International Journal of Hybrid Information Technology
Vol.9, No.4 (2016), pp. 367-374
http://dx.doi.org/10.14257/ijhit.2016.9.4.31

Improving Classification Accuracy Using Missing Data Filling Algorithms for the Criminal Dataset

Cuicui Sun, Chunlong Yao, Lan Shen and Xiaoqiang Yu

School of Information Science and Engineering
Dalian Polytechnic University
No.1, Qinggongyuan, Gan Jing Zi District, Dalian, 116034, P.R.China
yaocl@dlpu.edu.cn

Abstract

Predicting crime types by using classification algorithms can help to find factors affecting crimes and prevent crimes. Due to various reasons in the process of data collection, there are often a large number of missing values in actual criminal dataset, which seriously affects the classification accuracy. Therefore, based on mutual KNNI (K nearest neighbor imputation) algorithm and combined with GRA (Grey Relational Analysis) theory, a novel data filling algorithm called GMKNN is proposed in order to improve the classification accuracy. The algorithm replaces the Euclidean distance formula used in mutual KNNI algorithm with the Grey relational grade formula to eliminate the effect of noise from the nearest neighbors and effectively deal with the discrete attributes. By comparing with several popular data filling algorithms based on a real criminal dataset with lots of missing values, higher classification accuracy can be obtained by using GMKNN algorithm, which is up to 77.837%.

Keywords*: Crime types, Classification algorithms, Classification accuracy, Data filling algorithms.*

1. Introduction

A vast amount of criminal behavior has seriously affected people's normal lives. Classification methods of data mining play decisive role on criminal analysis. By using classification algorithms to predict crime types, the relationship among crime characteristics can be mined to guide the police in tracing the source of crime, which can help to infer the crime trends from different criminal characteristics and fight against criminal activities in time. Due to various reasons in the process of data collection, there are often a large number of missing values in actual criminal dataset, which seriously affects the classification accuracy. Prior to building the classification model, an effective data filling algorithm is necessary to fill the missing values of the initial dataset in the data preprocessing phase.

Missing data is pervasive and even inevitable in most fields. There are many factors causing missing values such as human negligence, faulty equipment, the information omitting, the higher cost, etc. How to effectively deal with missing values affects greatly the quality of data mining results. Thus methods of treat missing data received more attention in recent decades. At present, a number of missing data filling algorithms have emerged. Case deletion method [1] deletes simply those rows with missing values. The method is easy to implement, but in the case that values of the attributes have a higher missing rate, a lot of waste of resources can be caused and an amount of useful information hidden in these objects can be easily discarded. Special values filling method [2] create a new data value (such as ' missing value') to represent missing values. Though

ISSN: 1738-9968 IJHIT

the method is simple for dealing with missing values, this unreliable operation may lead to serious data deviation, which is likely to be inappropriate. Artificial filling method [3] is also a simple way to fill the missing values, but it is very time-consuming, and only suitable for the datasets containing fewer amount of data. For the method, the presence of a large number of missing values will lead to impractical outputs. Mean value substitution method [4] replaces the missing value with the average value for each attribute. This simple method is not applicable to discrete attributes, and may falsely increase precision of the estimates, and cause the uncertainty and biased results. Maximum class method [5] uses attribute value whose frequency of occurrence is the highest to replace respectively the missing values for each attribute. The method tends to deal with the discrete types. *k* nearest neighbor (KNN) imputation method [6] imputes the missing values of an instance considering a given number of instances that are most similar to the instance of interest. KNN imputation method is simple, fast and efficient [7] for dealing with missing data. Qin [8] proposes an extended information gain (IG) based algorithm by utilizing fully the underlying relationship between attributes of the dataset. It is difficult for the algorithm to obtain high accuracy when the dataset has a large amount of missing values. Chen et al [9] proposes a new data filling algorithm based on distinguishing the importance of attributes. In order to ensure the accuracy of data, instantaneity and practicality, the algorithm distinguishes important attributes and unimportant attributes by attribute reduction, and then applies the improved Mahalanobis distance algorithm to fill the missing values of the important attributes, and uses the similarity probability method to fill missing values of the unimportant attributes. One disadvantage of this algorithm is that it is only appropriate for small datasets. Xi [10] took advantage of rough set theory to present a filling algorithm based on importance of kernel values. The algorithm uses kernel values to construct the discernibility matrix, so that the data filled can better adhere to the decision rules and eliminate the noise data. But the algorithm has high computational complexity, and is only applicable to treat datasets with a small amount of missing values.

Due to simple operation and high filling accuracy, KNN imputation (KNNI) method has received wide attention for filling missing data. At present, there have been a number of KNNI method based data filling algorithms. In order to effectively dealing with discrete attributes and the impact of noise, a weighted KNN (GBWKNN) data filling algorithm [11] is proposed by replacing two distance formulas of KNNI algorithm with Gray Relational Grade. As there may be noise in choosing the K nearest neighbors of the missing data using the KNNI algorithm, a new imputation algorithm—Mutual *k* Nearest Neighbor Imputation (MKNNI) algorithm [12] is presented to effectively prevent noise. A new APT-KNN algorithm [13] uses the relationship among the attributes to estimate the missing values according to a couple of attribute values which are most similar to the object, so as to guarantee higher accuracy of the imputed results. For extending the KNNI method with mixed types of variables, statistical correlation measures between different data types are established to measure the distance among different types of variables [14]. An improved KNN algorithm [15] is proposed for filling missing data by replacing the existing Euclidean distance with Mahalanobis distance and Gray analysis. The algorithm is inapplicable to large number of missing values, and the more missing values are, the lower filling accuracy will be.

The actual collected criminal datasets used in predicting crime types usually have lots of missing values and involve many discrete attributes. For this kind of datasets, some of the above algorithms cannot work well to improve classification accuracy. Therefore, in order to improve classification accuracy for predicting crime types, a new data filling algorithm--GMKNN algorithm is proposed by combining MKNNI method with Grey System Theory to prevent noise in choosing *k* nearest neighbors and effectively deal with the discrete attributes.

The rest of paper is organized as follows: Section 2 describes the GMKNN algorithm. Section 3 illustrates the experimental results of comparison with several popular algorithms. Section 4 presents conclusions.

2. GMKNN Algorithm

GMKNN algorithm is the extension of KNNI algorithm, which combines MKNNI [12] with GBWKNN [11] algorithms. KNNI algorithm finds the nearest *k* cases of a incomplete case based on this consideration that similar cases have similar attribute values. Finally, according to maximum class principle, the discrete missing values are replaced by the values which occur most frequently in each column of *K* cases. The distances between the cases need to be calculated using distance formulas, such as Euclidean distance formula.

2.1. MKNNI Algorithm

MKNNI algorithm is used to estimate some nearest neighbors of missing instances. At the same time, their nearest neighbors include the corresponding missing instances. Generally, the range of the final nearest neighbors will decrease. Let dataset $D=\{ x_0, x_1,\cdots, x_n\}$, n is the number of cases, $x_i=\{ X_{i1}, X_{i2},\cdots, X_{in}\}$, (m is the number of the types of condition attributes in each case and $i=0, 1, 2,\cdots, n$). For each case x_0 with missing values and x_i with complete case, the k smallest values can be easily calculated using Euclidean distance, and $Dist(x_0, x_1)$ is defined to be:

$$Dist(x_0, x_1) = \sqrt{\sum_{k=1}^{m} (X_{0k} - X_{ik})^2} , \quad i=1, 2, 3... n \tag{1}$$

Hereinto, $i=0, 1, 2\ldots n$, $k=1, 2\ldots m$. Then the most similar K cases set will be identified, namely $N_k(x)=\{ x_1,x_2,x_3,\ldots, x_k\}$, ($x_1,x_2,x_3,\ldots, x_k$ are all complete cases). For each complete case $x_k \in N_k(x)$, K nearest neighbors of each x_k can be calculated respectively by using Euclidean distance formula, namely $N_k(x_k)$. And the Mutual k Nearest Neighbors set is determined according to formula (2). Finally, the discrete missing values will be replaced by the values which occur most frequently in each attribute column in $M_k(x)$.

$$Mk(x)=\{x_k \in D|x_k \in N_k(x)\cap x_0 \in N_k(x_k)\} \tag{2}$$

2.2. GBWKNN Algorithm

GBWKNN algorithm combines the idea of Grey System Theory with k nearest neighbor method. It is a measuring method of confirming the similarity between two data records using Grey System Theory. Let dataset $D=\{ x_0, x_1,\cdots, x_n\}$, n is the number of cases, $x_i=\{ X_{i1}, X_{i2},\cdots, X_{in}\}$, (m is the number of the types of condition attributes in each case and $i=0, 1, 2,\cdots, n$). For each case x_0 with missing values and x_i is compelete case, the values are computed on the gray relativity between this case x_0 and each case x_i with no missing values , then the Grey relationship coefficient formula of the two cases on attribute A is defined to be:

$$GRC (X_{0k}, X_{ik}) = \frac{\min_{\forall j} \min_{\forall k} | X_{0k} - X_{jk} | + \alpha \max_{\forall j} \max_{\forall k} | X_{0k} - X_{jk} |}{| X_{0k} - X_{ik} | + \alpha \max_{\forall j} \max_{\forall k} | X_{0k} - X_{jk} |} \tag{3}$$

Hereinto, $\alpha \in [0, 1]$, (generally $\alpha = 0.5$, $i=j=1, 2, \ldots, n$, $k=1, 2, \ldots, m$) and $GRC(X_{0k}, X_{ik}) \in [0, 1]$ represents the level of similarity of cases x_0 and x_i on attribute A, so the calculation formula (4) for Grey similarity of the similarity level between cases x and x_i is determined to be :

$$GRG(x_0, x_i) = \frac{1}{m}\sum_{k=1}^{m} GRC(X_{0k}, X_{ik}), \quad i=1, 2, 3 \ldots n \tag{4}$$

If $GRG(x_0, x_1) > GRG(x_0, x_2)$, it shows that the level of similarity between x_0 and x_1 is smaller than the level between x_0 and x_2. After the k smallest values of $GRG(x_0, x_i)$ are computed, the most similar k cases can be identified. Finally, the missing values can be replaced by the values which occur most frequently in each attribute column.

2.3. GMKNN Algorithm Description

Euclidean distance is confined to some continuous attributes rather than discrete attributes and the correlation between each variable are not taken into account. Therefore, in order to overcome the shortcoming of Euclidean distance, and strengthen the degree of similarity between missing data and complete data, Gray Relational Grade formula is used to replace Euclidean Distance formula of MKNNI algorithm in GMKNN algorithm. The Pseudo code of GMKNN algorithm is as follows:

Algorithm: GMKNN

Input: α, k , D.
Output: *A complete dataset D'.*
1. Divide the dataset into several parts, $D=\{D_1, D_2 \ldots D_k\}$;
2. ***for*** *each x_0, $x_i \in D_k$* ***{*** *// x_0 is imcomplete case*
3. Calculate $GRG(x_0, x_i)$ according to formula (4), and let $g_1 = GRG(x_0, x_i)$;
4. Sort g_1 from min to max;
5. Select the kth g_1 cases, $N_k(x)=\{x_1, x_2, x_3, \ldots, x_k\}$;
6. ***For*** *each neighbor $x_k \in N_k(x)$ { // x_k is complete case*
7. Calculate $GRG(x_k, x_i)$ according to formula (4), and let $g_2 = GRG(x_k, x_i)$;
8. Sort g_2 from min to max;
9. Select the kth g_2 cases, $N_k(x_k) = \{x_1', x_2', x_3', \ldots, x_k'\}$;
10. ***if*** *($x_k \in N_k(x)$ and $x_0 \in N_k(x_k)$) fill in $M_k(x)$ according to formula (2); }*
11. ***End*** *foreach;*
12. x_0 is replaced by most frequently values in each column in $M_k(x)$; }
13. ***End*** *foreach; // End of cycle*
14. ***Return*** *$D' = \{D_1', D_2', \ldots, D_k'\}$.*

3. Experiments and Results

3.1. Dataset Description

The criminal dataset in this experiment contains 69819 instances, which includes 15 condition attributes and 1 decision attribute in every instance. The condition attributes include sex, age, height, nationality, marital status, profession, cultural level, politics status and other essential information; Criminal-type is the decision

attribute. There are 6 kinds of condition attributes with missing values, which are all discrete types. For example, the missing values of attributes Professional, Religion, Marital-status, Cultural-level etc, are up to 54084, 26354, 20170 and 13476 respectively. In order to get suitable classification results, the attribute types are proposed by different law-enforcement agencies in various ways, including theft, traffic violations, fraud, sex crime, gang/drug offenses and violent crime [16].

3.2. Result Analysis

In this paper, data processing and classification methods are implemented by using Java language based on Weka platform. Experimental environment include an Intel Celeron 1.7GHz machine with 6GB memory.

The experiments are carried out to fill this criminal dataset using five filling algorithms, including Maximum Class algorithm, Information Gain algorithm, GBWKNN algorithm, MKNNI algorithm and GMKNN algorithm. Meanwhile, after the dataset is filled completely, three kinds of classifiers are built to train these datasets, including Naive Bayesian, C4.5 and KNN classifiers. Then ten-fold cross-validation is used to train models. Finally, the optimal model can be obtained by comparing classification accuracy of three classification models.

The experimental results show that values of *k* affect greatly the classification results for three KNNI based algorithms. Therefore, the optimal values of *k* in these three filling algorithms need to be achieved by analyzing classification accuracies, and 30 values of *k* are chosen to fill the data respectively. As shown in Figure 1, when the value of *k* is 70 for GWBKNN algorithm, the accuracy is higher, which reaches 72.96%. As shown in Figure 2, when the value of *k* is 80 for MKNNI algorithm, the accuracy is higher, which reaches 69.2605%. As shown in Figure 3, for GMKNN filling algorithm, the optimal value of *k* is determined to be 120, and the higher accuracy can be obtained, which is up to 77.837%. Finally, it is obvious that GMKNN filling algorithm can obtain higher classification accuracy than the others. According to the figures, compared with Naive Bayesian classification algorithm, KNN and C4.5 classification algorithms can get higher classification accuracy.

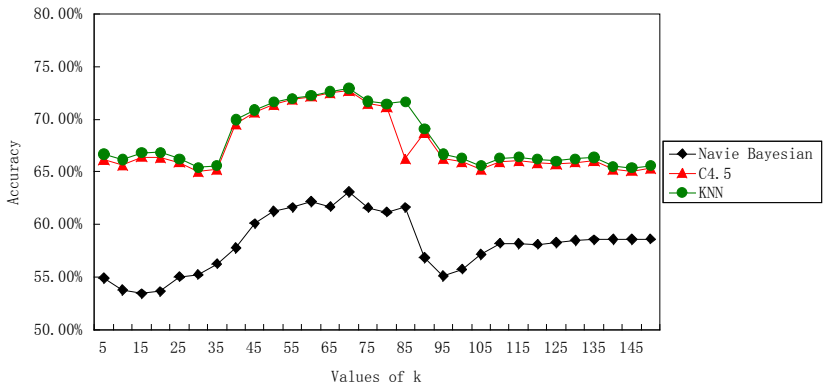

Figure 1. Effect on Classification Accuracy with Different Values of k for GBWKNN Algorithm

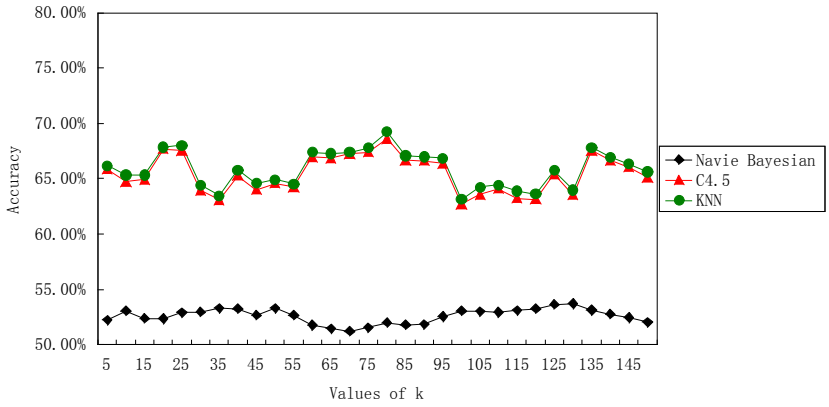

Figure 2. Effect on Classification Accuracy with Different Values of k for MKNNI Algorithm

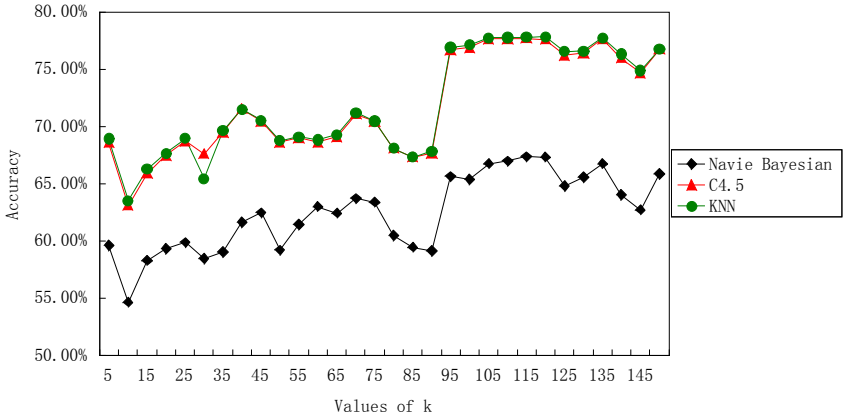

Figure 3. Effect on Classification Accuracy with Different Values of k for GMKNN Algorithm

As shown in Table 1, when existing vast missing values, dataset without any processing has lower classification accuracy. Therefore, dataset needs to be filled to get higher classification accuracy. According to Table 1, compared with several popular missing data filling algorithms, if GMKNN algorithm is used to fill missing data, C4.5 and KNN classification algorithms can offer higher classification accuracy, which is more than 77%.

Table 1. Comparison of Classification Accuracy of GMKNN Algorithm and Several Popular Data Filling Algorithms

Data filling algorithms	Classification algorithms		
	Navie Byeasian	C4.5	KNN
No Processing	53.4213%	56.3887%	56.9552%
Maximum Class	53.9065%	56.6293%	56.7911%
Information Gain	55.0552%	59.9378%	60.3704%
GBWKNN	63.0903%	72.7152%	72.9587%
MKNNI	53.7218%	68.626%	69.2605%
GMKNN	67.3642%	77.7066%	77.837%

4. Conclusions

Predicting crime types using classification algorithms of data mining plays an important role in analysis of criminal behavior. In fact, there are often lots of missing values in actual collected criminal datasets, which seriously reduces the classification accuracy. In order to improve the classification accuracy for predicting crime types, in this paper, a new data filling algorithm—GMKNN algorithm is proposed combining MKNNI algorithm and Grey System Theory. Compared with several popular data filling algorithms based on a real criminal dataset, GMKNN can work better to improve the classification accuracy for selected classification algorithms including Bayesian, C4.5 and KNN algorithm. The highest accuracy is up to 77. 837 %.

The criminal dataset used in this paper comes from the actual, and missing values only exist in the discrete attributes. Therefore, the algorithm proposed pays more attention to deal with the discrete attributes with the aim of getting higher classification accuracy. In future work, for the purpose of improving the adaptability of the algorithm, it will be considered that the filling strategies are adaptively adjusted according to types of the attribute.

Acknowledgement

The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] R. J. Little and D. B. Rubin, "Statistical Analysis with Missing Data", New York: John Wiley, 2nd, **(2002)**.
[2] G. Jagannathan and R. N. Wright, "Privacy-preserving imputation of missing data", Data & Knowledge Engineering, vol. 65, no. 1, **(2008)**, pp. 40-56.
[3] X. Y. Liu, "A hybrid method of missing value filling", Information Technology (Academic), no. 27, **(2007)**, pp. 418-420.
[4] Z. F. Qiao, F. Z. Tian, H. K. Huang and J. N. Chen, "A Comparison Study of Missing Value Datasets Processing Methods", Journal of Computer Research and Development, vol. 43, no. 1, **(2006)**, pp. 171-175.
[5] X. Y. Liu and G. C. Nong, "Comparing Several Popular Missing Data Imputation Methods", Journal of Nanning Teachers College, vol. 24, no. 3, **(2007)**, pp. 148-150.
[6] J. K. Dixon, "Pattern recognition with partly missing data", IEEE Transactions on Systems, Man and Cybernetics, no. 10, **(1979)**, pp. 617-621.
[7] J. V. Hulse and T. M. Khoshgoftaar, "Incomplete-case nearest neighbor imputation in software measurement data", Information Sciences, vol. 259, no. 2, **(2014)**, pp. 596-610.
[8] Z. Qin, "Information Gain based Algorithm for Filling Missing Data", Microcomputer Information, no. 12, **(2007)**, pp. 180-186.
[9] Z. K. Chen, A. L. Lv and Q. C. Zhang, "A New Algorithm for Imputing Missing Data Based on Distinguishing the Importance of Attributes", Microelectronics & Computer, no. 7, **(2013)**, pp.167-172.
[10] N. Xi, "A Filling Method of Rough Set based on Kernel Values", Silicon Valley, no. 8, **(2014)**, pp. 61-63.
[11] G. M. Sang, K. Shi, Z. Liu and L. J. Gao, "Missing Data Imputation Based on Grey System Theory", International Journal of Hybrid Information Technology, vol. 7, no. 2, **(2014)**, pp. 347-355.
[12] M. L. Zhu, "MkNNI: New Missing Value Imputation Method Using Mutual Nearest Neighbor", Modern Computer, no. 11, **(2012)**, pp.8-10.
[13] Y. M. Xu and C. Chen, "APT-KNN: An Efficient Missing Value Imputation Method", Computer Applications and Software, vol. 28, no. 4, **(2011)**, pp. 135-139.
[14] S. G. Liao, Y. Lin and D. D. Kang, "Missing value imputation in high-dimensional phenomic data: imputable or not, and how?", BMC Bioinformatics, vol. 15, no. 11, **(2014)**, pp. 346- 348.
[15] X. Y. Liu, "Improved KNN Algorithm Based on Mahalanobis Distance and Gray Analysis", Journal of Computer Applications, no. 9, **(2009)**, pp. 2502-2504.
[16] H. Chen, W. Y. Chung and J. J. Xu, "Crime Data Mining: A General Framework and Some Examples", Computer, vol. 37, no. 4, **(2004)**, pp. 50-60.

