method, K-means clustering method and fuzzy C-means clustering method are in common use,and these methods all have very good clustering effect. Among them, the system clustering method is the most widely used currently.

The basic idea of system clustering method is that the two closest data are merged into a class according to a certain distance criterion between all the data, and then calculate the distance between class and the other data, ane the nearest data would be merged into the class. The Distance between class and other data is calculated iteratively and the nearest one is merged into the class until all of the data are merged in a class. Finally determine the results of classification according to the clustering tree's contents and some relevant sample experience .

For the convenience of the calculation,we use mathematical software -- MATLAB.And we can use linkage function to select different system clustering method by changing the parameters of "method". In most cases, the class average method is a relatively stable method in several system clustering method.

D.Analysis and improvement to MGNN algorithm

Paper [12] improves that instances of the data can not be too many,there may be dimensional disaster after using the nearest neighbor algorithm if the number of instances is over 50.Take cluster analysis as the pretreatment of MGNN algorithm, and classify similar data in accordance with the characteristics of the data itself, only fill the data which containing the missing data in the class. This mode improves the accuracy as well as reduces dimension.Give attributes contain missing data assigned weights $\partial 1$ and the attribute does not contain missing data $\partial 2$ after cluster analysis on the case attributes. The calculation method of filling in the value such as type (6) shows:

$$X(filling)=\frac{1}{n}\partial 1(a_1+a_2+\cdots+a_n)+\frac{1}{m}\partial 2(b_1+b_2+\cdots+b_m) \qquad (6)$$

As regards the missing values to fill , the relationship of properties belong to the class I is closer than the other properties belong to the class II , so $\partial_1 > \partial_2$. $a_1, a_2, \cdots a_n$ and $b_1, b_2, \cdots b_m$ are the elements of properties from the class I and class II each other.

In gray relational analysis theory, if the similarity of $x_0(p)$ with $x_1(p)$ is greater than $x_0(p)$ with $x_2(p)$, then $GRG(x_0(p), x_1(p)) > GRG(x_0(p), x_2(p))$. $GRG(x_0, x_1) = 1$ saies the two cases are the same and the distance is shortest between them.Conversely, $GRG(x_0, x_1) = 0$ saies two cases without any links and the distance is longest[6].Take following formula (7) instead of formula (5) to improve the algorithm.New algorithm Abbreviates is abbreviated as GMKNN (Gray-Mahalanobis' and the k-there Nearest neighbor algorithm).

$$Dist(x_0(p),x_i(p))=\lambda(1-GRG(x_0(p),x_i(p)))+(1-\lambda)Mahd(x_0(p),x_i(p)) \qquad (7)$$

in which $\lambda \in [0,1]$, $i = 1, 2, 3, \cdots, n$

So the process by using GMKNN algorithms to fill the missing data can be described as follows:

①Pre-fill the missing data.If the data is continuity or discreteness is not obvious, then take the median or the average of the other elements of the event for filling the value first time ; if the data is discrete, use the largest class method to do firstfill, that is the value which appears most as the fill value

②Eliminate the impact of data dimensionless.This paper chooses the initial value of the transformation method, i.e. each value of a set of sequences Divide of the the first number of the corresponding sequence respectively to obtain a new set of sequences which has not dimension.

③Delete the examples which has little relationship with the missing data instances to lower dimension.

④Calculate the distance of the examples need to be filled and the other examples According to the formula (7).

⑤ Cluster the properties of the sample. Give weight in accordance with the Table 1.

⑥Sort the distance get from step ④ in ascending order and select K minimum distances.With , find the value of the original elements corresponding to the the size of the missing data distance. $a_1, a_2, \cdots, a_n, b_1, b_2, \cdots b_m$, where n is the number of attributes of the class I, m is the number of attributes of the class II.

⑦ If the data is the continuity or discrete nature is not obvious, fill the missing data in accordance with the equation (6);if the data is discrete, use the largest class method to fill missing value by these K minimum distance.

III. The experimental results and analysis

To test the performance of GMKNN algorithm to fill the missing data , take data from a bank last two years every month as example.The data is the account number of personal e-banking, corporate e-banking, telephone wallet, message services, mobile banking on 24 months .Measure variables are the the account number of electronic banking products from eighteen areas every month.Remove a certain proportion as test data in these data randomly, and compare the estimated value with the true value which is deleted.

Experiment is through by Matlab simulation. Divide data into five parts by the five banking products and select one part of the data to input 18×24 matrix.Each row of the matrix corresponds to an observation event and each column corresponds to a variable property.In order to facilitate fast clustering,this article uses a clusterdata function by Matlab and orders $\partial 1 = 0.6$, $\partial 2 = 0.4$ .In order to verify the performance of the algorithm, choose the traditional KNN algorithm, MGNN algorithm and GMKNN algorithms to fill the missing data and make the average error rate as the error criterion.The average error rate is abbreviated as rate1 and defined as follow:

$$\text{rate1} = \frac{\frac{1}{n}\sum|truth - \text{estimate}|}{truth} \times 100\%$$

,n is the number of filling values (8)

The experimental results and error as shown in Figure 1 below:

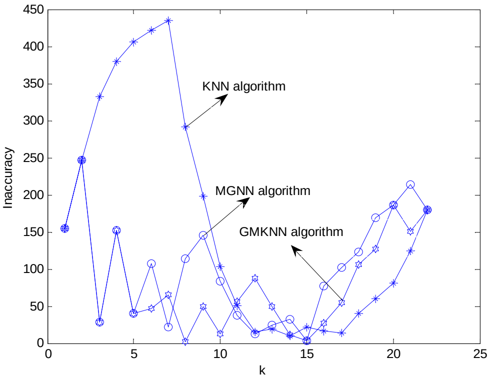

Figure 1 Error plots of the three algorithms for one missing data

Calculate the average error rate of the three algorithms by using formula (8) and the error is shown in Fig1. The average error rate of traditional KNN algorithm is 18.75%, the average error rate of MGNN algorithm is 11.74%, and the average error rate of GMKNN algorithm is 9.5%.

The comparison can be seen by the average error rate,when the amount of missing data is small, GMKNN algorithm is better than MGNN algorithm.Even the error is less than 2 when K = 8 by using GMKNN algorithm.Actually missing a small part of the data often appears in most cases , such as the data of 5% is lost.In order to test the advanced nature of the algorithm, collect 20 datas in the raw data and then restore the selected 20 datas with three algorithms.Finally,take error square root (RMSE) as a standard of filling accuracy, and the smaller RMSE is, the more similar filling data and real data are.The experimental results are shown in figure II.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e(i)-true(i))^2}$$

As can be seen from Figure 2, in the different values of K, the accuracy of GMKNN algorithm is better than MGNN algorithm and traditional KNN algorithm.

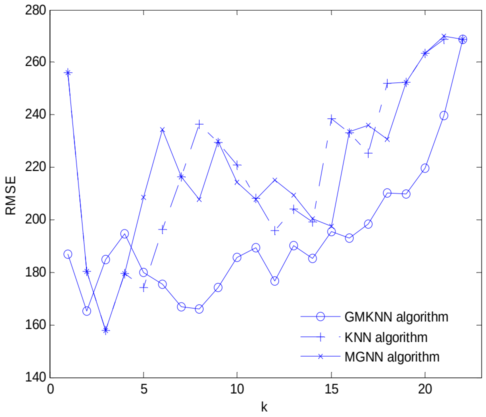

Figure II Error plots of the three algorithms for a partof missing data

Ⅳ.Conclusion

This paper presents a nearest neighbor algorithm based on cluster analysis to fill missing data,and the algorithm has the ability to handle all types of data proved by experiments.First, the algorithm is looking for a nearest neighbor based on the same class data, not only improve the accuracy but also greatly reduces the number of dimensions; Secondly, the accuracy of the improved KNN algorithm based on the Mahalanobis distance and gray analysis is improved; Finally,the accuracy is further improved by using cluster analysis assigned the weight instead of the mean or median in the calculation process of filling value for continuous or discrete not obvious data.

Ⅴ.References

[1]Liu Xingyi Zeng Chunhua. The Handling and challenges of missing data[J].Qinzhou University, 2008, (6) :25-29.

[2]Liu Xingyi, Nong Guocai. The comparison of several different filling missing values method[J]Nanning Teachers College, 2007, (3): 148-150.

[3]Yu Yuemeng, Huang Xiaobin.A algorithm based on KNN text classification[J].Computer Knowledge and Technology, 2012, (3) :1564-1566.

[4]Liu Xingyi, Tan Yao, Zeng Chunhua.Filling missing data method based on the Mahalanobis distance [J].Microcomputer Information, 2010, (9) :225-226.

[5]Liu Xingyi.Filling missing value algorithm based on Mahalanobis distance and gray analysis[J].Journal of Computer Applications, 2009, (9): 2502-2506.

[6]SunWensheng.Economic forecasting methods[M]Beijing: China Agricultural University Press, 2005

[7]Su Yijuan.Multiple imputation method for missing values by gray relation analysis[J].Computer Engineering and Applications, 2009, (15) :169-172.

[8]Yao Guangqun, Wang Yongsheng. Intrusion detection algorithm based on fuzzy evaluation and clustering analysis[J].Computer Engineering and Applications,2012, (21) :169-172 .99-103.

[9]LITTLE R, RUBIN D. Statistical analysis with missing data [M] .2 nd ed. New York: John Wiley and Sons, 2002.

[10]HUANG CC, LEE H M. A grey-based nearest neighbor approach for missing attribute value prediction [J]. Applied Intelligence 2004, 20 (3): 239-252.