MATEC Web of Conferences **25**, 03013 (2015)
DOI: 10.1051/matecconf/20152503013

Applied Research of Decision Tree Method on Football Training

Jinhui Liu
College of Physical Education, Langfang Teachers University, Langfang, Hebei, China

ABSTRACT: This paper will make an analysis of decision tree at first, and then offer a further analysis of CLS based on it. As CLS contains the most substantial and most primitive decision-making idea, it can provide the basis of decision tree establishment. Due to certain limitation in details, the ID3 decision tree algorithm is introduced to offer more details. It applies information gain as attribute selection metrics to provide reference for seeking the optimal segmentation point. At last, the ID3 algorithm is applied in football training. Verification is made on this algorithm and it has been proved effectively and reasonably.

Keywords: decision tree; football training; ID3 algorithm

1 INTRODUCTION

With the arrival of big data age, and in order to provide more decision proofs and method support for decision makers, people are paying more attention to data and making effort to extract more important knowledge and information.

At present, there are many information collection systems existing in football training. These systems can make systematic records of weather condition and athlete's status in football training. The records can help accomplish football training in a smooth way and accumulate abundant data for reaching the best training result. In the meantime, it can be seen that an effective tool is required in the transformation between mass data and information. It means that looking for an effective data mining technology is very important in providing certain decision support for football decision makers.

Within the wide research range of data mining technology algorithm, decision tree is a common method. Due to its briefness and integrality, more scholars and decision makers are studying it. The conception of decision tree method was proposed by Hunt (1966). Early-stage algorithm was called CLS. Following decision tree learning methods are the supplement and development of CLS. Quinlan (1979) proposed ID3 algorithm that makes up the defect of CLS. ID3 algorithm takes information gain as the standard in text attribute selection, and it can solve the problem of large-scale data operation. It is the optimization and innovation of CLS to some extent. Currently, C4.5 system is widely used due to 2 advantages: First, it can solve the problem of continuous deference; second, its form of rule presentation can realize equivalent transformation of decision tree. The decision tree method as an inductive learning method based on living examples contains certain advantages in data mining and classification.

Based on this, this paper will make analysis of decision tree method in combination with football training data so as to provide related decision proof and method support for decision-makers.

2 ANALYSIS MODEL OF DECISION TREE

Among all classification algorithms, the selection of a predictive analytical model with good interpretability to directly acquire knowledge without ambiguity has more practical value. The decision tree is a simple and comprehensive system mode. Currently, it has become the most commonly-used data mining technology. The study of the decision theory and related algorithms can play a key role in applying them into other areas.

2.1 *Decision tree theory*

The decision tree can be used to conclude classification rules of certain expression forms from irregular examples. In general, they are mainly conclusions based on reality. Decision tree makes comparison of attribute value from top to bottom and from the inside and out so as to judge different attributes. It takes further classification from nodes until reaching the final conclusion at the end. Users only need to express their conclusion and there is no need for them to get acquainted with background knowledge which is the uppermost advantage of decision tree.

Internal node of decision tree can be an attribute or a corresponding collection, referring to the texted attribute. Leaf nodes are the types for classification. A decision tree will occur after certain training which can classify unknown values.

2.2 *Decision tree algorithm*

The most commonly-used algorithm among classification technologies is the decision tree inductive method. The main calculation process is shown in Figure 1:

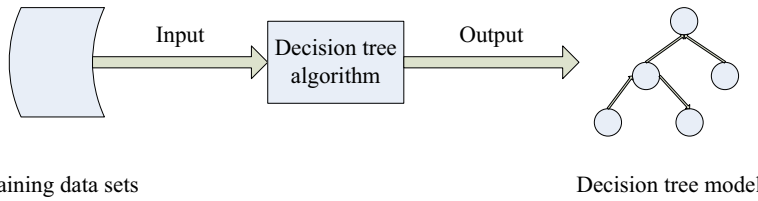

Figure 1. Work flow of decision tree

The most significant distinction in decision tree classification lies in the different decision algorithms of decision tree generation process. The following part will give a brief introduction to the CLS algorithm and the ID3 algorithm.

2.2.1 *CLS algorithm*

The CLS algorithm starts from an empty decision tree. During its process, the CLS algorithm will replace the original decision tree by continuously adding and judging nodes. At last, the algorithm will end with the right training examples of decision tree. The main steps to complete the process are as follows:

1. If Q refers to the collection of test attribute and X refers to the entire training data set, (X, Q) can refer to one tree root of the decision tree T in its original state.

2. If leaf node (X') of decision tree T has two statuses—component Q' is empty or X' refers to the same category of training data, the entire algorithm will be stopped. The corresponding decision tree is T.

3. Otherwise, select a node (X', Q') corresponding to other steps;

4. For component Q', select a text attribute b according to certain rules. Take X' as the various values obtained by b. X'_i refers to the m different intersecting subsets among which $1 \le i \le m$. If m branches appear from $(X', Q' - \{b\})$, each branch can refer to the various values of b. Hence m new leaf nodes (X', Q') can be formed in this way with range of $1 \le i \le m$.

5. Return to step 2 to continue iteration until related standards can be met.

2.2.2 *CLS algorithm analysis*

In step 2, if all data sharing the same attribute exist in same categories, then one of the above conditions can be satisfied and other can also be met. However, this conclusion is unsuitable for condition when any pair of data sharing the same attribute exists in different categories.

In step 4, the process only has certain significance when $m > 1$. When there's any problem in training data, it cannot be guaranteed that $m > 1$. Moreover, CLS algorithm does not specify the selection standards for test in details.

It is actually a process of specializing hypothesis to obtain the establishment of decision tree according to CLS algorithm. One feature of the algorithm is that it is a learning algorithm with only one operator. Therefore, the functioning process of CLS algorithm in invoking this operator is actually the process of establishing decision tree.

2.3 *ID3/C4.5 Algorithm*

The CLS algorithm has clear thinking mechanism. Many researchers are studying this algorithm. However, CLS algorithm contains certain defects as well. One of the major problems is that training data are the foundation of algorithm running and the algorithm may slow down due to the massive training data stored in internal storage. System failure may also occur and thus impact the process of problem solving. In the meantime, certain problems may also exist in the resolution results. As a result, a new algorithm, ID3 algorithm, is generated to eliminate the defects of CLS. ID3/C4.5 algorithm can solve the bottleneck problem of CLS algorithm to some extent. It was proposed by Quinlan (1979). Compared with other algorithms, ID3/C4.5 algorithm has two unique advantages: 1) The test attribute standard of this algorithm is the descent rate of information entropy (the descent rate of information entropy refers to the descent rate of probabilistic information); 2) this algorithm introduces incremental learning technology. Incremental learning technology can effectively solve the problem of algorithm running slow-down resulting from data redundancy.

2.3.1 *Algorithm steps*

1. An important feature for why ID3 algorithm is superior to other algorithms is that its incremental learning method can effectively solve the problem of data running. The key solution is to introduce windows. First, select a random window X_i for training data X with window scale of W. X_i , which is to be called a window, refers to a random subset.

2. The decision tree is established in this step. The specific establishment method is consistent with CLS. However, the main difference between these two algorithms which are also important features of incremental learning method is that their measure standards are different. Standard of ID3 algorithm is the incremental attribute of information. This attribute can hold more data capacity in running process. As a result, the selection made by the current window in attribute is to select the maximum information incremental attribute, which is not random. Meanwhile, the running of decision tree can only be stopped when two important

requirements can be satisfied; or it will continue until the requirements can be met. At first, consistence of sample sources shall be guaranteed. Second, an important requirement for stopping decision tree running is to divide samples until no more division is available. In this situation, regard the leaves as the transformation method of nodes. Mark, replace and store the leaves by means of majority voting.

3. This step mainly judges whether there are some "special individuals"—whether any exception exists. If there're "special individuals" existing in decision tree, scanning of the training data shall be continued. If there is no "special individual" existing in decision tree, scanning of the training data can be stopped.

4. A new window can be formed in this step. The window is composed of two parts: 1) the "special individuals" existing in step 3—exceptions; 2) some parts of training data. These two parts constitute the new window. Skip to step 2 to continue the running of the algorithm.

2.3.2 *Algorithm characteristics and step analysis*

Compared with other algorithms, ID3 algorithm has one significant advantage that it can effectively solve the problem of large-scale data running. The key of solving this problem is to apply incremental learning method and generate new windows. Through the introduction to the above algorithm steps, we can know that new windows are mainly formed in step 4. There're two ways to form new windows. 1) The core idea of this method is to use exceptions to replace training examples. Of course, the selection of training examples is not random. Each leaf node shall be careful. Only one node will be preserved while all the others will be deleted. The running results of this method are good; however, there're also some defects. When multiple conceptions overlap with each other, windows with scale of W are not easy to be found during the running of the algorithm. As a result, convergence is not applicable while the effectiveness of the results may be impacted. 2) Compared with the first method, this one has better running results. It can eliminate the result defects existing in algorithm running with multiple conceptions and complicated influence. The major part of this method is to expand the existing windows. On one hand, it reserves the "special individuals"—exceptions found in step 3; on the other hand, it can also maintain all the window examples. From these two parts, we can see that this method can expand windows at the greatest extent.

Furthermore, compared with CLS algorithm, there're two significant advantages in this algorithm. 1)It can solve the problem that the large-scale data may impact running effect. When the data scale is large, there's certain limitation in CLS algorithm. However, ID3 algorithm can start incremental learning by generating new windows. The core of this algorithm is to combine two ideas, including rough set and windows. With the guidance of these two ideas, the ID3 algorithm can avoid the slow-down or even failure caused by insufficient internal storage resulting from the large-scale data in decision tree algorithm. 2)Compared with CLS algorithm, this algorithm is superior in test attribute selection. CLS algorithm selects test attributes by rules which have certain subjectivity while scientific proof is not a must in selection standard. Nevertheless, in ID3 algorithm, information gain is the main selection standard which has more scientific proofs.

2.3.3 *Selection of attribute metrics*

From the analysis given above, we can know that compared with CLS algorithm, one significant advantage of ID3 algorithm is its selection of test attribute. ID3 algorithm takes information gain as test attribute which is also called the split advantage/ disadvantage measurement. When data scale is large and occupies massive storage, algorithm running may be impacted. Taking information gain as test attribute can classify, divide and arrange samples while the required information content is the least. Hence the storage occupancy can be greatly reduced. Moreover, this method can also show the minimum random classification. The specific method to select test attributes for current window nodes is as follows: search each attribute during algorithm running, and use applicable information gain to select the optimal (or the maximum) attribute. The maximum entropy compression selected by this means contains much information. From the above analysis, we can find that this method only needs the minimum test quantity to greatly reduce workload and accelerate the entire algorithm during running. It can be ensured that the decision tree will be found out at the same time. In addition, this algorithm contains more metrics of information gain which can be taken as the test attributes of nodes. The principle of taking information gain as selection standard makes ID3 algorithm superior to other algorithms to some extent.

Propose the following hypothesis:

1. Suppose there are K sets with k samples.

2. Suppose different label attributes have different values. Suppose there are s values and s inhomogeneity C_i ($i = 1,...,s$) here.

3. Suppose the sample quantity of inhomogeneity C_i is k_i.

4. Suppose attribute B contains different values. There're n values in the form of $\{b_1,...,b_n\}$.

5. Suppose the sample quantity in inhomogeneity C_i of subset is k_{ij}.

Based on the above hypothesis, classify one selected sample to show the required information:

$$I(k_1,k_2,....k_m) = -\sum_{i=1}^{s} p_i \log_2(p_i) \quad (1)$$

Among which $p_i = \frac{k_i}{k}$ and p_i refers to the frequency of $\in C_i$. Moreover, binary system is a coded

form of information. Therefore, the form of log_2 occurs in formula.

K refers to the form of B's classification into n subsets. The detailed form is $\{K_1,...,K_n\}$. The sample quantity in K is also included in K_j. Values in b_j correspond with attribute B. If B is chosen as the optimal test attribute, the corresponding subset will present one-to-one corresponding relation with the branch. The branch is formed by nodes including set K. Entropy of subset will be classified according to certain standards. The classification according to B standard is as follows:

$$E(B)=\sum_{j=1}^{n}\frac{k_{1j}+...+k_{sj}}{k}I(k_{1j,...,}k_{sj}) \tag{2}$$

The weight of the j subset is $\frac{k_{1j}+...+k_{sj}}{k}$, among which the numerator refers to the sample quantity of b_j while the denominator refers to the total sample quantity of K. From the above formula, it can be seen that the detailed planning degree of subset classification is inversely proportional to entropy value. With the increase of entropy value, the detailed planning degree (fineness) will decrease, among which the subset K_j is presented as follows:

$$I(k_{1j},k_{2j},....k_{sj})=-\sum_{i=1}^{s}p_{ij}\log_{2(p_{ij})} \tag{3}$$

Among which $p_{ij}=\frac{k_{ij}}{|K_j|}$

The branch binary coding information of attribute B can be shown in the following formula:

$$Gain(B)=I(k_1,k_2,....k_m)-E(B) \tag{4}$$

The highest information gain $Gain(B)$ can be obtained by the above formula.

Based on this, the method to classify samples can be concluded as follows: 1 node will generate during the running process of the algorithm. Use the attribute of information gain to classify and distinguish it. A branch will appear at the same time. Calculate the expected compression of attribute entropy during the running process of the algorithm. Select the attributes according to certain standards. Here, the expected compression of the optimal entropy is taken as the test attribute of set K.

From the above study, we can know that compared with other algorithms, ID3 algorithm is superior in solving the problem of large-scale data. As a result, this algorithm has high application significance.

However, the system that is widely accepted and applied by most scholars is C4.5. The major advantage of this system lies in its data processing. It has certain advantage in processing continuous attribute data. In addition, this system can transform equivalent rules to decision tree which can significantly improve the running speed of the system.

3 APPLICATION OF DECISION TREE ON FOOTBALL TRAINING

3.1 *Football training data table*

Climate factors can have certain influence on football training. This paper selects four climate factors to study their influence:

1. Wind: calm; medium wind; gale
2. Temperature: moderate; hot; cold
3. Weather: sunny; rainy; cloudy
4. Humidity: high; normal

Beyond that, whether the comprehensive results of the four factors are suitable for football training is presented in "Influence". This paper studies the existing 10 groups of data as shown in Table 1.

3.2 *Construction decision tree*

Establishment process of ID3 algorithm:

1. Conduct systematic calculation of information gain of various attributes;
2. Select the attribute Ak with the most information gain in calculation;
3. Gather the examples with the same Ak value into the same category. Take the following values as corresponding subsets;
4. Process the subsets with positive and negative examples by recursive algorithm;
5. Return if no positive example and negative example exist simultaneously.

S refers to the sample set. $P(\mathrm{ui})$ refers to the occurrence frequency of category i. The gain can be presented as follows:

$$gain(\mathrm{S},\mathrm{A})=En\,tropy(\mathrm{S})-\sum_{u}\frac{|S_u|}{|S|}En\,tropy(\mathrm{S}_u) \tag{5}$$

The information gain of climate factors can be calculated from the above steps. Corresponding results are shown in Table 2.

From the Table 2, we can reach the conclusion that the most information gain can be obtained by dividing training sets according to weather. As a result, weather shall be selected as the root node of decision tree. After making further division by taking sunny, cloudy and rainy weather as the nodes, the set that is entirely suitable for training is sunny weather. The corresponding decision tree is as Figure 2.

From the decision tree graph given above, we can reach the conclusions that temperature shall be considered in rainy days and wind shall be considered when temperature is too high; calm days are suitable for football athlete's training and gale is unsuitable;

Table 1. Football training data sheet

Wind	Temperature	Weather	Humidity	Influence
Gale	Moderate	Sunny	High	Appropriate
Medium wind	Moderate	Cloudy	Normal	Appropriate
Calm	Moderate	Rainy	High	Inappropriate
Medium wind	Cold	Rainy	Normal	Inappropriate
Medium wind	Hot	Sunny	High	Appropriate
Calm	Hot	Sunny	High	Appropriate
Medium wind	Moderate	Sunny	High	Appropriate
Gale	Moderate	Rainy	High	Inappropriate
Medium wind	Moderate	Cloudy	High	Appropriate
Gale	Hot	Cloudy	High	Inappropriate

Table 2. Information gain table

Wind	$g(A1)=0.097$	Temperature	$g(A2)=0.179$
Weather	$g(A3)=0.181$	Humidity	$g(A4)=0.143$

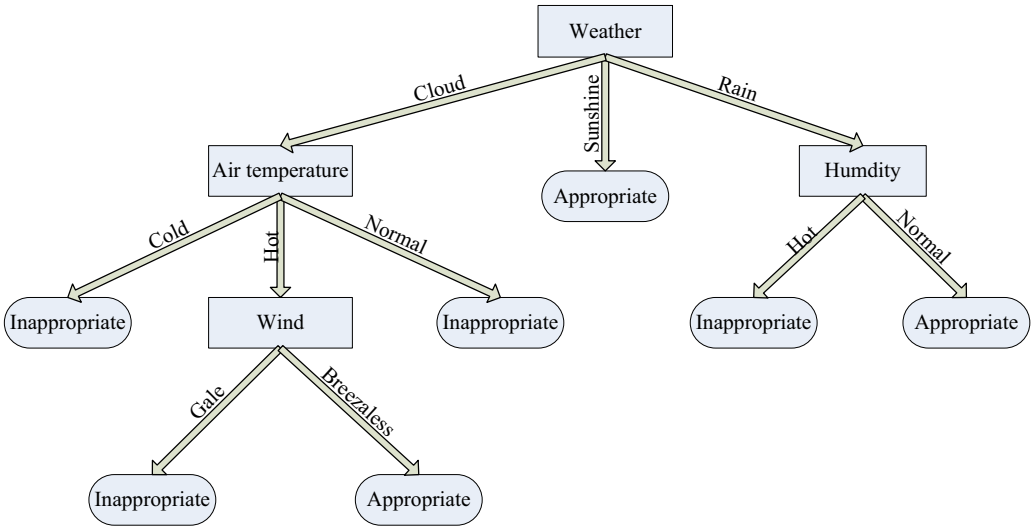

Figure 2. Decision tree forming graph

Table 3. Experimental test result table

Group Number	Wind	Temperature	Humidity	Weather
1st group	Gale	Cold	Normal	Sunny
2nd group	Medium wind	Moderate	High	Rainy
3rd group	Medium wind	Cold	Normal	Cloudy

moderate or low temperature is unsuitable for training; humidity shall be considered in cloudy days; normal humidity is suitable for training while too high humidity is unsuitable; sunny days are suitable for football athlete's training.

3.3 *Decision tree model inspection*

In order to certify the validity of the model, this paper selects three groups of data to analyze as shown in Table 3.

Weather of the first group data is sunny which is ensured to be applicable for football athlete's training according to the decision tree graph. Weather of the second group data is rainy which is ensured to be inapplicable for football athlete's training according to the decision tree graph. Weather of the third group data is cloudy which is ensured to be applicable for football athlete's training according to the decision tree graph. By comparison, it is easy to find that these

conclusions conform to Table 1. The validity of decision tree model can be certified.

4 CONCLUSION

By exploring decision tree algorithm and applying it in football training, main conclusions given below can be reached:

(1) The main advantages of decision tree algorithm include briefness, high practicability, clear and understandable theory, and wide application range;

(2) With the arrival of big data age, the accuracy and precision of decision tree will be improved. With continuous data calculation, the deletion and updating can make results more accurate.

(3) The application of decision tree on football training can provide theoretical proof for football training. Moreover, the decision tree can be also used in volleyball and other areas.

REFERENCES

[1] Zhang, L., Chen, Y. & Li, T.Y. et al. 2011. Study of Decision Tree Classification Algorithm. *Computer Engineering*, (13):63-65.

[2] Wang, X.; Wang, X.Z. & Chen, J.K. 2013. Comparative Study of Ordered Decision Tree. *Journal of Frontiers of Computer Science and Technology*, 7(11):1016-1023.

[3] Zhang, P. & Ma, M.C. 2011. *Advanced Training of Tennis*. Beijing: Jilin Science and Technology Press,

[4] Li Boyang, Wang Qiangwei, & Hu Jinglu. 2013. Multi-SVM classifier system with piecewise interpolation. *IEEJ Trans on Electrical and Electronic Engineering*, 8(2):130-136.

[5] Wu, W.J. & He, F. 2013. Learning Method of Tennis Selection Index Based on Decision Tree. Harbin Normal University: *Journal of Natural Science*, (6):87-91.

[6] Wang, L.M. 2007. Study of Decision Tree Learning and Its Pruning Algorithm. Wuhan University of Technology. 4:14-17.

[7] Wang, H.R. & Li, W. 2011.Decision Tree Algorithm Based on Association Rules. *Computer Engineering*, 37 (9):101-107.