Research Journal of Applied Sciences, Engineering and Technology 5(22): 5217- 5222, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 09, 2012 Accepted: December 13, 2012 Published: May 25, 2013

Research on Decision Tree Application in Data of Fire Alarm Receipt and Disposal

[1]Wen Ma and [2]Fajun Dong
[1]Department of Fire Engineering,
[2]Key Laboratory of Ministry of Public Security for Fire-Fighting and Rescue, Chinese People's Armed Police Force Academy, Langfang 065000, China

Abstract: The study adopts decision tree algorithm in SPSS to mine the collected data sample of fire alarm receipt and disposal in a city in 2007, primarily verifies the validity of decision tree method in the analysis of data of fire alarm receipt and disposal and briefly analyzes the mining results. The results show that, there exists categorical relationship in descriptive attributes (fields) of the data of city fire alarm receipt and disposal, which has guiding significance for fire alarm receipt and disposal, meanwhile, makes primary exploration into application of data mining in analyzing data of fire alarm receipt and disposal.

Keywords: CHAID algorithm, data mining, decision tree, fire alarm receipt and disposal

INTRODUCTION

Recently, the concept of "Three Stations Merged into One" has been carried out for receiving alarm in most domestic cities, "Administrative System of the Alarm Receive and Disposal" is widely used among fire fighting forces, huge amount of data information related to receive-disposal alarm and commanding dispatch is accumulated, which lays foundation for data mining application in fire safety domain. However, concerning these data, "sufficient data while deficient information" is the current situation, so how to use data mining technology to find valuable knowledge from mass data as well as mine inherent regularity of data become an important research topic. Hu (2008) study the theory and application of data mining. John *et al*. (2005) have a research of the decision tree technique and its current research. Xue (2006) study the statistical analysis and application of spss. Liu (2004) study the spss for windows statistical analysis.

Decision Tree is a major technique for classification and prediction. Decision tree learning is an inductive learning algorithm on the basis of examples. It infers the classification rules of decision tree's structure according to the disordered and random examples. Decision tree model has many algorithms, such as CART (classification and regression trees), C4.5 and CHAID (chi-squared automatic interaction detector). CHAID algorithm is widely used in market research and social inspection. This study mainly studies the data of fire alarm receipt and disposal by using decision tree analysis and SPSS. Our preliminary step has verified the validity of decision tree in the analysis of data of fire alarm receipt and disposal and analyzed the significance of classified results. Furthermore, it provides scientific support for fire-fighting forces to be prepared against the fire and do fire-fighting and rescue work.

This study adopts decision tree algorithm in SPSS to mine the collected data sample of fire alarm receipt and disposal in a city in 2007, primarily verifies the validity of decision tree method in the analysis of data of fire alarm receipt and disposal and briefly analyzes the mining results. The results show that, there exists categorical relationship in descriptive attributes (fields) of the data of city fire alarm receipt and disposal, which has guiding significance for fire alarm receipt and disposal, meanwhile, makes primary exploration into application of data mining in analyzing data of fire alarm receipt and disposal.

BASIC ALGORITHM

The origination of CHAID algorithm: CHAID algorithm is derived from the two models of AID and thaid. AID only can divide many original variable levels into two categories, but the research always requires more than two categories for the many levels of one variable. So, AID model has the limitation for the further research. The thaid model developed by Messenger and Mandell doesn't resolve the issue that the resolutions are affected by the samples size. CHAID algorithm was proposed by Kass in 1975. The basic analysis concept is χ^2 automatic interaction detector. χ^2 detector resolves the issue that the results are affected by the sample size and can get the significance test on classification.

Corresponding Author: Wen Ma, Department of Fire Engineering, Chinese People's Armed Police Force Academy, Langfang 065000, China

CHAID algorithm is widely used in segmentation research. Firstly, based on the exploration of causation relation, it classifies the level of independent variables for one dependent variable and realizes the classification. Additionally, CHAID algorithm offers the interaction effects between many independent variables. The process is not the simple classification, but presents the description of each category. It can also provide the most important influencing variable, the result of the main effect test of independent variables and interactive influence just like Logit model.

CHAID model algorithm: CHAID algorithm can mine the major influence factors quickly and effectively. It not only can be used for nonlinear and highly relevant data, but also considers missing values and overcomes the limitation of traditional parameter test method in these aspects. The result explanation is also very simple. The core concept is optimal segmentation of the samples on the basis of result variable and explanatory variable and automatic judging and grouping of multivariate contingency table according to $\chi 2$ test results.

Pruning and cross-validation: CHAID algorithm doesn't provide afterwards pruning method, so the tree height and minimum P value should be defined before tree building to avoid excessive growth and fitting. It realizes the purpose of ex-ante pruning. CHAID algorithm achieves the optimal tree through cross-validation to improve the classification precision.
CHAID algorithm is just as follows:

Step 1: For each explanatory variable X, find the two categories with the minimum distribution difference with respect to target variable Y. The method for P value depends on the type of target variable Y. If target variable Y is continuous, use F test; if target variable Y is nominal, build a 2-D cross classification table by the column of X category and the row of Y category and use Pearson $\chi 2$ (Pearson chi-squared) statistic test and Likelihood-ratio test; if target variable Y is ordinal or discrete, use likelihood ratio test.

Step 2: Find the two categories of X with maximum P value and compare P value with pre-set merge level α_{merge}. Firstly, if P value is less than α_{merge}, then merger the two categories of X and form the category of new series of X and then repeat Step One. Secondly, if P value is larger than α_{merge}, continue with Step Three.

Step 3: Use Bonferroni method to calculate the P value adjusted by contingency table by explanatory variable X and target variable Y. See Formula (1) for Bonferroni multiplier calculation method, in which c is the initial category amount of explanatory variable, r is the merged category amount of explanatory variable:

$$B = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{r!(r-i)!} \tag{1}$$

Step 4: Select the explanatory variable X with the minimum adjusted P value and compare the P value with the pre-set split level α_{split}. Firstly, if P value is less than α_{split}, then split the node by the category of X series; secondly, if P value is larger than α_{split}, then don't split the node. The node is the terminal node.

Step 5: Continue with the decision tree growth until it meets the stop rules.

DATA AND CATEGORICAL VARIABLE

The data of fire alarm receipt and disposal comes from the database of 119 or "Three Stations Merged into One" receive-disposal alarm center of every city. It contains abundant information about receiving and disposing fire-fighting alarms. We use this information as the target of decision tree analysis. We must be from the fire-fighting and rescue work of fire fighting forces and select the basic elements which is closely related to fire-fighting and rescue for analysis.

The complete process of fire alarm receipt and disposal include three phases of alarm receipt, fire fighting forces scheduling (organize and dispatch fire fighting vehicles, equipments and staff) and disposal of disaster accidents. Mine the categorical rules by decision tree method and offer scientific evidences for receive-disposal alarm and commanding dispatch.

According to the analysis, primarily confirm seven descriptive attributes (fields) of data of fire alarm receipt and disposal which are closely related to fire-fighting and rescue as categorical variables. They are the time of receiving alarms (by Var1), divided by 24 time phases for each alarm; the duration of disposal (by Var2), means the length of the time that fire fighting forces dispose disaster accidents; the number of turnout vehicles (by Var3); the number of the public security firefighters (by Var4); accident type (by Var5), means the type of the accident; accident target type (by Var6), means the subject type of the accident; arrival time (by Var7) means the time that public fire fighting forces takes from alarm reception to the arrival of the accident site.

EXAMPLES ANALYSIS

Initial settings: To research the categorical rules of the 7 variables, the study collects the data sample of fire alarm receipt and disposal in a city in 2007. 543 data records are achieved through arrangement and standardization and input them to SPSS. Then select and set the model, Table 1.

Effectiveness analysis of variable categorical result: To avoid the categorical function of the missing

Table 1: CHAID model settings

	Settings
Tree	Display = Topdown Nodes = Statistics Branchstatistics = Yes Nodedefs = Yes Scale = Auto Summarytable = No
Gain	Type = [Node] Sort = Descending Cumulative = No Nodes = Terminal
Rules	Syntax = Generic Labels = Yes
Growth limit	Maxdepth = Auto Minparentsize = 100 Minchildsize = 50
Validation	Type = None Output = Bothsamples
Chaid	Asplit = 0.05 Amerge = 0.05 Alphasplit = 0.05 Alphamerge = 0.05 Splitmerged = No Chisquare = Pearson Converge = 0.001 Maxiterations = 100 Adjust = Bonferroni Intervals = 10
Missing	Nominalmissing = Missing

variables during the process of decision tree classification, classify the variables as dependent variables and carry out CHAID algorithm classification (Table 2).

From Table 2 we can see that variables of Var1, Var2 and Var7 have no contribution on the decision tree classification; as dependent variables, other variables all have categorical results. SPSS software can generate the decision tree and categorical rules.

Classification process and results: According to the above analysis, as dependent variables, variables of Var3, Var4, Var5 and Var6 all have categorical results. On the basis of the experience of fire-fighting alarm scheduling, fire-fighting command center determines the scheduling strength (basically the scheduling of fire fighting vehicles) according to the alarm situation. Alarm situation information includes time, address, target type, accident type and scale and so on. So, it is reasonable to select Var3 as dependent variable and others as independent variables. Compare the categorical results further, screen out the categorical results with practical meanings. Decision tree, categorical rules and the results are just as below:

- Var3 is a dependent variable; Var4, Var5 and Var6 are independent variables
  Categorical results in Fig. 1 indicate that Var4 is an effective categorical variable and other variables' categorical functions are not showed out. According to Table 3, Risk statistic estimate is 0.123, which means that Var4's accuracy of categorical prediction to Var3 is 87.7%. Predicted categorical results consist with the real classification. The model fitting results are very good.

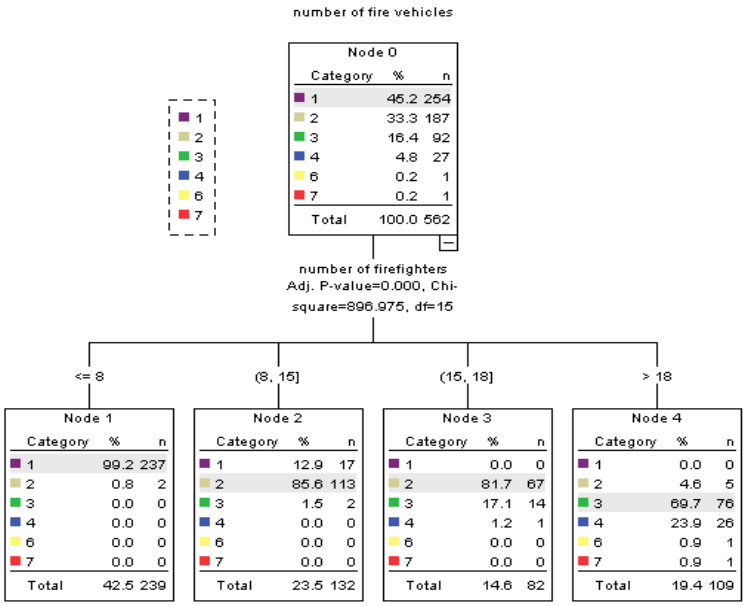

Fig. 1: CHAID decision tree for the number of fire fighting vehicles

Table 2: Classification effectiveness of each variable

Dependent variable	Var1	Var2	Var3	Var4	Var5	Var6	Var7
Var 1	○	×	×	×	×	×	×
Var 2	×	○	×	×	×	×	×
Var 3	×	×	○	√	√	√	×
Var 4	×	×	√	○	√	√	×
Var 5	√	√	√	√	○	√	×
Var 6	×	×	√	√	√	○	×
Var 7	×	×	×	×	×	×	○

Table 3: CHAID misclassification matrix and risk statistics

	Classification						
	Predicted						
Observed	1	2	3	4	6	7	Percent correct
1	237	17	0	0	0	0	93.3%
2	2	180	5	0	0	0	96.3%
3	0	16	76	0	0	0	82.6%
4	0	1	26	0	0	0	0.0%
6	0	0	1	0	0	0	0.0%
7	0	0	1	0	0	0	0.0%
Overall percentage	42.5%	38.1%	19.4%	0.0%	0.0%	0.0%	87.7%
Risk	Estimate＝0. 123			S.E.＝0.014			

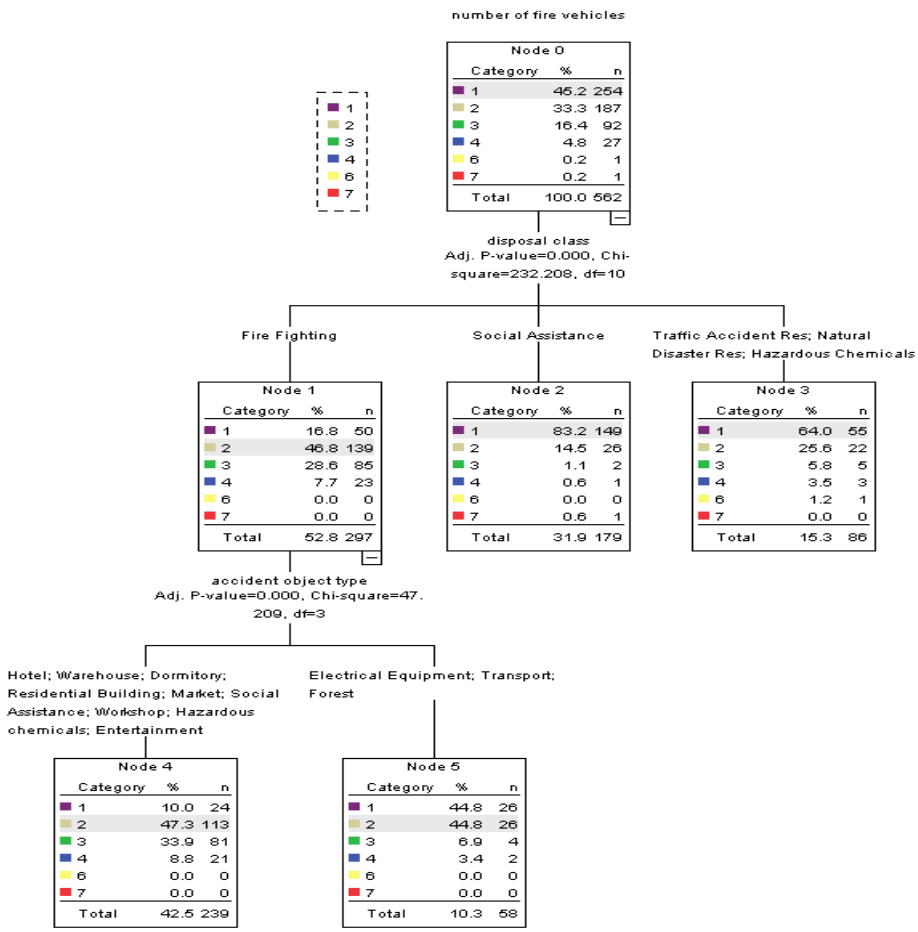

Fig. 2: CHAID decision tree for the number of fire fighting vehicles

The proportion of fire fighting vehicles to the firefighters in public fire fighting forces is about 1：7, so the categorical result of Var3 and Var4 is inevitable. It also proves the effectiveness of CHAID decision tree model, but such result only can be used for the test. It has no real application value.

To get the categorical results with real application value, remove Var4 and carry out CHAID algorithm further.

- Var3 is a dependent variable; Var5 and Var6 are independent variables

According Fig. 2, Var5 and Var6 all have categorical function on Var3. Results in Table 4 shows that Risk statistic estimate is 0. 390, which means that prediction accuracy of this model is 61.0%. Predicted categorical results basically consist with the real classification. The model fitting results are up to the average.

Categorical rules analysis: Table 2 indicates that city accident type has certain categorical influence on fire fighting vehicles scheduling. It has certain guidance meaning for fire alarm receipt and disposal, so decision tree categorical rules in Fig. 2 is analyzed further.

Table 4: CHAID misclassification matrix and risk statistics

	Classification						
	Predicted						
Observed	1	2	3	4	6	7	Percent correct
1	204	50	0	0	0	0	80.3%
2	48	139	0	0	0	0	74.3%
3	7	85	0	0	0	0	0.0%
4	4	23	0	0	0	0	0.0%
6	1	0	0	0	0	0	0.0%
7	1	0	0	0	0	0	0.0%
Overall Percentage	47.2%	52.8%	0.0%	0.0%	0.0%	0.0%	61.0%
Risk	Estimate＝0.390			S.E.＝0.021			

- **Conditions on node 4:** The rule is to dispatch 2 fire fighting vehicles, but there is still 10.0% possibility to dispatch 1 fire fighting vehicle and 33.9% possibility to dispatch 3 fire fighting vehicles and even 8.8% possibility to dispatch 4 fire fighting vehicles. According to the experience, if the condition at node 4 turns out after the alarm, it is proper to dispatch 3 fire fighting vehicles.
- **Conditions on node 5:** The rule is to dispatch 2 fire fighting vehicles, while the possibility of 2 fire fighting vehicles covering Node 5 is 89.6 %, so it is proper to dispatch 2 fire fighting vehicles for Node 5.
- **Conditions on node 2:** The rule is to dispatch 1 fire fighting vehicle, while the possibility of 1 fire fighting vehicle covering Node 2 is 83.2%, so it is proper to dispatch 1 fire fighting vehicles for Node 1.
- **Conditions on node 3:** The rule is to dispatch 1 fire fighting vehicle, but still has 25.6% possibility to dispatch 2 fire fighting vehicles. The possibility of 2 fire fighting vehicles covering Node 3 is 89.6%, so it is proper to dispatch 2 fire fighting vehicles for Node 3.
- **Categorical rules for decision tree improvement in Fig. 2:**

/* Node 4 */

IF (disposal class! = "Social Assistance" AND disposal class! = "Traffic Accident Res" AND disposal class! = "Natural Disaster Res" AND disposal class! = "Hazardous Chemicals") AND (accident object type! = "Electrical Equipment" AND accident object type! = "Transport" AND accident object type! = "Forest")

THEN

Node = 4

Prediction = '3'

Probability = 0.912

/* Node 5 */

IF (disposal class! = "Social Assistance" AND disposal class! = "Traffic Accident Res" AND disposal class! = "Natural Disaster Res" AND disposal class! = "Hazardous Chemicals") AND (accident object type = "Electrical Equipment" OR accident object type = "Transport" OR accident object type = "Forest")

THEN

Node = 5

Prediction = '2'

Probability = 0.896

/* Node 2 */

IF (disposal class = "Social Assistance")

THEN

Node = 2

Prediction = '1'

Probability = 0.832

/* Node 3 */

IF (disposal class = "Traffic Accident Res" OR disposal class = "Natural Disaster Res" OR disposal class = "Hazardous Chemicals")

THEN

Node = 3

Prediction = '2'

Probability = 0.896

According to the above analysis, categorical results of Var3 divided by Var5 and Var6 are obtained by using CHAID algorithm of decision tree, which offer useful decision support for the fire-fighting command center to dispatch fire fighting vehicles. For categorical generation rule, increment (coverage >80%) in the criterion analysis is more suitable for the application of fire alarm receipt and disposal.

CONCLUSION

The research of this study shows that CHAID model of decision tree is effective for analyzing data of fire alarm receipt and disposal and makes certain improvements for the rule generation criteria. Therefore, the research promotes the application of data of fire alarm receipt and disposal actively.

Mine the data of fire alarm receipt and disposal further, not only limited within the decision tree method application, but also cooperated with other data mining skills (such as cluster analysis, time series and association rules) and form complete mining solution. In fire alarm receipt and disposal database, the attributes of data is very rich. This study only selects partial closely associated with fire- fighting and rescue for cluster analysis, so full data arrangement and diversified data mining skills are the research in the future, as well as the

necessary way for the application of data mining skills in fire safety domain.

REFERENCES

Hu, K., 2008. Theory and Applic ation of Data Mining. Tsinghua University Press, China, Beijing.

John, D., J. Cai and Z. Cai, 2005. Decision tree technique and its current research. Cont. Eng. China, 1: 15-18.

Liu, Z., 2004. SPSS for Windows Statistical Analysis. Northeastern University Press, China, ShenYang.

Xue, W., 2006. Statistical Analysis and Application of SPSS. Renmin University Press, China, Beijing.