2014

ISSN : 0974 - 7435 Volume 10 Issue 24

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(24), 2014 [15451-15457]

Detecting crime types using classification algorithms

Cuicui Sun, Chunlong Yao*, Xu Li, Xiaoqiang Yu
School of Information Science and Engineering, Dalian Polytechnic University,
Dalian116034, (CHINA)
E-mail: chunlongyao@163.com

ABSTRACT

Criminal behavior reflects the characteristics of the criminals. To infer the types of unknown criminals from vast amounts of different crime characteristics is an important part of criminal behavior analysis. It is a good solution to classify the criminals using classification algorithms. Three typical classification algorithms are used to analyze the criminal datasets in this paper, including C4.5 algorithm, Naive Bayesian algorithm and K nearest neighbor (KNN) algorithm. However, quite a lot of missing data values can result in a seriously effect on the classification accuracy. Therefore, the missing data filling method which fills missing data based on grey relational analysis (GRA) theory is used. The experimental results on the criminal dataset show that higher classification accuracy can be obtained using this missing data filling method.

KEYWORDS

Crime; Classification algorithm; Missing data filling; Classification accuracy.

INTRODUCTION

A vast amount of criminal data has been accumulated by police offices and other criminal agencies as an essential component of criminal investigation. The data contains a large number of knowledge which is useful to help preventing and combating crime. Therefore, it is important about how to analyze effectively the data to obtain the relationships and rules of criminal information. In recent years, the application of data mining in crime analysis has been paid more attention. For example, using the association rule mining, the association information among criminal characteristics can be obtained to guide the police in tracing the source of crime. The information with different criminal characteristics can be analyzed more effectively using data mining, which can help the police infer the criminal propensity and fight against crime activities in time.

At present, there exists several typical crime data mining methods, including classification analysis, association rule analysis and clustering analysis etc. Association rule analysis is usually used to find the relationship among criminal behaviors from the criminal dataset, which can help the police translate the information resource advantage into realistic detection ability. For example, the application of association rule in economic crime analysis, e.g. Apriori algorithm is applied for funds fraud analysis based on criminal data and case data[1], can improve the economic efficiency of law enforcement and crime detection level. Clustering analysis can help detect crimes and crimes' relationships with criminals. In[2], K-means clustering algorithm is presented to help detect the crimes patterns and speed up the process of solving crime incidents, and the crime data obtained by a sheriff's office can be analyzed to learn technique for knowledge discovery and classify crimes. In[3], the SOM clustering method is put forward to identify crime characteristics and categorize crimes in intelligent crime analysis.

Different from above, the classification method and model plays a decisive effect on criminal analysis, it aims to find the factors affecting the crime and help the police officers strengthen crime preventions. In[4], Id3 algorithm is used for criminal behavior classification based on a simulated criminal behavior dataset to find out the relationship between criminal behavior and the economic foundation, age, education level and family environment etc, in order to find the crime trends and the source of crime. Yu et al.[5] used some classification methods to predict crime hot spots. The datasets contain aggregated counts of crime, the location and time of crime-related events, and these events are categorized by the police department of a United States city in the Northeast. The criminal attributes and the predicted types of crime are used for finding the crime trends. In[6], by using criminal population convictioSn histories of recent offenders, prediction models are developed based on K nearest neighbor classification algorithm and Linear SVM algorithm in order to predict three types of criminal recidivism: general recidivism, violent recidivism and recidivism, an sexual d we can get the criminogenic needs of the offender. But some classical types of criminals such as traffic violations, fraud etc, are not involved in[6]. In[7], some classification methods including ID3 algorithm, C4.5 algorithm and Naive Bayesian algorithm are selected to analyze the dataset of criminals in order to find the factors affecting crime, which are introduced to excavate knowledge from the criminals of background information, the psychological information and the genetic information of criminals. It is clearly that the application of classification mining for crime prevention is meaningful. In addition to the classification algorithms mentioned above, there are also BP Neural Network algorithm, Genetic algorithm and any other typical algorithm[8]. Although these algorithms are rarely used on the analysis of crime data, they can be effectively applied in other areas to get a better result. Generally, the datasets are used to test the classification algorithms' performances, but the actual datasets collected usually have missing values that can affect the classification accuracy. Therefore, it is very important how to fill the missing values to improve the classification accuracy.

However, the quality of the dataset directly influences the result of classification. Any efficient classification algorithm has lost its original advantages without full datasets. It will result to unreliable output when the missing values exist. Therefore, it is important about how to handle data missing problem by choosing appropriate data preprocessing methods. At present, there are some processing methods handling with missing values. For example, case-wise deletion method[9] is a common method. However, it causes a lot of waste of resources and discards a number of useful information hidden in these objects. The mean value substitution method[10] replaces the missing value with the average value calculated over all the values for each attribute. This method falsely increases the stated precision of the estimates due to failing to account for the uncertainty caused by the missing data and give generally biased results. The ANO (average nearest observation) method[11] is used to replace missing values with the average of the nearest previous and following observations. It can only describe local variations while ignoring the global effects. Maximum class algorithm[9] aims to find out the attribute values which has the highest frequency, and then the missing values can be replaced respectively with these substituted attribute values. In[12], the method is proposed which combines the KNN algorithm and the filling method based on the kernel function filling method. This method can dealing with both of the discrete missing data and the continuous missing data. In[13], a new weighted KNN data filling algorithm is proposed, 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.

In fact, this section has been devoted to a brief survey on related researches and existing intelligent crime analysis methods. In this paper, our objective is to test several popular imputation methods and classification methods so that we can find good approaches to get higher classification accuracy for our project based on a real dataset from the police bureau in our city. Different from mentioned above, the dataset have a large amount of missing values, which brings great difficulties to obtain accurate classifications.

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

CONSTRUCTING CLASSIFIER

In general, two steps are needed for building a classifier, including data preprocessing and the selecting appropriate classification methods. Data preprocessing can improve the quality of the data mining model and reduce the time required. The methods of data preprocessing include missing value processing, noise data processing, data transformation, data reduction and data discretization etc. After data preprocessing is done, the optimal classification method are selected from some classical classification methods, and the classification rules can be derived based on types of crime by analyzing contents data which belongs to a certain classification. The criminal data in this paper has a large number of missing values which is used to construct the classifier, in this case, by using this data can result to unreliable output. So missing value processing is necessary before comparing the classification methods.

Data filling algorithms

We choose three data filling algorithms for missing value processing from the current popular algorithms as below, including Maximum class algorithm[9], Roulette algorithm[14] and GBWKNN algorithm[13].

1) Maximum class algorithm

Considering the missing attribution values are all discrete. For each attribute, the first step is to find out the attribute values which occurs the highest frequency, and then the missing values can be replaced respectively with these substituted attribute values.

2) Roulette algorithm

This is a common random selection of the traditional genetic algorithm which is similar to the gambling game of roulette. The individual fitness proportionally converts into the choosing probability. Dividing some sectors in a disc based on the proportion of individuals. When the disc stops after each rotating, the pointer will stop to point the sector which is corresponding to the selected value. Finally, we can use this value to replace the missing value. Obviously, the larger the probability of the individual is, the greater the area occupied by the roulette is, which is selected more chances for filling data.

3) GBWKNN algorithm

This algorithm combines with the advantages of the Grey System Theory and 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 $D=\{x_0, x_1, \ldots, \mathrm{x_n}\}$, 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:

$$GRC(x_0(A), x_i\,(\mathrm{A})) = \frac{\min_{\forall j} \min_{\forall k} |x_0(k) - x_j(k)| + \alpha \max_{\forall j} \max_{\forall k} |x_0(k) - x_j(k)|}{|x_0(A) - x_i(A)| + \alpha \max_{\forall j} \max_{\forall k} |x_0(k) - x_j(k)|} \quad \textbf{(1)}$$

Hereinto, $\alpha \in [0, 1]$, (generally $\alpha = 0.5$, $i=j=1, 2, \ldots, n$, $A=k=1, 2, \ldots, m$) and $GRC(x_0(A), x_i(\mathrm{A})) \in [0, 1]$ 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:

$$GRG(x_0, x_i) = \frac{1}{m}\sum_{A=1}^{m} GRC(x0(A), xi(A)),\ i=1, 2, 3, \ldots, n \quad \textbf{(2)}$$

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 that between x_0 and x_2. In fact, the smaller the value of $GRC(x_0(A), x_i\,(\mathrm{A}))$ is, the more similar the two cases are.

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 cases of K 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. And the complete dataset can be obtained by using this algorithm.

Classification algorithms

Classification builds up a model and utilizes it to predict the categorical labels of unknown objects to distinguish between objects of different classes. These categorical labels are predefined, discrete and unordered[15]. Three classification algorithms are selected below, based on their advantages, 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 know the prior probability and conditional probability, based on Bayesian theory of probability statistics. To compute which belongs to a specific class, we choose the final category which probability is the maximum of the sample[17]. This algorithm has the following advantages: the simple implementation, stable classification effect and the higher rate. However, this algorithm generally is assumed that every attribute is independent of each other. In fact, this assumption is not always right which can have an impact on the classification performance.

2) C4.5 Classification algorithm

C4.5 builds decision trees from a set of training data, using the concept of information entropy. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits data set of samples into subsets that can be one class or the other[18]. And this attribute is considered with the highest normalized information gain. C4.5 algorithm has the following advantages: the classification rules are easy to understand and always have a high accurate rate. Its shortcomings are as follows: 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 works based on minimum distance from the query instance to the training samples to determine the K neighbors[19]. It predicts the test sample’s category according to the *K* training samples which are the nearest neighbors to the test sample, and judge it to that category which has the largest category probability[20]. The algorithm has the following advantages: It can reduce the adverse effects caused by improper classification feature and minimize the error term in the process of classification. Its shortcomings are as follows: KNN Classification algorithm has a high computing complexity. It is inapplicable to a large number of training samples, when the training samples are more, the accurate rate will be lower.

EXPERIMENTS AND RESULTS

In order to detect more accurately types of the criminals, 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.

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 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].

Result analysis

In order to test the algorithms used in this paper, the methods and data processing in this paper are implemented by using Java language and Weka platform. Experimental environment include a PC machine, Oracle11g, Eclipse, Weka etc.

TABLE 1 : Comparison of classification performance indicators on criminal dataset

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
	Naive Bayesian	0.05	53.2147	37154	32665	0.0615

Roulette	C4.5	0.94	56.2884	39259	30560	0.089
	KNN	0.05	53.9724	37625	32194	0.1122
	Naive Bayesian	0.04	54.8833	38319	31500	0.1778
GBWKNN (K=5)	C4.5	0.83	66.1639	46195	23624	0.3884
	KNN	0.01	66.6939	46565	23254	0.4079

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 datasets are filled completely using these algorithms, three kinds of classifiers are built to train 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, so that we show clearly the most effective consequence. Finally, the optimal model can be found by comparing classification accuracies of three classification models.

Some performance indicators are used to check every proposed model performance, which includes building times, classification accuracy and Kappa Statistic etc. In the experiment, classification accuracy is taken as the standard to evaluate the strength and weakness of algorithm. As is shown in TABLE 1, when the missing data were replaced by using the GBWKNN filling algorithm, it presents the higher classification accuracies in C4.5 and KNN classification algorithms, and both of the accuracy rates reach more than 66%. It is also obvious to find that the model of KNN Classification algorithm is built faster than others. The Kappa Statistic is the main metric used to measure how good or bad an attribute measurement system is, and it represents the level of agreement between the predicted results of the classifier and the actual classification results. The experimental result in TABLE 1 shows that by using *GBWKNN* filling algorithm and KNN classification algorithm always have an advantage which reaches 0.4079. In a conclusion, the filling effect of GBWKNN algorithm in the paper is better than the other filling algorithms, and in this case, the accuracy of KNN classification algorithm is usually higher than others.

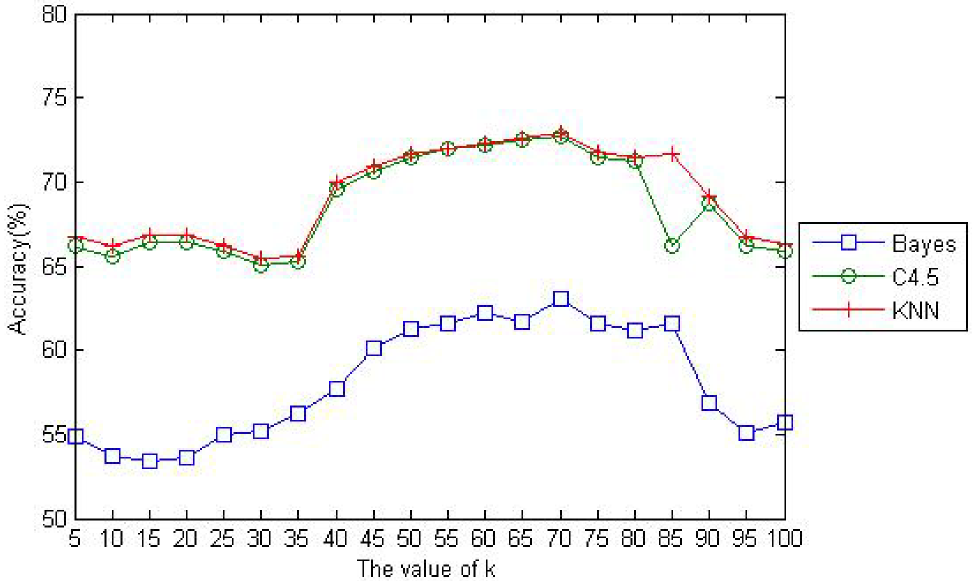

Figure 1 : Influence on classification accuracy of GBWKNN algorithm with different values of K

It is clear that GBWKNN algorithm mentioned above in this paper is convergent under different missing attribute types and the value of k needs experiment to be determined, and in the formula (1), we set $\alpha = 0.5$ as a general value. Results of the experiment indicate that the selection of the best *K* in GBWKNN algorithm is important which affect the result a lot. *K* represents the number of nearest cases, and we choose 20 values of *K* to fill the data respectively. The optimal value of *K* can be determined by analyzing the classification accuracy with three classification algorithms. As is shown in Figure 1, when the value of *K* is 70, the accuracy rates reach higher than others, especially the KNN classification algorithm, which reaches 72.9587 %. Finally, the optimal value of K in the GBWKNN is determined. Predict class.

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 the fact that whether the system is confusing two classes. As is shown in TABLE 2, classification methods has been trained to distinguish between traffic violations, theft, fraud, sex crime, gang/drug offenses and violent crime. This resulting confusion matrix could look like the table below: For example, in this confusion matrix, of the 38229 actual thefts, the system predicted that 138 were traffic violations, 272 were fraud, 126 were sex crime, 198 were drug offenses and 1871 were violent crime etc. For example, it has trouble distinguishing between sex crimes and thefts. But it makes the distinction between thefts and other types of criminals pretty well. Perhaps this case occurs because the instances of attribute theft are too many while the others are relatively less.

TABLE 2 : Confusion matrix on criminal data

Predict Class **Actul 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

CONCLUSIONS

Classification of data mining plays an important role for the criminal analysis. In this paper, there are many missing values in actual criminal dataset. So these filling algorithms are used to deal with the dataset. After filling missing values of the original data set, the complete dataset can be easily chalked up. Finally, three classification algorithms are selected to classify the criminal dataset, comparing the accuracy of three classification methods. Experimental results show that using GBWKNN method to fill the dataset has the higher classification accuracy.

In fact, these filling algorithms have deficiencies. Only filling the discrete missing attributes rather than continuous missing attributes, the filling algorithms can be further improved. Meanwhile, the classification algorithms also can be ameliorated in view of the large number of missing values to improve the classification accuracy.

REFERENCES

[1] Qinchuan Xie; The research and application of data mining technology in economic crime investigation, Netinfo Security, **12**, 36-38 **(2012)**.

[2] Shyam Varan Nath; Crime pattern detection using data mining, In proceeding of Web Intelligence and Intelligent Agent Technology Workshops, 41-44 **(2006)**.

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

[4] Jianshe Huang, Qifu Yao; The application of data mining technique on crime analysis, Journal of Zhejiang Business Technology Institute, **4(3)**, 45-47 **(2005)**.

[5] Chung-Hsien Yu, Max W.Ward et al.; Crime forecasting using data mining techniques, In proceeding of 2011 IEEE 11th nternational Conference on Data Mining Workshops (ICDMW), 779-786 **(2011)**.

[6] N.Tollenaar, P.G.M.van der Heijden; 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 **(2013)**.

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

[8] Lingli Li; A review on classification algorithms in data mining, Journal of Chongqing Normal University: Natural Science Edition, **28(4)**, 44-47 **(2011)**.

[9] Xingyi Liu, Guocai Nong; Comparing several popular missing data imputation methods, Journal of Nanning Teachers College, **24(3)**, 148-150 **(2007)**.

[10] Jinsheng Huo, D.Cox Chris, L.Seaver William et al.; Application of two-directional time series models to replace missing data, Journal of Environmental Engineering-asce-J ENVIRON ENG-ASCE, **136(4)**, 435-443 **(2010)**.

[11] Karahalios Amelia, Baglietto Laura, B.Carlin John et al.; 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 **(2012)**.

[12] Xingyi Liu; A hybrid method of missing value filling, Information Technology (Academic), **27**, 418-420 **(2007)**.

[13] Guoming Sang, Kai Shi, Zhi Liu, Lijun Gao; Missing data imputation based on grey system theory, International Journal of Hybrid Information Technology, **7(2)**, 347-355 **(2014)**.

[14] Geng Zhu; C + + implementation of genetic algorithms and selection of roulette, Journal of Dongguan University of Technology, **14(5)**, 70-74 **(2007)**.

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

[16] Yong Zhou, Youwen Li, Shixiong Xia; An improved KNN text classification algorithm based on clustering, Journal of Computers, **4(3)**, 230-237 **(2009)**.

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

[18] Xiaoliang, Zhu, Hongcan Yan, Jian Wang, Shangzhuo wu; Research and application of the improved algorithm C4.5 on decision tree, In proceeding of International Conference on Test and Measurement (ICTM), 184-187 **(2009)**.

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

[20] Yong Zhou, Youwen Li, Shixiong Xia; An improved KNN text classification algorithm based on clustering, Journal of Computers, **4(3)**, 230-237 **(2009)**.