ATLANTIS PRESS
Advances in Engineering Research, volume 118
2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)

The Application of C4.5 Algorithm in the Analysis of Police Training

Chen Zhe[1, a]

[1]School of Beijing University of Technology, Beijing 100000, China;

[a]czczwin@126.com

Keywords: C4.5, Data Mining, Police Training

Abstract. With the deepening of police training, the traditional training results analysis method can not adapt to the needs of scientific training.In this paper,an example of police shooting performance is introduced, and the C4.5 decision tree learning algorithm of data mining is applied to classify and analyze the shooting situation, and get the factors that influence the performance of special police training and some other conclusions.

Introduction

Police training is the central task of the police force, the efforts to improve training performance are a constant goal.The results of traditional training are analyzed by mean, variance and other statistics, only from the surface to obtain the effectiveness of training and can not be obtained from the implicit factors that affect the training performance of the core link.In order to find out the potential correlation between data, we can find out the true reason that affects the training result through the decision tree induction learning method in data mining, and improve the training level.

C4.5 Decision Tree Algorithm

Decision Tree Induction.

Decision tree induction is a learning of decision trees from labeled training tuples.A decision tree is a tree structure similar to a flowchart in which each internal node (non-leaf node) represents a test on a property,each branch represents an output of the test, and each leaf node (or terminal node) stores a class label.The topmost node of the tree is the root. Given a tuple X with unknown class labels and test the tuple's attribute values on the decision tree. Tracing a path from root to leaf node, which holds the class prediction of the tuple.

Attribute Selection Measures.

The attribute selection metric is a splitting criterion that divides the data partition D of the training-tuple labeled into a single class heuristic.Attribute selection measures are also called split criteria because they determine how the tuples on a given node split.The attribute with the best metric score is selected as the splitting attribute for a given tuple.The tree nodes created for partition D are marked with splitting criteria, branches are generated from each output of the criteria, and the tuples are divided accordingly.Three commonly used attribute selection measures——Information gain, gain rate and Gini index .C4.5 algorithm of the attribute selection measure is the gain rate, because the information gain is the basis of the gain rate, the following first to introduce the concept of information gain.

Information Gain.

The symbols used are as follows.Data partition D be the training set of labeled tuples.Assuming that the class label attribute has m different values, m different classes Ci (i = 1, ..., m) are defined.$C_{i,D}$ is the set of C_i tuples in D，|D|and|$C_{i,D}$| are respectively the number of tuples in D and $C_{i,D}$.The ID3 algorithm uses the information gain as the attribute selection metric.The attribute with the highest information gain is selected as the splitting attribute of the node N.This attribute minimizes the amount of information needed to categorize tuples in the result partition and reflects the minimum randomness in these partitions.The expected information needed to classify tuples in D is given by:

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

p_i is the nonzero probability that any tuple in D belongs to class C_i, and is estimated by $|C_{i,D}|/|D|$.The logarithmic function based on 2 is used because the information is encoded in binary. Info(D) is also known as the entropy of D.

It is assumed that the tuples in D are partitioned by an attribute A, where the attribute A has v different values {a 1, a 2, ..., a v} depending on the observation of the training data.D can be partitioned into v partitions or subsets {D 1, D 2, ..., D v} by the attribute A. (After this division) In order to obtain accurate classification, we also need the information measured by the following formula:

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \text{Info}(D_j) \tag{2}$$

$\frac{|Dj|}{|D|}$ acts as the weight of the jth partition.$\text{Info}_A(D)$ is based on the desired information needed to classify tuples of D by A. The smaller the desired information, the higher the purity of the partition.The information gain is defined as the difference between the original information requirement and the perceived information requirement (after a division)

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \tag{3}$$

Gain (A) tells us how much we got by dividing by A.The attribute A with the highest information gain is selected as the splitting attribute of the node N.

C4.5 Algorithm Description.

The information gain metric is biased towards tests with many outputs.It tends to select attributes with a large number of values.For example, a property that acts as a unique identifier, such as product_ID.The division in the product_ID will result in a large number of partitions (as many as the values), each containing only one tuple.Since each partition is pure, the information needed to classify the dataset D based on the partition is Info product_ID (D) = 0.Therefore, the information gain obtained by dividing the attribute is the largest. Obviously, this division is not useful for classification.

C4.5 uses an information gain expansion called the gain ratio to try to overcome this bias.It normalizes the information gain with split information .Split information is defined as follows:

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2\left(\frac{|Dj|}{|D|}\right) \tag{4}$$

This formula represents the information generated by the v partition of the training data set D divided into v outputs corresponding to the attribute A test.The gain rate is defined as:

$$\text{GrainRate}(A) = \frac{Grain(A)}{SplitInfoA(D)} \tag{5}$$

select the attribute with the highest gain rate as the splitting attribute.

Police Training Example

Police combat effectiveness assessment is mainly carried out through the usual training .In training, the leader usually evaluates the unit's training level by calculating the pass rate, mean, and variance of the training results.But in the multi-personnel, multi-subjects, multi-level case, the traditional means of performance analysis is clearly unable to achieve positive role in the training needs，C4.5 decision tree algorithm in data mining can quickly establish decision tree induction learning method, and from the large number of data samples to summarize the potential rules of impact training, so as to find the effective way to improve the training level scientifically, Pointing out the key points of the training session, effectively feedback and shooting training.

Determine The Business Object.

Shooting training in special police training is a basic training courses, the evaluation of shooting results are hoped to get a comprehensive shooting performance and the relationship between individual training programs and to improve the bottleneck, which targeted to start special training jobs.

Shooting training is divided into precision shooting, short pause shooting, perforated target shooting and moving target shooting.Select a Xinjiang military police unit daily shooting training as a data sample library, build data model: including police officers name and the results.The sample data are shown in Table 1

Table 1 Shooting Performance Sample Data

Police Name	Precision shooting	Short Pause Shooting	Perforated target shooting	Moving target shooting	Comprehensive Result
Li Bing	3	Off target	Off target	Hit	Unqualified
Wang Hao	6	Hit	Off target	Off target	qualified
Sun Hao Yang	2	Off target	Hit	Hit	Unqualified
Chen Kai	4	Hit	Off target	Hit	qualified
…	…	…	…	…	…
Yang qi	4	Off target	Off target	Hit	qualified
He Jun	7	Hit	Hit	Hit	qualified
Xie Hui	3	Hit	Hit	Off target	qualified
Feng Yuan Zhen	2	Hit	Off target	Hit	Unqualified

Data preprocessing.

First of all, the above-mentioned shooting results for data normalization processing, so that the next step in data mining.

The name of the police is replaced with an ID of 1-85.

Shot comprehensive results to 1 on behalf of qualified; 0 on behalf of failure.

Precision shooting to 1 on behalf of the target hit 4 or more, defined as qualified; 0 on behalf of unqualified

The remaining fields are marked with a 1 hit the target, 0 off target

The standard data are shown in Table 2

Table 2 Shooting Performance Specification Data

ID	Precision Shooting	Short Pause Shooting	Perforated target shooting	Moving target shooting	Comprehensive Result
1	0	0	0	1	0
2	1	1	0	0	1
3	0	0	1	1	0
4	1	1	0	1	1
…					
82	1	0	0	1	1
83	1	1	1	1	1
84	0	1	1	0	1
85	0	1	0	1	0

Use the information gain rate to select the attributes that best differentiate the training instances.

This stage uses C4.5 algorithm to establish the corresponding decision tree, the integrated results of 60 qualified students, the overall results of unqualified students 25. For the sake of convenience, The formulas are abbreviated:

Precision Shooting:PS

Short Pause Shooting:SPS

Perforated target shooting:PTS

Moving target shooting:MTS

1)Calculation of information entropy

information entropy：$\text{Info(D)}= -\sum_{i=1}^{m} p_i \log_2(p_i)= -\frac{60}{85}\log_2\frac{60}{85}-\frac{25}{85}\log_2\frac{25}{85}=0.874$

2)Conditional entropy calculation

In the accuracy of shooting, qualified personnel have 75 (comprehensive score of 70 qualified, unqualified 5), 10 unqualified personnel (comprehensive score qualified 8, unqualified

$$\text{Info(PS)}=\frac{75}{85}\times(-\frac{70}{75}\log_2\frac{70}{75}-\frac{5}{75}\log_2\frac{5}{75})+\frac{10}{85}\times(-\frac{8}{10}\log_2\frac{8}{10}-\frac{2}{10}\log_2\frac{2}{10})=0.397$$

$$\text{SplitInfo(PS)}= -\frac{78}{85}\times\log_2\frac{78}{85}-\frac{7}{85}\times\log_2\frac{7}{85}=0.409$$

In the short pause shooting, hit personnel 67 (50 comprehensive qualified, unqualified 11), off target personnel have 18 (11 comprehensive results qualified, unqualified 7)

$$\text{Info(SPS)}= \frac{67}{85}\times(-\frac{50}{67}\log_2\frac{50}{67}-\frac{17}{67}\log_2\frac{17}{67})+\frac{18}{85}\times(-\frac{11}{18}\log_2\frac{11}{18}-\frac{7}{18}\log_2\frac{7}{18})=0.848$$

$$\text{SplitInfo(SPS)}= -\frac{67}{85}\times\log_2\frac{67}{85}-\frac{18}{85}\times\log_2\frac{18}{85}=0.745$$

In the perforating shooting, hit personnel 72 (comprehensive score of 46 qualified, unqualified 26), offtarget personnel 13 (comprehensive score of 6 qualified, unqualified 7)

$$\text{Info(PTS)}= \frac{72}{85}\times(-\frac{46}{72}\log_2\frac{46}{72}-\frac{26}{72}\log_2\frac{26}{72})+\frac{13}{85}\times(-\frac{6}{13}\log_2\frac{6}{13}-\frac{7}{13}\log_2\frac{7}{13})=0.851$$

$$\text{SplitInfo(PTS)}=- \frac{72}{85}\times\log_2\frac{72}{85}-\frac{13}{85}\times\log_2\frac{13}{85}=0.617$$

In moving target shooting, 71 hit personnel (56 were qualified, 15 were unqualified), 14 offtarget (2 were qualified and 12 were unqualified)

$$\text{Info(MTS)}= \frac{71}{85}\times(-\frac{56}{71}\log_2\frac{56}{71}-\frac{15}{71}\log_2\frac{15}{71})+\frac{14}{85}\times(-\frac{2}{14}\log_2\frac{2}{14}-\frac{12}{14}\log_2\frac{12}{14})=0.718$$

$$\text{SplitInfo(MTS)}=- \frac{71}{85}\times\log_2\frac{71}{85}-\frac{14}{85}\times\log_2\frac{14}{85}=0.646$$

3)Calculation of Information Gain Rate

$$\text{GainRatio(PS)}=\frac{Info(D)-Info(PS)}{\text{Split}Info(PS)}=\frac{0.874-0.397}{0.409}=1.166$$

$$\text{GainRatio(SPS)}=\frac{Info(D)-Info(SPS)}{\text{Split}Info(SPS)}=\frac{0.874-0.848}{0.745}=0.034$$

$$\text{GainRatio(PTS)}=\frac{Info(D)-Info(PTS)}{\text{Split}Info(PTS)}=\frac{0.874-0.851}{0.617}=0.037$$

$$\text{GainRatio(MTS)}=\frac{Info(D)-Info(MTS)}{\text{Split}Info(MTS)}=\frac{0.874-0.718}{0.646}=0.241$$

C4.5 algorithm get the maximum information rate of the "shooting accuracy" as the root node, The two values of " precision shooting " in 85 sample data,hit and offtarget, are branched and extended recursively according to the above algorithm. The final decision tree is shown as the figure below. Rectangular box represents the split of the property, oval box on behalf of the overall results of qualified or not.

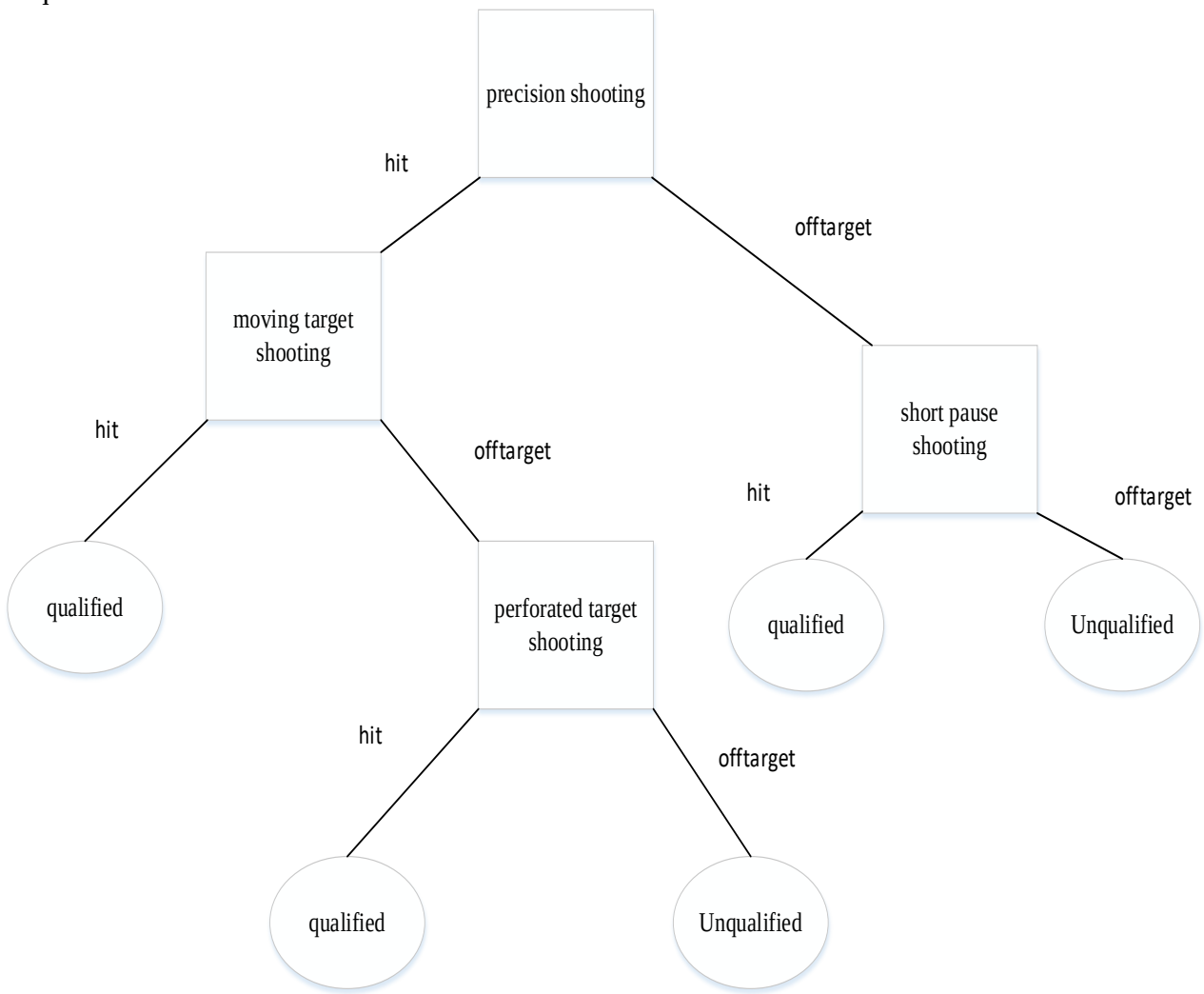

Fig. 1 C4.5 Decision Tree Generated by Shooting Training

Result Analysis.

After the calculation of the C4.5 decision tree, observation and analysis can draw the following conclusions:

1)As the basic training of the shooting training, precision shooting is the basis of precision shooting, also known as the static shooting, precision shooting poorly commanded officers of its shooting training is not very good comprehensive performance, so to enhance the accuracy of shooting is very important for a police officer.

2)In the case of good precision shooting, sport shooting directly affects the comprehensive performance, so dynamic and static combination can effectively improve the shooting training results.

3)For the static and sports shooting are not good police officers often poor comprehensive training results, perforated shooting and pause shooting effects on the random comparison, the police leaders can set up other individual shooting training content, worked out a set to improve overall performance The best solution.

Conclusion

In this paper, a police shooting examples show that C4.5 data mining algorithm for training performance analysis of the impact of great significance. Through the C4.5 algorithm, the data mining of the daily training data can be applied in various training subjects, and a large number of rules which are difficult to detect but have practical significance can be obtained, which can effectively guide the special police training and promote the whole training level.

References

[1] Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques[J]. Biomedical Engineering Online, 2011, 5:51(1):95–97.

[2] Han J, Kamber M. Data Mining Concept and Techniques[M]. 2006.

[3] Tan P N, Steinback M, Kumar V. Introduction to Data Mining[J]. Intelligent Systems Reference Library, 2006, 22(6):753-754.

[4] Quinlan J R. C4.5: programs for machine learning[J]. 1993, 1.

[5] Li Jian Ping.The Analysis and Research of Decision Tree Technology in Military Training Achievements [J]; Journal of Kunming University of Science and Technology;

[6] Hssina B, Merbouha A, Ezzikouri H, et al. A comparative study of decision tree ID3 and C4.5[J]. International Journal of Advanced Computer Science & Applications, 2014(2).

[7] Bashir S, Qamar U, Khan F H, et al. An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5, & CART Ensembles[C]// International Conference on Frontiers of Information Technology. 2014:226-231.