Improvement of Decision Tree ID3 Algorithm

Lin Zhu(✉) and Yang Yang

Beijing University of Posts and Telecommunications, Beijing, China
zlzl.zl@l63.com, yyang@bupt.edu.cn

Abstract. This paper describes the basic concepts of the ID3 algorithm and its principles as well as the construction process. Because ID3 algorithm tends to select values for more attributes shortcomings, we introduce threshold, properties information gain rate and parameters to compensate for the lack of ID3 properties selected standard. Based on the above two points to achieve new property selection standard, the original ID3 algorithm is improved. Through the experiment, the improvements of the improved algorithm were compared. Experiment results show that the improved algorithm is effective.

Keywords: ID3 algorithm · Information gain · Information gain rate · Classification property

1 Introduction

With the rapid development of computer technology and network technology, amount of data information is also invaluable in multiples of growth. The purpose of data mining technology is how to effectively find potentially useful knowledge from these data. Data mining is a field of study and gradually developed at the end of the 1980s. It is from a lot of, incomplete, noisy, fuzzy actual application data. We extracted that people do not know in advance but is potentially useful information and knowledge in them.

2 ID3 Algorithm

ID3 algorithm is a typical decision tree learning algorithm. Its core is the tree nodes at all levels, with the information gain attribute selection method as a standard to help determine the appropriate property to generate each node. So you can choose when having the highest information gain attribute as an attribute of the current node to use subset of samples obtained by the division of the property, the minimum information required to classify.

This work was supported by Open Subject Funds of Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory (ITD-U15002/KX152600011). NSFC (61401033, 61372108, 61272515). National Science and Technology Pillar Program Project (2015BAI11B01).

S. Wang and A. Zhou (Eds.): CollaborateCom 2016, LNICST 201, pp. 595–600, 2017.
DOI: 10.1007/978-3-319-59288-6_59

Suppose the training set size m, the set $S = A_1 * A_2 * \ldots * A_n$ is finite-dimensional vector space, and each vector space is $n_1, n_2, \ldots n_n$ dimensional subspace respectively.

Suppose $s_1, s_2, \ldots s_r$ is a subset of a vector space S. Its size is $m_1, m_2, \ldots m_r$, and $m = m_1 + m_2 + \ldots + m_r$. Then a decision tree to correctly determine Expectation needs:

$$I(S, m) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m} \qquad m = m_1 + m_2 + \ldots + m_r \tag{1}$$

If Property A_k is the root, the information entropy is:

$$E(I(A_k, m)) = \sum_{i=1}^{m} \frac{m_i}{m} I(A_k, m) \tag{2}$$

The information gain:

$$Gain(A_k) = I(S, m) - E(I(A_k, m)) \tag{3}$$

By selecting the maximum information gain attributes extension test attributes diffusion principle ID3 classification, choose each calculated maximum information gain attributes as a new node in the tree, and each attribute value of the property to build branch, according to this thinking divide training data sample set.

While ID3 is a typical decision tree classification algorithm, but there are still shortcomings: Since the ID3 algorithm is to choose the maximum entropy as an attribute selection standard, therefore, it will be more inclined to attribute more value, but not the property more value is optimal properties.

3 The Improved Algorithm

ID3 algorithm selects the maximum entropy as property standards, will be more inclined to more value property, but the property more value is not optimal properties. In this paper, ID3 algorithm for multi-valued attribute bias problems to improve, through the decision tree information gain rate, using the classification tree to determine the parameter values and re-establish the decision tree, thereby improving the value of the multi-algorithm bias problem.

Based on ID3 algorithm, we introduce the concept of information gain rate, using information gain rate instead of information gain attributes as selection standard, the formula of which the rate of information gain:

$$GainRate(A_k) = \frac{Gain(A_k)}{I(S, m)} \tag{4}$$

But the rate of information gain may be the case over compensation is approximately zero, thus introducing bias threshold r. For set S, there are n attributes, attribute bias threshold value of r is typically the average of all the attribute information entropy:

$$r = \frac{1}{\mathrm{n}} \sum\nolimits_{1}^{n} E(I(A_k, m)) \tag{5}$$

Information gain ratio generated a decision tree, which effectively solve the multi-valued attribute bias problem. Meanwhile, the importance of property refers to all the properties of the importance of information gain contribution comparative results. The value can be defined as a subset of the properties A branch share Instances A proportion, by comparison to distinguish between information entropy gain contribution calculated to select the optimal properties. Importance of the role is to distinguish between different information attributes importance or dependence. The value can be defined as a subset of A proportion shares Instances A proportion, by comparison to distinguish between information entropy gain contribution calculated to select the optimal properties.

Thus, for the classification of different properties have different degrees of importance. Introducing parameter indicates the degree of importance, in order to increase the degree of importance of important properties, the improvement of information entropy formula:

$$\mathrm{E}(\mathrm{I}(\mathrm{A_k}, \mathrm{m})) = \sum\nolimits_{\mathrm{i}=1}^{\mathrm{m}} \left(\frac{\mathrm{m_i}}{\mathrm{m}} + \alpha \right) \mathrm{I}(\mathrm{A_k}, \mathrm{m}) \tag{6}$$

The resulting tree root attribute value is the proportion of first generation of a decision tree yes/no, non-root attribute value is 0.

Thus, the resulting decision tree to determine a parameter value, and by generating a decision tree again raised the importance of important attributes. Through a combination of both methods to make up for the shortcomings of traditional ID3 algorithm.

The improved algorithm steps:

(1) Calculate the Expectation information and information entropy of each attribute.
(2) For the set S, there are n attributes. R usually tend to the threshold value is the average of all the attribute information entropy.
(3) Computing the root attribute information gain.
(4) The information entropy of each attribute and the threshold value r are compared. If the entropy value is lower than the threshold r, we select the gain ratio standard; if higher than the threshold r, we select the information gain standard.
(5) Create node recursively until you've selected all properties.
(6) Determine the impact factor and parameter values.
(7) Replace the formula (3) is calculated using the formula of information entropy improved, decision trees again.

4 Analysis of Results

To further demonstrate the effectiveness of the algorithm can be used to provide data sets UCI comparison test.

The Fig. 1 is the original ID3 algorithm and improved algorithm comparison in accuracy and time in different amount of samples in different data sets.

Seen from Fig. 1, as data collection increases, the accuracy of the algorithm increases. The improved accuracy of the algorithms has greatly improved than the original ID3 algorithm. Seen from Fig. 2, when dealing with the same amount of records, the improved algorithm consumes more time than the original ID3 algorithm. In this paper, the improved algorithm sacrifices some time in exchange for a substantial increase accuracy. Time consuming optimization is still needed to continue the research work.

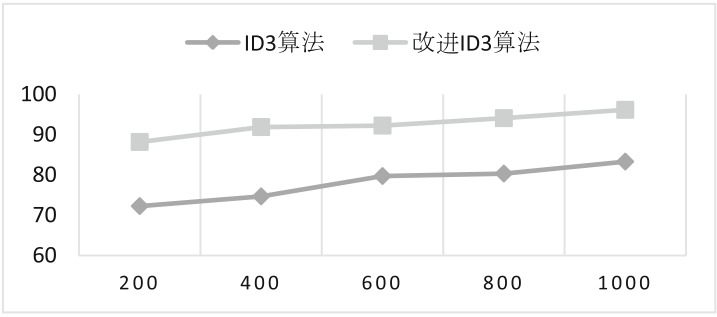

Fig. 1. Different sample sizes algorithm accuracy comparison

Figures 3 and 4 are in different data sets, the original ID3 algorithm and improved ID3 algorithm comparison results in terms of accuracy and computational time. We can see in this paper accuracy of the improved algorithm is significantly better than the accuracy of the original ID3 algorithm, but in the time-consuming slightly larger than the original ID3 algorithm. Thus, less time-consuming exchange for accuracy significantly improved, overall improved algorithm is better than the original ID3 algorithm.

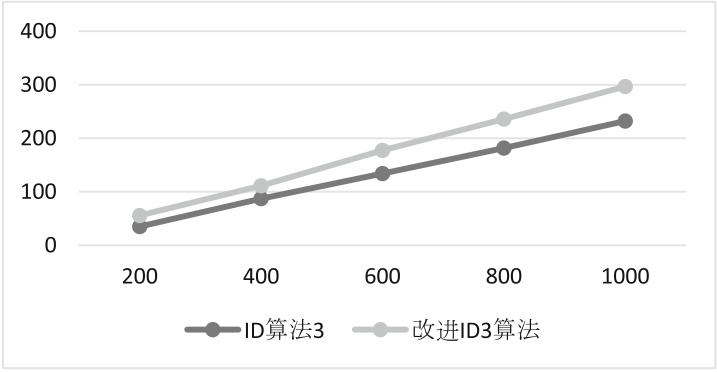

Fig. 2. The comparison of different sample sizes

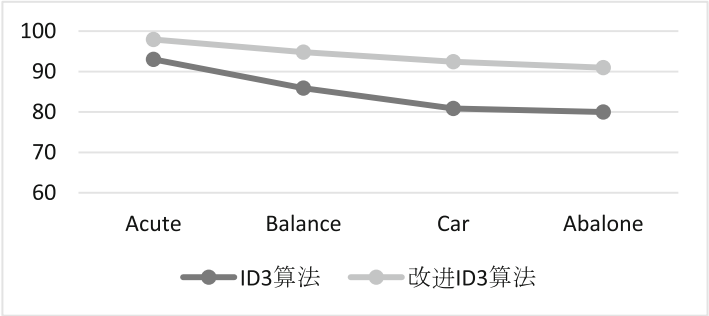

Fig. 3. The comparison of the accuracy level for different data

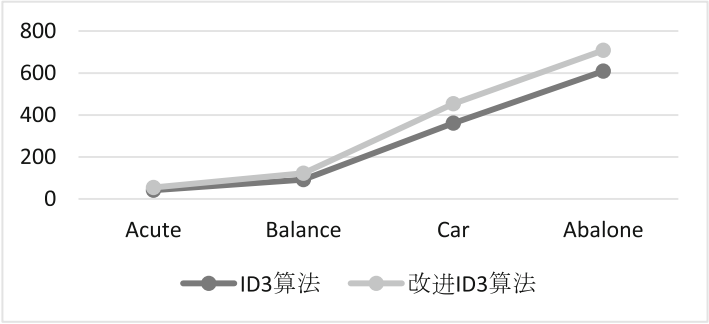

Fig. 4. Algorithm time under different sets of data comparison

5 Conclusion

ID3 algorithm is a decision tree algorithm, the most typical method, a large number of scholars have studied and analyzed. This paper describes the ID3 algorithm and its improved algorithm. First, we describe the basic principles of ID3 algorithm. The biggest shortcoming that is multi-valued attribute bias problem for this algorithm, by introducing information gain rate and parameters, to a certain extent overcome this major drawback. Experimental results show that: the new algorithm overcomes the disadvantages of ID3 algorithm tends to properties of more value than ID3 algorithm and has better classification performance on classification accuracy.

References

1. Wang, Y., Xuegang, H.: Decision tree ID3 algorithm research. J. Anhui Univ. (Nat. Sci. Edn.) **03**, 71–75 (2002)
2. Wang, X., Jiang, Y.: Analysis and improvement of decision tree ID3 algorithm. Comput. Eng. Des. **09**, 3069–3072 + 3076 (2011)

3. Liu, Q.: Improvement of Decision Tree ID3 Algorithm. Harbin Engineering University (2009)
4. Huang, Y., Fan, T., Wang, Y.: Decision tree ID3 algorithm improved algorithm. Comput. Knowl. Technol. **01**, 96–98 (2012)
5. Wang, S.: Decision tree ID3 algorithm and implementation. J. Qiqihar Univ. (Nat. Sci. Edn.) **03**, 64–68 (2012)
6. Wang, S.: ID3 decision tree algorithm analysis and improvement. J. Yichun Univ. **04**, 7–9 (2012)
7. Huang, Y., Fan, T.: Analysis and optimization decision tree ID3 algorithm. Comput. Eng. Des. **08**, 3089–3093 (2012)
8. Wang, M.: Data Mining Algorithm Optimization Research and Application. Anhui University (2014)
9. Wang, B.: Research and Application of Decision Tree Algorithm. Donghua University (2008)