Data Mining in the Application of Criminal Cases Based on Decision Tree

Ruijuan Hu[1]

Abstract

A briefing on data mining technology used in criminal investigation work and the importance of using ID3 decision tree to structure the Decision Tree algorithm method is given by this paper. It makes a combination of criminal cases criminal suspects training data sets, using its decision tree analysis of the classification and uses Microsoft SQL Server 2005 Office 2007 Add to the data mining of Visio 2007 graphics. It shows and shares the form of mining model.

Keywords: data mining; decision tree; ID3 algorithm; Visio 2007

1. Introduction

With the rapid development of politics, economy and scientific technology, modern crime shows its high-speed, intelligent and high-tech features. Judging from the overall trend of crime, we can observe that the crime case number increases significantly. Since 1949, there have been five peaks of crime. Particularly in the fifth peak, which emerged after the reform and opening up policy coming out. What's more, every crime wave brought new kinds of sins(Li Xian, 2009, p.592-595). The rate of economic crimes, financial crimes and intelligent high-tech crimes rose dramatically – the high rising speed surpassed a vicious, primitive, human instinct of crime, while the proportion of major cases also has increased. Under such circumstances, it is an urgent problem for the police department to figure out how to employ the data mining technology in criminal investigation work, to discover new rules to improve efficiency and rapid responding capability of law enforcement, and to prevent and strike criminal actions in time(Luo Yi,2009.pp.133-135). Data mining will serve as a tool for the leaders and police officers of basic unit to find the hidden information.

2. Data Mining

Data Mining (DM) is the process of extracting useful information and knowledge from large quantities of structured and unstructured data, which is also an effective means of discovering knowledge(Han J,2002,Chapter 2). Data mining sometimes is also known as knowledge discovering or knowledge extracting. Data mining sometimes can be defined as the process of extracting implicit, previously unknown and useful information and knowledge from large quantities of incomplete, noisy, ambiguous, random data for practical application. This definition includes the following layers of meanings: the data source must be real in large quantity; the information that is discovered must arouse the users' interests; the knowledge must be acceptable, understandable and applicable; the knowledge that is discovered is required only to cover specific business problem.

In the process of data mining, the most important step is the data preprocessing, because there exist such bad data as noisy data, void data and inconsistent data due to the diversity and heterogeneity of historical data.

[1] Dept. of Language Engineering, PLA University of Foreign Languages, 2 Guang Wen Road, Luoyang 471003, China

The bad data must be cleaned up because they will affect the accuracy of data mining. And the secondary process is the correlation analysis since many properties of data may concern themselves with the data classification and mission prediction. For example, the height of the suspect may indicate the irrelevance with certain cases. Therefore, the relevance analysis can be conducted to delete irrelevant or redundant properties during the studying process. And at last, the process of data exchanging, which may generalize the data onto a higher level of concept. And the concept hierarchy can be used for this purpose. This step is very important for the property with continuous values. For example, the numeral value of the attribute of age may be generalized to the discrete intervals such as youth, middle-aged and old-aged. Similarly, the nominal values, such as domicile, can be generalized to high-level concepts, such as city.

2.1. ID3 decision tree algorithm

The basic algorithm of decision tree is greedy algorithm which makes use of recursion from the top down to construct the decision tree. In decision tree learning, ID3 (Iterative Dichotomiser 3) is a well-known algorithm used to generate a decision tree invented by Ross Quinlan. The ID3 algorithm can be summarized as follows(Kantardzic M,2002):

Algorithm: Generate_decision_tree. Generate a decision tree by given training data.
Input: samples(training samples) which are indicated by discrete attributes. attribute_list which is the set of attributes.

Output: a decision tree.

Method:

1) Create a root node for the tree.
2) If all samples are in the same class C ,then
3) Return N as a leaf node and mark by class C.
4) If attribute_list is null then
5) Return N as a leaf node and mark by the commonest class of samples.// majority vote
6) Set GainMax=max(Gain1,Gain2…Gainn), if(GainMax)<Criticality threshold (e.g.0.001).
7) Return N as a leaf node and mark by the commonest class of samples.
8) Select test_ attribute having the most information gain of the attribute_list.
9) Make node N as test_attribute.
10) For each known value ai of test_attribute.//to divide samples
11) Grow a branch by node N which meet test_attribute=ai .
12) Set si a sample set of test_attribute=ai.
13) If si is null then
14) Add a leaf, mark by the commonest class of samples.
15) Else add a return node by Generate_decision_tree(si,attribute_list,test_attribute). End ror
16) Return N.

Make use of the attribute of the most information gain as the test_attribute of current node in each node of the tree. The expression of information gain which is a measure of information in attributes is information entropy. The higher the information gain is, the larger the differences among the attributes are, and the stronger the capability of distinguishment is.

The method of information gain is as follows (XianMin Wei, 2011, p.78-80).

Set S be a set of s data samples. Suppose the attribute having m different values and define m different classes $C_i(i=1,2,\ldots,m)$.S_i is the number of samples in C_i. The needing expected information of a giving sample can be given by formula (1) (J. Han, 1992, pp. 547-559).

$$I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{m} P_{ij} * \log(P_{ij}) \quad (1)$$

In formula (1), P_i is probability that any sample belong to C_i. $P_i=|S_i|/|S|$, and 2 is the bottom of the logarithm because we use binary system to code the information.

The entropy of the subsets divided by A is as follows (ZOU Yuan,2010).

Set attribute A having V different values (A_i,a2,…av). Divide S into V subsets (S_1,S2,…Sv) among which S_j contains the samples of S while their values are aj of A. If A is test attribute, the subsets correspond to the branch growing by nodes of set S. Suppose Sij be the number of samples that are the subset sj in class Ci. The entropy or the expected information of the subsets divided by A is:

$$E(A) = \sum((S_{1j} + \cdots + S_{mj}) / S) * I(S_{1j}, \ldots, S_{mj}) \quad (2)$$

In formula (2),(S_{ij}+…+Smj)/S serves as the authority of the jth subset, and equals to the number of the samples in subsets dividing the total of samples in S. The smaller the entropy is, the higher the pureness of divided subsets is.

$$I(S_{ij}, \ldots, S_{mj}) = -\sum_{i=1}^{m} P_{ij} * \log(P_{ij}) \quad (3)$$

In formula (3), Pij = Sij/|Sj| is probability that samples in Sj belong to Ci.
Information gain:

$$Gain(A) = I(S1,S2,\ldots,Sm) - E(A) \quad (4)$$

2.2. ID3 algorithm in the work of criminal cases

Here is an example of decision tree classification algorithm. A certain area criminal cases training data sets listed as table 1.

From table 1, the attribute of the sample sets is “criminal level” which has two values as light and serious. Set C1 corresponding to light and C2 corresponding to serious. There are 9 samples in C1 while 6 samples in C2. To compute the information gain of each attribute, we should compute the needing expected information of the given sample classes.

I(S1,S2) = I(9,6) = -(9/15)*log(9/15)-(6/15)*log(6/15) = 0.971

Next, compute information entropy of each attribute. Let’s begin at “family background” and observe the distributions of light and serious. Then compute the expected information in connection with each distribution.

Family background =”middle”, S11=6, S21=0, I(S11,S21)=0
Family background = “poor”, S12=3, S22=6, I(S12,S22)=0.918

If samples are divided according to family background, the needing expected information of the given sample classes is:

E(family background) = (0/15)*I(S11,S21) + (9/15)*I(S12,S22) = 0.551
The information gain:
Gain(family background) =I(S1,S2) - E(family background) =0.420
The similar compution:
Gain(criminal experiences or not)=0.116
Gain(specialty or not)=0.146
Gain(foreigner or not)=0.249

In this way, Gain(“family background”) has the largest value. It shows this attribution perform an important function in composing data to subclasses. We can create the first node “family background” and divide it to two parts, and so forth, we can get figure 1 as follows.

From figure1,we can see one rule created in each route from root to leaf. When the family background is good, the criminal level of the suspect is light; when the family background is poor and he is foreigner having no specialty, the criminal level of the suspect is light; when the family background is poor and he has specialty, the criminal level of the suspect is serious; when one has specialty and he is native, the criminal level of the suspect is serious too.

3. Demonstration of the mining Results

Office 2007 Visio is data mining template which has the function of prediction and analysis. Using Microsoft SQL Server 2005 Office 2007 data mining external connection program and the draw form of Office Visio 2007, the results of DM based on decision tree can be appeared and shared as Figure 2 below (Zhu Deli. 2007, Chapter 11).

4. Conclusions

Data mining technology is a new kind of science. Its effective application in the criminal investigation is the trend of times and the necessity of the public security work. In face of information sea, Data mining application in criminal cases can help to pick out the intelligence that reflects the security situation timely, comprehensively and accurately, and the intelligent staff will be freed from the massive statistic work, so they can devote more energy and time to focusing on information analysis, which will bring much more improvement to the judgment and efficiency.

References

Luo Yi ,&Chen Lian.(2009). Research on crime case using rough set theory and data mining technology. *Microcomputer Information*,133-135. doi: CNKI: SUN: WJSJ.0.2009-21-057

Du Haizhou, &Ma Chong. (2009). Study on Constructing Generalized Decision Tree by Using DNA Coding Genetic Algorithm. *Web Information Systems and Mining, 2009.WISM 2009. International Conference on*. 31, 163-167. doi: 10.1109/WISM.2009.41, http://dx.doi.org/10.1109/WISM.2009.41

XianMin Wei(2011). Research and application of conditional probability decision tree algorithm in data mining. *Circuits,Communications and System (PACCS), 2010 Second Pacific-Asia Conference on.* 2,78-80. doi: 10.1109/PACCS.2010.5626993.

Li Xian,& Chen Ye gang(2009). Computer Crime Data Mining Based on FP-array. *Journal of University of Electronic Science and Technology of China.*(38),592-595

Han J, Kamber M. (2002). *Data mining: Concept and technologies.* (Chapter 2).

AGRAWAL R, IMIELINSKI T,&SWAMI A(1993). Mining association rules between sets of items in large databases. *Proc of the ACM SIGMOD Conf on Management of Data(SIGMOD '93).New York:ACM Press*,207-216

HAN J, PEI J, &YIN Y(2000).Mining frequent patternswithout candidate generation. *Proc of the 2000 ACM SIGMOD Int'l Conf on Managenent of Data(SIGMOD 2000).New York:ACM Press*,1-12.

J. Han, Y. Cai, &N. Cercone. (1992). Knowledge discovery in databases: An attribute oriented approach. *In Proc. of the VLDB Conference*, Vancouver, British Columbia, Canada, 547-559.

Kantardzic M. (2002). Data Mining Concept, Models, Methods and Algorithms. *IEEE Press*.

S. Muggleton, &C. Feng. (1992). E-cient induction of logic programs. In S. Muggleton, editor, Inductive Logic Programming. Academic Press.

LU S F,&LU Z D(2001) Fastmining maximum frequent itemsets. *Journal of Software*.12(2):293-297

ZHOU Yan, LIU Jie, &SUN Ke. (2011). Improved Decision Tree Algorithm Based on Decision Attribute Selection Strategy. *Journal of Shenyang Normal University (Natural Science Edition)*, (01),60-64

ZOU Yuan. (2010).Data Mining Algorithm Based on Decision Tree Application and Research. *Science Technology and Engineering*, (18)

Zhu Deli. (2007). *SQL Server 2005 Data Mining complete solutions and business intelligence*. Beijing: Electronic Industry Press, (Chapter 11).

Table 1: a certain area criminal cases training data sets

No	Family background	Criminal experience	Specialty	Foreigner	Criminal level
R01	middle	yes	yes	yes	light
R02	poor	no	yes	yes	serious
R03	poor	no	yes	yes	serious
R04	middle	yes	no	yes	light
R05	poor	no	no	no	serious
R06	poor	no	no	yes	light
R07	middle	no	yes	no	light
R08	poor	yes	yes	no	serious
R09	middle	yes	no	yes	light
R10	middle	no	no	yes	light
R11	poor	no	yes	no	serious
R12	poor	yes	no	yes	light
R13	poor	yes	no	yes	light
R14	poor	no	no	no	serious
R15	middle	no	no	yes	light

Figure 1: criminal cases based on Decision Tree

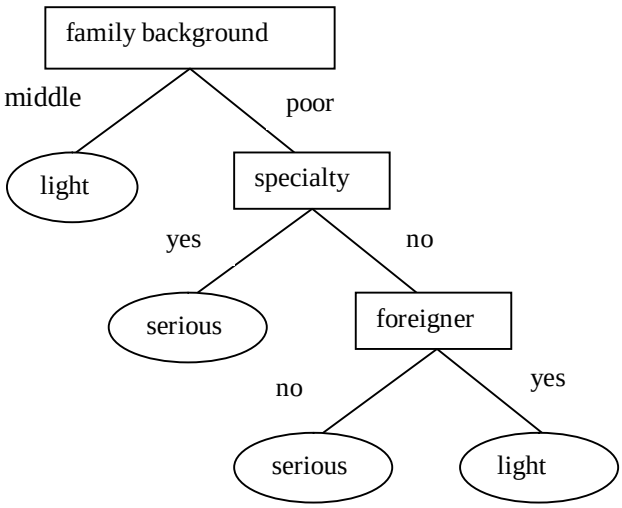

Figure 2: the results of DM based on Decision Tree