Detecting Crime Types Using Classification Algorithms

Cui-cui Sun[1], Chun-long Yao[1], Xu Li[1], Kejun Lee[2]
[1]School of Information Science and Engineering
Dalian Polytechnic University, Dalian, China
[2]Technical University of Varna
Varna 9010, Bulgaria
yaocl@dlpu.edu.cn, kjlee59@gmail.com

Journal of Digital
Information Management

ABSTRACT: *Criminal behaviors can reflect the characteristics of the criminals to a great extent. To predict the crime types according to characteristics of vast amounts of criminals is an important part of criminal behavior analysis. In order to get high classification accuracy, three typical classification algorithms, including C4.5 algorithm, Naive Bayesian algorithm and K nearest neighbor (KNN) algorithm, are compared using several popular missing data filling algorithms respectively based on a real crime dataset with lots of missing data. The experimental results show that higher classification accuracy can be obtained by combining KNN classification algorithm and GBWKNN missing data filling algorithm which is based on grey relational analysis (GRA) theory.*

Subject Categories and Descriptors
H.2.8 [Database Management]: Database Applications-Data Mining

General Terms: Algorithms, Criminal Analysis

Keywords: Crime, Classification algorithm, Missing data filling, Classification accuracy

Received: 11 November 2012, Revised 13 March 2013, Accepted 19 March 2013

1. Introduction

A large amount of criminal data has been accumulated by police offices and other criminal agencies for criminal investigation. The criminal data contains a large amount of knowledge which is useful to prevent and combat crime. Therefore, it is important to analyze effectively the data to obtain the relations and rules implied in criminal information. In recent years, the application of data mining in criminal analysis has been received more attention. For example, using the association rule mining, the relationships among crime characteristics can be obtained to guide the police for tracing the source of crime. Large crime datasets can be analyzed more effectively using the data mining technology, which can help the police to predict the criminal inclinations and fight against crime in time.

At present, there exists several typical crime data mining methods, including classification analysis, association rule analysis and clustering analysis. Association rule analysis is usually used to find the relationships among criminal behaviors from the criminal dataset, which can help the police to translate the data resource into practical detection ability. For example, Applying association rule to economic criminal analysis can improve the efficiency of law enforcement and crime detection. Xie [1] applied Apriori algorithm to funds fraud analysis based on criminal data. Clustering analysis can help to detect relationship among criminals. Nath [2] used K-means clustering algorithm to help to detect the crime patterns and speed up solving crime incidents. The SOM clustering algorithm [3] is put forward to identify crime characteristics and categorize crimes in intelligent crime analysis.

Different from the above mentioned methods, the classification algorithms and models aim to find the factors affecting the crime and help the police officers to strengthen crime preventions. In order to find the crime trends and the source of crime based on a simulated criminal behavior dataset, Huang et al. [4] used Id3 algorithm to classify criminal behaviors and discover relationship between criminal behaviors and criminal characteristics, such as economic foundation, age, education level and family environment. Yu et al. [5] used some classification algorithms to predict crime hot spots based on dataset that contains aggregated counts of crime, location and time of the crime-related events. But the work needs to find the crime trends according to criminal features and the predicted types of crime. Based on criminal population conviction histories of recent offenders, Tollenaar et al. [6] proposed a prediction model combining K nearest neighbor algorithm and Linear SVM algorithm to predict three types of criminal recidivism, including general recidivism, violent recidivism and sexual recidivism. Some classical types of criminals such as traffic violations, fraud are not involved in [6]. Zhou [7] selected some classification algorithms including ID3, C4.5 and Naive Bayesian algorithm, to analyze the dataset of criminals in order to find out the factors affecting crime according to criminal backgrounds, psychological characteristics and genetic characteristics. Although the classification analysis plays an important role in the analysis of criminal behavior, few classification algorithms are currently applied to this field. In addition to the above-mentioned classification algorithms, there are also BP Neural Network algorithm, Genetic algorithm and some other typical algorithms [8]. Although these algorithms are rarely used for the analysis of crime data, they can be effectively applied to other areas to get good results. Generally, the dataset is used to test the performance of the classification algorithms, but the actual dataset collected usually has missing values that can affect the classification accuracy. Therefore, it is important for improving the classification accuracy to fill the missing values of a dataset to make it complete. The quality of a dataset directly influences the result of classification. If there is no a complete dataset, any effective classification algorithm will lost its original advantages. The presence of missing values can cause an algorithm to give unreliable results. Therefore, it is significant to adopt proper data preprocessing methods for solving the problem of data missing. At present, there are some methods dealing with missing values. For example, case-wise deletion method [9] is a common method. The method can causes a great waste of resources and discards an amount of useful information hidden in these objects. The mean value substitution method [10] replaces a missing value with an average value for each attribute. This method may increase falsely the precision of the estimates, and cause the uncertainty and give biased results. The ANO (average nearest observation) algorithm [11] replaces a missing value with the average of the nearest previous and next observations. The ANO algorithm can only describe local variations while ignoring the global effects. Maximum class algorithm [9] replaces a missing value with the most frequent value for each attribute. Liu [12] proposed a filling method by combining the KNN algorithm and the kernel function method. This method can deal with discrete and continuous missing values. Sang et al. [13] proposed a new weighted KNN data filling algorithm based on Grey correlation analysis (GBWKNN) by combining with the nearest neighbor algorithm. This algorithm can effectively work when missing data is not sensitive to noise data.

In fact, this section has been devoted to a brief survey on related researches and existing intelligent crime analysis methods. In this paper, the objective is to test several popular missing data filling methods and classification algorithms using a real dataset collected by Dalian Police Bureau, so that good approaches can be found to get higher classification accuracy. This dataset has a large number of missing values, which will bring great difficulties to get accurate classifications.

The rest of paper is organized as follows: Section 2 describes the process of building classifiers as well as several popular data filling algorithms and classification algorithms. Section 3 illustrates the results and performance of selected algorithms. Section 4 presents conclusion.

2. Building Classifier

In general, building a classifier need two steps, including data preprocessing and selecting a proper classification algorithm to train the classification model using a dataset got by preprocessing. Data preprocessing can improve the quality of the data mining model and reduce the time required. The methods of data preprocessing include missing values processing, noise data processing, data transformation, data reduction and data discretization etc. The dataset produced by data preprocessing will be provided to the classification algorithm for training the classification model. A best classification algorithm should be selected to draw the classification rules by analyzing the training dataset.

2.1 Data Filling Algorithms

The data filling algorithm aims to adopt effective ways to complement the missing attribute values in order to provide a complete dataset. In this paper, for missing values processing, three data filling algorithms are selected from the current popular algorithms, including Maximum class algorithm [9], Roulette algorithm [14] and GBWKNN algorithm [13].

1. Maximum class algorithm

For each attribute, considering the missing attribute values are all discrete, the first step is to find out the attribute values that occur with the highest frequency. Then, the missing values can be replaced respectively with these substituted attribute values.

2. Roulette algorithm

Similar to roulette game, this is a common way of random selection in genetic algorithms. The selection probability [14] of each value needs to be calculated for each attribute. For any attribute, the selection probability of an attribute value is the frequency of occurrence of the value accounted for the proportion of all values of the attribute. Obviously, an attribute value with greater selection probability has more chances to be selected to fill the missing value.

3. GBWKNN algorithm

This algorithm combines the advantages of the Grey System Theory with the K nearest neighbor algorithm [13]. It is a measuring method of confirming the similarity between two data records using Grey System Theory. Initially, we can divide the dataset into several parts based on the types of decision attribute. In every part, the dataset $D = \{x_0, x_1, \ldots, x_n\}$, n is the number of cases, $x_i = \{x_i(1), x_i(2), \ldots, x_i(m)\}$, $i = 0, 1, 2, \ldots, n$, and m is the number of the types of condition attributes in each case. For each case x_0 with missing values, the values are computed on the gray relativity between this case and each case x_i with no missing values, then the Grey relationship coefficient of the two cases on attribute A is:

$$GRC\,(x_0(A), x_i(A)) = \frac{min\ \forall j\ min\ \forall k\, |\, x_0(k) - x_j(k) + \alpha\, max\ \forall j\ max\ \forall k\, |\, x_0(k) - x_j(k)\,|}{|\, x_0(A) - x_i(A)\,| + \alpha\, max\ \forall j\ max\ \forall k\, |\, x_0(k) - x_j(k)\,|} \quad (1)$$

Here into, $\alpha \in [0, 1]$, (generally, $a = 0.5$, $i = j = 1, 2, \ldots, n$, $A = k = 1, 2, \ldots, m$) and $GRC\,(x_0(A), x_i\,(A)) \in [0, 1]$ represents the level of similarity of cases x_0 and x_i on attribute A, so the calculation formula for grey similarity of the similarity level between cases x_0 and x_i is determined to be:

$$GRG\,(x_0, x_i) = \frac{1}{m} \sum_{A=1}^{m} GRG\,(x_0(A), x_i(A)),\ i = 1, 2, 3, \ldots, n \quad (2)$$

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. When $GRG\,(x_0, x_i) = l$, it shows that the two cases have no relationship, and $GRG\,(x_0, x_i) = 0$, it shows that the two cases are almost the same.

Finally, according to the K nearest neighbor algorithm, the value of K is uncertain. The K smallest values of $GRG\,(x_0, x_i)$ can be computed, and then the most similar K cases can be identified. The missing values can be replaced by the value which occurs most frequently in each attribute column based on the maximum class principle.

2.2 Classification Algorithms

Classification algorithms establish a model and utilize it to predict the categorical labels of unknown objects in order to distinguish different classes. These categorical labels are predefined, discrete and unordered [15]. Based on their advantages, three classification algorithms are selected, including Naive Bayesian Classification algorithm [8], C4.5 Classification algorithm [8] and KNN Classification algorithm [16].

1. Naive Bayesian Classification Algorithm

This classification algorithm predicts the possibility of a class relation pattern when the prior probability and conditional probability are known, which is based on Bayesian theory of probability statistics. To compute which one belongs to a specific class, we choose the final category that its probability is the maximum of the sample [17]. This algorithm has the following advantages: easy to implement, stable classification results, high efficiency. However, this algorithm generally is assumed that every attribute is independent of each other. In fact, this assumption is not always correct. Therefore, the classification performance may be affected.

2. C4.5 Classification Algorithm

C4.5 builds decision trees from a set of training data using the concept of information entropy. For C4.5, splitting attributes are selected by computing information gain ratio, and discrete and continuous attributes can be processed. [18-19]. C4.5 algorithm has the following advantages and disadvantages: the classification rules are easy to understand and always have a high accurate rate; in the tree structure, a number of data sets are needed to be scanned and sorted, thus leading to reduce the effectiveness of the algorithm.

3. KNN Classification Algorithm

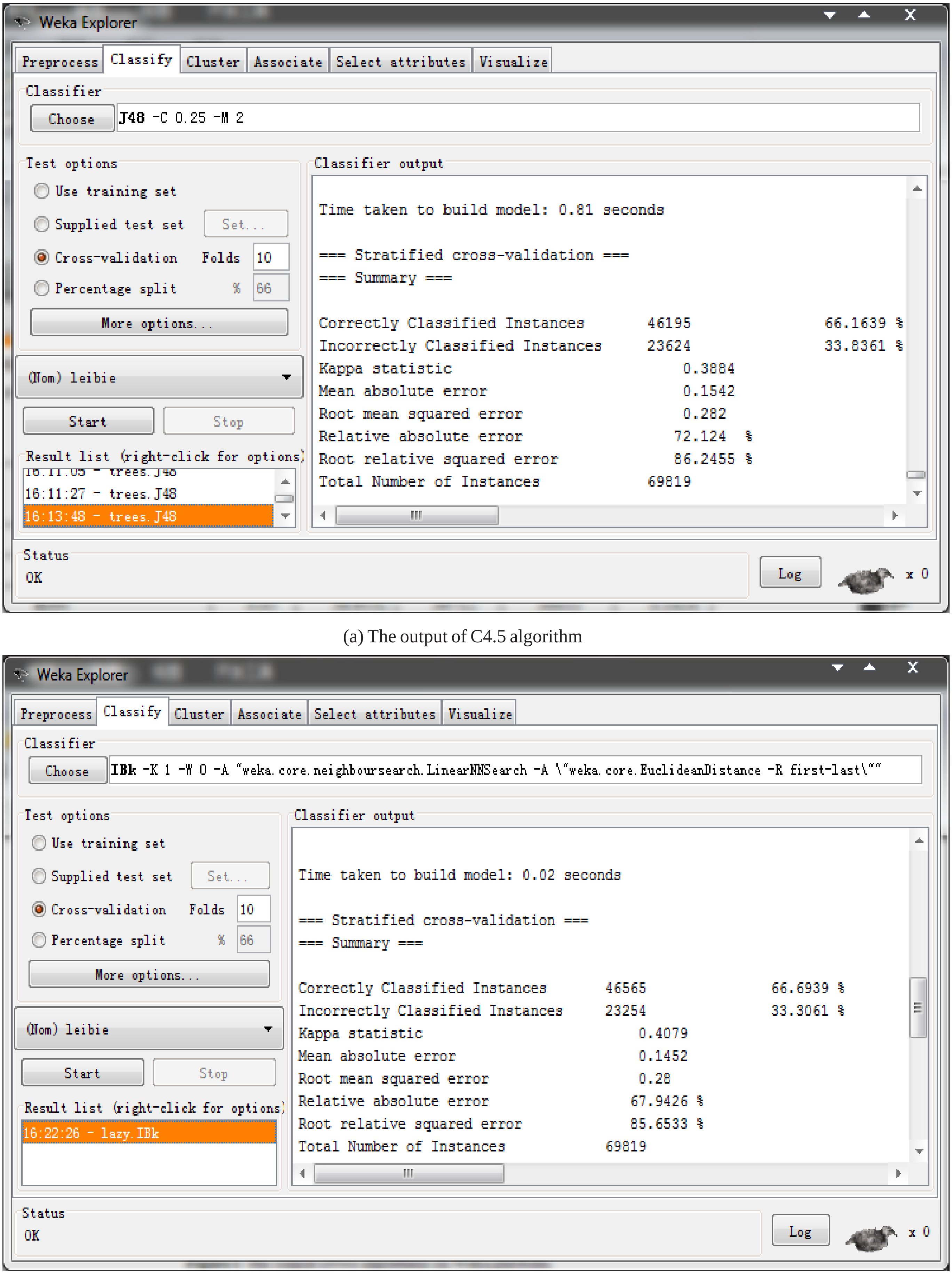
This algorithm determines the K neighbors according to the minimum distance from the query instance to the training samples [20]. It predicts the category of test samples according to the nearest K training samples, and judges the largest probability of those belonging to a category. The algorithm has the following advantages and disadvantages: It can reduce the adverse effects caused by improper classification feature and minimize the error term in the process of classification. KNN classification algorithm has a high computing complexity, and the accurate rate will be lower when the training samples are huge.

3. Experiments and Results

In order to detect more accurately crime types, experiments are carried out by the comparison of several popular classification algorithms as well as classical data filling algorithms based on the real criminal dataset collected by the police system.

3.1 Dataset Description

In the experiments, the criminal dataset contains 69819 instances, which has 1 decision attribute and 15 condition attributes in every instance. Criminal-type is the decision attribute and the condition attributes include age, height, nationality, sex, profession, cultural level, politics status, marital status and other essential information. There are 6 kinds of condition attributes with missing values, and all of these condition attributes are discrete types. For



Classifier: J48 -C 0.25 -M 2

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) leibie

Result list (right-click for options)
16:11:05 - trees.J48
16:11:27 - trees.J48
16:13:48 - trees.J48

Classifier output

```
Time taken to build model: 0.81 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       46195               66.1639 %
Incorrectly Classified Instances     23624               33.8361 %
Kappa statistic                          0.3884
Mean absolute error                      0.1542
Root mean squared error                  0.282
Relative absolute error                 72.124  %
Root relative squared error             86.2455 %
Total Number of Instances            69819
```

Status
OK

(a) The output of C4.5 algorithm

Classifier: IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) leibie

Result list (right-click for options)
16:22:26 - lazy.IBk

Classifier output

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       46565               66.6939 %
Incorrectly Classified Instances     23254               33.3061 %
Kappa statistic                          0.4079
Mean absolute error                      0.1452
Root mean squared error                  0.28
Relative absolute error                 67.9426 %
Root relative squared error             85.6533 %
Total Number of Instances            69819
```

Status
OK

(b) The output of KNN algorithm ($K = 1$)

Figure 1. The output of two algorithms on Weka platform

example, with regard to the attributes Cultural-level, Marital-status, Religion and Professional etc, missing values are up to 13476, 20170, 26354, and 54084 respectively. In order to get suitable classification results, the attribute types are proposed by different law-enforcement agencies in various ways, including traffic violations, theft, fraud, sex crime, gang/drug offenses and violent crime [17].

3.2 Result Analysis

In order to test the algorithms selected, the related algorithms are implemented using Java language based on the Weka platform, and experiments were performed on a Intel Celeron 1.7GHz machine with 6GB memory.

To analyze the effects about the processing of missing values, the experiments are carried out to fill the data among three filling algorithms, including GBWKNN algorithm, Roulette algorithm and Maximum Class algorithm. Meanwhile, when the dataset is filled completely using these algorithms, three kinds of classifiers are built totrain these complete datasets, including Naive Bayesian, C4.5 and KNN classifier. In KNN classification algorithm, the effect of K is not very obvious, so the value of K is set to the default value 1. Then ten-fold cross-validation is used to estimate the performance of each model, and the most effective consequence is shown. Finally, the optimal model can be found by comparing classification accuracy of three classification models.

Some performance indicators including building time, classification accuracy and Kappa Statistic, etc, are used to evaluate performance of the algorithms selected. Figure 1 shows the output of KNN algorithm and C4.5 algorithm on the WeKa platform when GBWKNN algorithm $(K = 5)$ is used to fill the dataset. According to figure 1, the performance indicators of the algorithms can be clearly displayed.

Data Filling Algorithms	Classification Algorithms	Building time (s)	Accuracy (%)	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic
No filling	Naive Bayesian	0.13	53.4213	37248	32477	0.0768
	C4.5	4.04	56.3887	39317	30408	0.0898
	KNN	0.05	56.9552	39712	30013	0.1424
Maximum class	Naive Bayesian	0.09	53.9065	37637	32182	0.0866
	C4.5	2.12	56.6293	39538	30281	0.1041
	KNN	0.04	56.7911	39651	30168	0.1406
Roulette	Naive Bayesian	0.05	53.2147	37154	32665	0.0615
	C4.5	0.94	56.2884	39259	30560	0.089
	KNN	0.05	53.9724	37625	32194	0.1122
GBWKNN (K=5)	Naive Bayesian	0.04	54.8833	38319	31500	0.1778
	C4.5	0.81	66.1639	46195	23624	0.3884
	KNN	0.02	66.6939	46565	23254	0.4079

Table 1. Comparison of classification performance on criminal dataset

Predict class Actual class	theft	traffic violations	fraud	sex crime	gang/drug offenses	violent crime
theft	35624	138	272	126	198	1871
traffic violations	678	725	91	106	104	751
fraud	1135	56	3827	54	160	1503
sex crime	1982	283	80	1089	144	183
gang/drug offenses	785	155	176	190	1969	1152
violent crime	5725	103	319	59	301	7705

Table 2. Confusion matrix on criminal data

In the experiment, classification accuracy is taken as the standard to evaluate the algorithms. As shown in table 1, when the missing data is replaced by using the GBWKNN filling algorithm, C4.5 and KNN classification algorithms present higher classification accuracy, and its accuracy rate reaches more than 66%. It is also obvious to find that a classification model can be faster built using KNN algorithm than other algorithms. The Kappa Statistic as the main metric is used to measure an attribute measurement system, and it represents the level of agreement between the predicted results of the classifier and the actual classification results. As shown in table 1, the experimental results show that when the GBWKNN filling

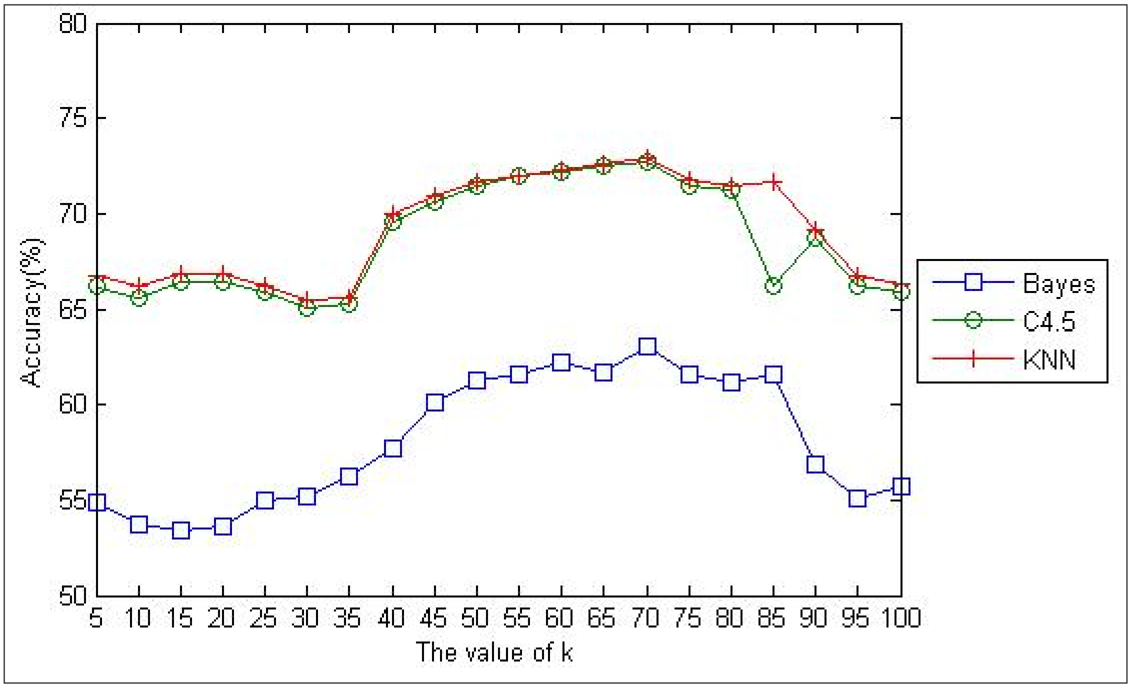


Figure 2. Influence on classification accuracy of GBWKNN algorithm with different values of K

algorithm and KNN classification algorithm are used for the dataset, Kappa Statistic value is up to 0.4079 and higher accuracy is obtained.

It is clear that GBWKNN algorithm is convergent under different missing attribute types and the value of K needs to be determined by experiment. Here, K represents the number of nearest cases. For the formula (1), set $a = 0.5$ as a general value. The experimental results show that values of K affect to a large extent the results of the GBWKNN algorithm. Therefore, it is very important for GBWKNN algorithm to find the optimal value of K. In the experiment, 20 values of K are chosen to fill the data respectively. The optimal value of K can be determined by comparing the classification accuracy achieved by three classification algorithms. As shown in Figure 2, when the value of K is 70, the accuracy rate is higher than other values of K. Especially the accuracy rate reached 72.9587 % for the KNN classification algorithm. Finally, the optimal value of K is determined to be 70 for the GBWKNN algorithm.

A confusion matrix, also known as a contingency table or an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The confusion matrix stems from whether the system is confusing two classes. As shown in table 2, classification methods has been trained to distinguish among traffic violations, theft, fraud, sex crime, gang/drug offenses and violent crime. For example, in this confusion matrix, 35624 cases were predicted accurately as theft in actual 38229 thefts, while the other 2605 cases are predicted falsely as other types such as traffic violations, fraud.

4. Conclusions

Classification analysis can find the factors affecting the crime and help the police officers to strengthen crime preventions. There are many missing values in actual criminal dataset, which affects the classification accuracy. In this paper, Maximum class filling algorithm, Roulette filling algorithm and GBWKNN filling algorithm are used to fill the real crime dataset in order to obtain the complete dataset. Finally, the classification accuracy of three algorithms, including C4.5 algorithm, Naive Bayesian algorithm and KNN algorithm, are compared based on dataset got by using above filling algorithms respectively. The experimental results show that higher accuracy can be achieved by combining GBWKNN filling algorithm and KNN classification algorithm.

At present, for the crime dataset with lots of missing values, classification accuracy of crime types still needs to be further improved. In fact, these filling algorithms used are inadequate to a certain extent. For example, only missing values of the discrete attributes are filled without considering continuous attributes. Therefore, the missing data filling algorithms that can improve further the classification accuracy still need further investigation.

References

[1] Qinchuan Xie. (2012). The Research and Application of Data Mining Technology in Economic Crime Investigation. Netinfo Security, (12) 36-38. (In Chinese).

[2] Shyam Varan Nath.(2006). Crime Pattern Detection Using Data Mining. *In*: Proceeding of Web Intelligence and Intelligent Agent Technology Workshops, p. 41-44.

[3] Reza Keyvanpoura Mohammad, Javideh Mostafa, et al. (2011). Detecting and investigating crime by means of data mining: a general crime matching framework, Procedia Computer Science, 03, p. 872–880.

[4] Jianshe Huang, Qifu Yao. (2005). The Application of Data Mining Technique on Crime Analysis. *Journal of Zhejiang Business Technology Institute*, 4 (3) 45-47. (In Chinese).

[5] Chung-Hsien Yu, Max W. Ward,et al. (2011). Crime Forecasting Using Data Mining Techniques, *In*: Proceeding of 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pages 779-786

[6] N. Tollenaar, P. G. M. van der Heijden. (2013). Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models, *Journal of the Royal Statistical Society: Series A* (Statistics in Society), 176 (2) 565–584.

[7] Shuai Zhou. (2012). Crime related factors analysis based on data mining technology. Master Thesis, Dalian Maritime University. (In Chinese).

[8] Lingli Li. (2011). A Review on Classification Algorithms in Data Mining. *Journal of Chongqing Normal University: Natural Science Edition*, 28 (4) 44-47. (In Chinese).

[9] Xingyi Liu, Guocai Nong.(2007). Comparing Several Popular Missing Data Imputation Methods. *Journal of Nanning Teachers College*, 24 (3) 148-150. (In Chinese).

[10] Jinsheng Huo, Cox Chris, D., Seaver William, L., et al.(2010). Application of Two-Directional Time Series Models to Replace Missing Data, *Journal of Environmental Engineering-asce-J ENVIRON ENG-ASCE*, 136 (4) 435-443.

[11] Karahalios Amelia, Baglietto Laura, Carlin John, B., et al. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures, BMC Medical Research Methodology, 12 (7) 96-105.

[12] Xingyi Liu. (2007). A hybrid method of missing value filling. Science Information (Academic), (27) 418-420. (In Chinese).

[13] Guoming Sang, Kai Shi, Zhi Liu, Lijun Gao. (2014). Missing Data Imputation Based on Grey System Theory, *International Journal of Hybrid Information Technology*, 27 (2) 347-355.

[14] Geng Zhu. (2007). C + + Implementation of Genetic Algorithms and Selection of Roulette. *Journal of Dongguan University of Technology*, 14 (5) 70-74. (In Chinese).

[15] Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques, Second edition, Morgan Kaufmann Publishers, 285–464.

[16] Gongde Guo, Hui Wang, David Bell, et al. (2006). Using kNN model for automatic text categorization, *Soft Computing,* 10(5) 423-430.

[17] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, et al. (2004). Crime Data Mining: A General Framework and Some Examples, *Computer,* 37 (4) 50-60.

[18] J. R. Quinlan.(1993). C4.5: programs for machine learning, Morgan Kaufmann.

[19] Wei Dai, Wei Ji.(2014). A MapReduce Implementation of C4.5 Decision Tree Algorithm, *International Journal of Database Theory and Application*, 7(1) 49-60.

[20] Bawaneh Mohammed, J., Alkoffash Mahmud, S., Al Rabea Adnan, S.(2008). Arabic text classification using K-NN and Naive Bayes. *Journal of Computer Science*, 4 (7) 600-605.