

# Medical Data Mining Based on Decision Tree Algorithm

Ruijuan Hu

Dep. of Foundation

PLA University of Foreign Languages

2 Guang Wen Road

Luoyang 471003, China

Tel: 86-151-3993-8631 E-mail: huruijuan01@126.com

Received: July 12, 2011

Accepted: July 21, 2011

doi:10.5539/cis.v4n5p14

## Abstract

Detailed elaborations are presented for the idea on ID3 algorithm of Decision Tree. An improved method called Improved ID3 algorithm that can improve the speed of generation is brought forward owing to the disadvantages of ID3 algorithm. Moreover, based on Improved ID3 algorithm, data mining for breast-cancers is carried out for primarily predicting the relationship between recurrence and other attributes of breast cancer by making use of SQL Server 2005 Analysis Services. Results prove the effectiveness of Decision Tree in medical data mining which provide physicians with diagnostic assistance.

**Keywords:** Data mining, ID3 algorithm, Improved ID3 algorithm, Breast-cancer

## 1. Introduction

Widely used of computer information management system in medical institutions promotes the digitization of medical information and expands the information capacity in the hospital database. These precious hospital information resources are valuable to medical diagnosis, treatment and medical research (Du Haizhou, 2009, pp.163-167). However, the new problem for promoting the development of hospital and service quality is that, how to automatically upgrade and process the medical database, to provide comprehensive and accurate diagnostic decision-making and health measures. In this context, medical data mining emerged (ZHAO Xiao-fan, 2011, pp.292-294).

As is well known, many algorithms including Association Rules, Decision Tree and Clustering for data mining were presented over time (Han J, 2002, Chapter 2). A trial of medical data mining was made on 285 cases of breast disease patients in HIS (Hospital Information System) using Decision Tree algorithm.

## 2. Medical data mining based on Decision Tree

According to the basic principle of building decision tree, the ID3 algorithm is prone to over-fit problem. An improved ID3 algorithm is put forward based on the threshold of the ID3 algorithm. Detail descriptions about how to build decision tree and propose improved ID3 algorithm are as follows.

### 2.1 The basic principle of Decision Tree

The basic principle of decision tree for constructing tree can be illustrated by ID3 algorithm. It uses the divide-and-conquer strategy in the construction of decision tree, which uses the information gain of characteristic as the heuristic function of attribute selection of a branch in each node of the tree, selecting the information gain as the characteristic of the branch.

ID3 algorithm is described as follows (Kantardzic M, 2002):

Let  $E = D_1 \times D_2 \times \dots \times D_n$  be finite-dimensional vector  $n$ , where  $D_j$  is a finite set of discrete symbols,  $E$  elements  $e = \langle v_1, v_2, \dots, v_n \rangle$  is the sample,  $v_j \in D_j, j = 1, 2, \dots, n$ . Let  $PE$  be the positive sample set,  $NE$  be the anti-sample set, and the number of samples which are  $p$  and  $n$ . According to the principle of information theory, ID3 algorithm is based on two assumptions:

- (1) In the vector space  $E$ , a decision tree classification probability for any sample and the probability for positive sample and anti-sample in  $E$  are the same.
- (2) The expected bits of information needed for making the correct identification by a decision tree are:

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n} \quad (1)$$

If attribute A is the root of the decision tree, A has n values  $\{u_1, u_2, \dots, u_n\}$ , which will divide the sample set E into n subsets  $\{E_1, E_2, \dots, E_n\}$ . Supposing that  $E_i$  contains  $p_i$  positive samples and  $n_i$  negative samples, then a subset of the information needed for the  $E_i$  is  $I(p_i + n_i)$ , and the expected information needed for the attribute A as the root node is:

$$E(A) = \sum_{i=1}^n \frac{p_i + n_i}{p+n} I(p_i + n_i) \quad (2)$$

Therefore, the information gain of classification attribute of A as the root node is  $\text{Gain}(A) = I(p, n) - E(A)$ . ID3 algorithm selection contributes the greatest attribute of  $\text{Gain}(A)$  to a branch of the node attributes, and each node of the decision tree is using this principle until the decision tree is completed (each node of the samples belong to the same class or all Category attributes are used up). One advantage of ID3 is its time of tree construction and difficulty of the task (such as the number of sample set samples, the number of attributes for each sample to study the complexity of the concept of the decision tree nodes) are steadily increasing in linear and the computation is relatively small.

### 2.2 Improved ID3 algorithm

The obvious deficiency of the ID3 algorithm is its over-fit problem (Zelic I, 2000, pp.799-803) (J. Han, 1992, pp.547-559). The training sample set is used to create the whole information between the attributes of samples and their category attributes. With the contribution of the tree, the available samples behind on each floor node are less, in this point, the computation result benefits the effects of training algorithm data classification, but it is bad for the classification of test data results, and then it will produce the over-fit problem.

Generally, the data in the HIS are large, of which the output will be a lot of correlated attributes, each attribute will be placed on the set of candidate attributes, we will determine each attribute of the set; in the case of massive data processing, it will consume a long time to build the decision tree, and the decision tree may not be the most effective product, even it would affect the results of the judgment. For the deficiency of ID3 algorithm, there is an improved method which is based on the threshold of ID3 algorithm (ZHOU Yan, 2011, pp.60-64). We conduct the correlation analysis of the attributes, just taking the attribute strongly correlated to the output (predicted) attributes (S. Muggleton, 1992). As follows:

- (1) Calculate the information gain of each attribute;
- (2) Calculate the filter (critical threshold):

$$\text{lim} = \frac{1}{s} * \sum_{i=1}^m \text{Gain}(A_i) \quad (3)$$

Where s is the number of attributes, m is the number of input attributes. When the information gain of input attributes  $\text{Gain}(A_i) < \text{lim}$ ,  $A_i$  is weakly correlated attribute; when  $\text{Gain}(A_i) > \text{lim}$ ,  $A_i$  is the strongly correlated attribute.

- (3) If the  $\text{Gain}(A_i) < \text{lim}$  in the process of building decision tree is set up, this attribute will be removed in the set of candidate attributes.

In this way, there is no need to judge the weakly correlated attribute in the recursive process again, which can reduce the calculation times and improve the speed of decision tree generation.

### 2.3 Implementation of medical data mining

#### 2.3.1 Data preparation

A Decision tree is essentially a classification method, selecting from the data in the sorted training set to build classification model, and classify the data what is not classified. Meanwhile, the decision tree can be used to prediction (Markey MK, 2002, pp.489-493). This paper uses 285 cases of breast cancer from People's Hospital of Puyang City in the HIS system as a data source for data mining trial, in which 201 patients without recurrence, 84 patients relapsed. The database table of cases are built by SQL Server 2005, and then we extract the patient's age (age), tumor size (tumor-size), number of lymph node invasion (inv-nodes), nodules with or without risk (node-caps), tumor extent (deg-malig), tumor location (breast), the quadrant tumor (breast-quad), radiotherapy (irradiat), 8 attributes with or without recurrence (Class) as the attributes of each cases, through data mining to establish tumor the relationship between recurrence and other attributes.

### 2.3.2 Concrete realization

Firstly, divide the samples of 285 patients cases into two, and then randomly select the 180 samples as the training sample set, so the remaining 105 samples will be the test sample set. The decision tree is built by the ID3 algorithm, and the process is as follows (ZOU Yuan, 2010):

- (1) Create a node N.
- (2) If the samples of the node belong to the same class C, then return N is a classification for the C of leaf nodes.
- (3) If all the attributes for classification have been used up, then return N as a leaf node, the node samples the classification under the maximum value for all classes.
- (4) If (2) and (3) the situation does not exist, then the method of using the maximum information gain from the property list, select the node of the classification attributes, the node N is marked as the classification of property.
- (5) For each selected value of the property the classification of  $a_i$ , the growth of the corresponding child nodes, computing nodes belonging to various sub-samples, the number of samples if a child node 0, then marks it as leaf node, classify the value whichever is parent node of all samples under the most class.
- (6) For each non-leaf node, repeat the above steps until the entire decision tree is completed.

### 3. Analysis of the mining results

Realize Decision Tree algorithm by making use of SQL Server 2005 Analysis Services (Zhu Deli, 2007, Chapter 11). Select Class as the predictable column, and other attributes as input columns, the resulting mining model shown in Figure 1.

As shown in Figure 1, the number of decision tree will take invaded lymph nodes, tumor size and the degree of malignancy as the main factor of determining the breast disease with or without the recurrence, if the number of invaded lymph node is or more than 4, the recurrence rate will be high. The tumor recurrence is also related to the tumor size and the degree of malignancy. When the number of invaded lymph node is less than 2, the tumor size will be 10-14 mm, and the possibility of recurrence will be low, however, the tumor size will be 35-39 mm, and the possibility of recurrence will be high. The smaller tumor is, the lower possibility of recurrence is; when the numbers of the invaded lymph nodes are equal, the greater degree of malignancy is, the higher possibility of recurrence is. This is consistent with the clinical diagnosis.

The dependency and the accuracy of results of mining are shown in Figure 2 and 3. Figure 2 shows the relationship between predicted attributes and other attributes, you can change the strength of the link to observe the dependence of nodes, whether the relationship between the number of invaded lymph node and the tumor with or without recurrence is the strongest, and if the tumor with or without recurrence relates to the tumor size and the degree of malignancy. From Figure 3, the accuracy rate of data mining conducted by decision tree algorithm is nearly 80%, close to or above the effect of clinical diagnosis, which fully proved that the effectiveness of decision tree algorithm in the data mining of breast cancer.

### 4. The comparison between the ID3 algorithm and the Improved ID3 algorithm

Prior to using the Microsoft decision tree algorithm, we conduct the correlation analysis of attributes by the ID3 algorithm, selecting the attributes which has the strongest correlation to the output attributes (predict attributes) as the input column of the Microsoft decision tree algorithm, and then conduct the data mining. We compare the processing time and the accuracy of evaluation in the digging model between the unanalyzed attributes and the analyzed attributes. The results obtained in Table 1.

As shown in Table 1, we can see that when the number of training set are 285, the processing time of improved algorithm by the attribute correlated analysis about 74% of the original algorithm. There is no need to judge the weakly correlated analysis again in the recursive process under the Improved ID3 algorithm, in this way it not only reduces the calculation times but also helps to improve the speed of decision tree generation

### 5. Conclusion

An Improved ID3 algorithm is put forward by studying on ID3 algorithm of Decision Tree and the deficiencies of ID3 algorithm. Conclusions are made by doing data mining on breast cancer patients provided by HIS using SQL Server 2005 Analysis Services that the main factors of the breast disease with or without recurrence and clinical diagnosis are the same, and the accuracy of data mining reached at 80%. So the decision tree algorithm has played a significant role in the medical data mining, which can classify and predict the various testing data and inspecting data in the medical database to help doctors make an objective and effective in patients with the diagnosis, and help doctors' effective and objective diagnosis. Meanwhile, the ID3 algorithm not only plays an important role in the field of data mining algorithm research, but also has practical significance to the

construction of the decision supporting platform.

## References

- 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>.
- Han J, Kamber M. (2002). *Data mining: Concept and technologies.* (Chapter 2).
- 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.
- Markey MK, LoJY & Floyd CE Jr. (2002). Differences between computer-aided diagnosis of breast masses and that of calcifications. *Radiology*, (223), 489-493. doi:10.1148/radiol.2232011257, <http://dx.doi.org/10.1148/radiol.2232011257>.
- S. Muggleton & C. Feng. (1992). E-cient induction of logic programs. In *S. Muggleton, editor, Inductive Logic Programming.* Academic Press.
- ZHAO Xiao-fan & NIU Cheng-zhi. (2011). Research on the Application of Data Mining in HIS Based on Decision Tree. *Computer Knowledge and Technology*, (02), 292-294.
- 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).
- Zelic I, Bercic B, Pikec M & Slavec S. (2000). Implementation and deployment of healthcare management information system. *Stud Health Techno Inform*, 2(77), 799-803.
- Zhu Deli. (2007). *SQL Server 2005 Data Mining complete solutions and business intelligence.* Beijing: Electronic Industry Press, (Chapter 11).

Table 1. Comparison table mining model results

Processing quantity (column)	Digging model of the unanalyzed attributes		Digging model of the analyzed attributes	
	Processing time(ms)	Accuracy of evaluation (%)	Processing time(ms)	Accuracy of evaluation (%)
50	752	52.80	750	52.80
100	998	56.90	920	57.60
150	1386	60.80	1193	64.90
200	1610	69.90	1208	74.30
250	1824	75.10	1445	79.90
285	2607	80.00	1933	85.20

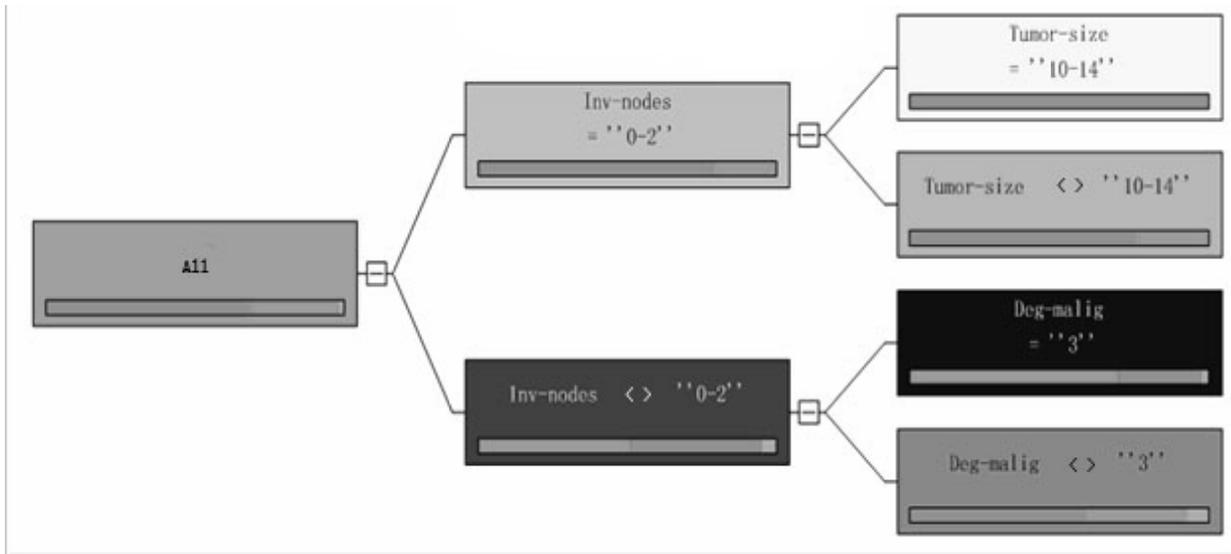


Figure 1. The Result of Data Mining Based on Decision Tree

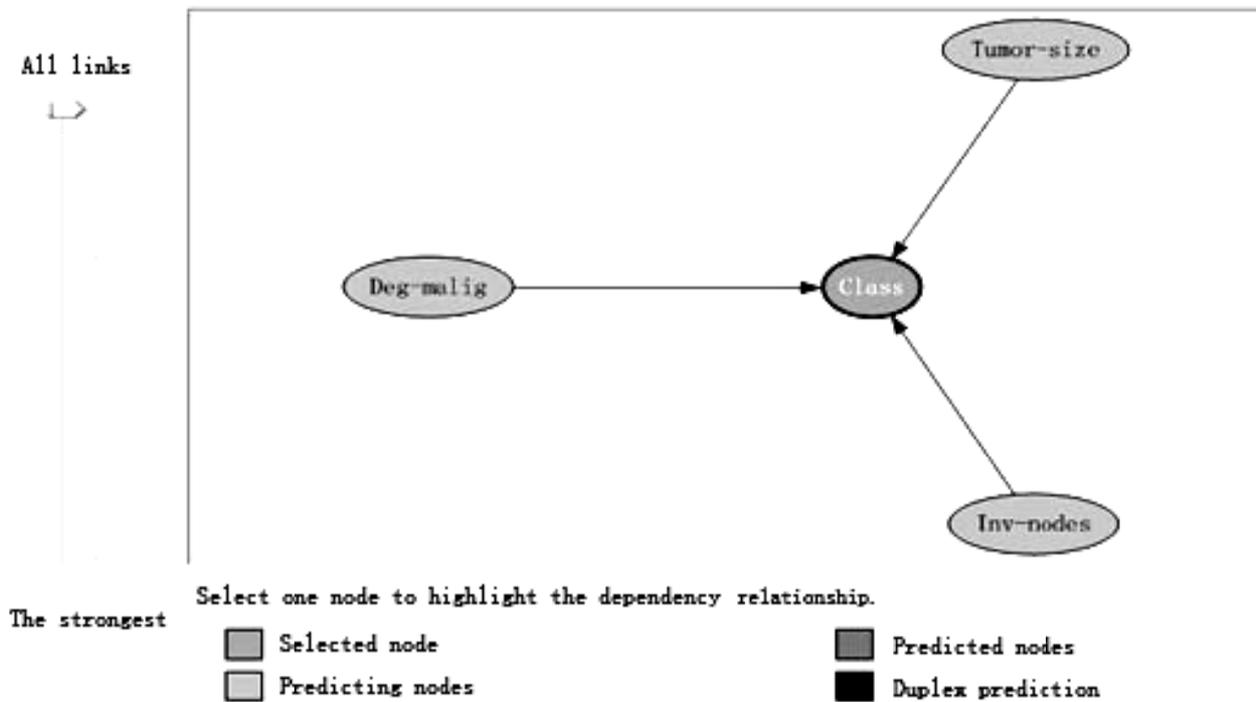


Figure 2. The Dependency Relationship Graph

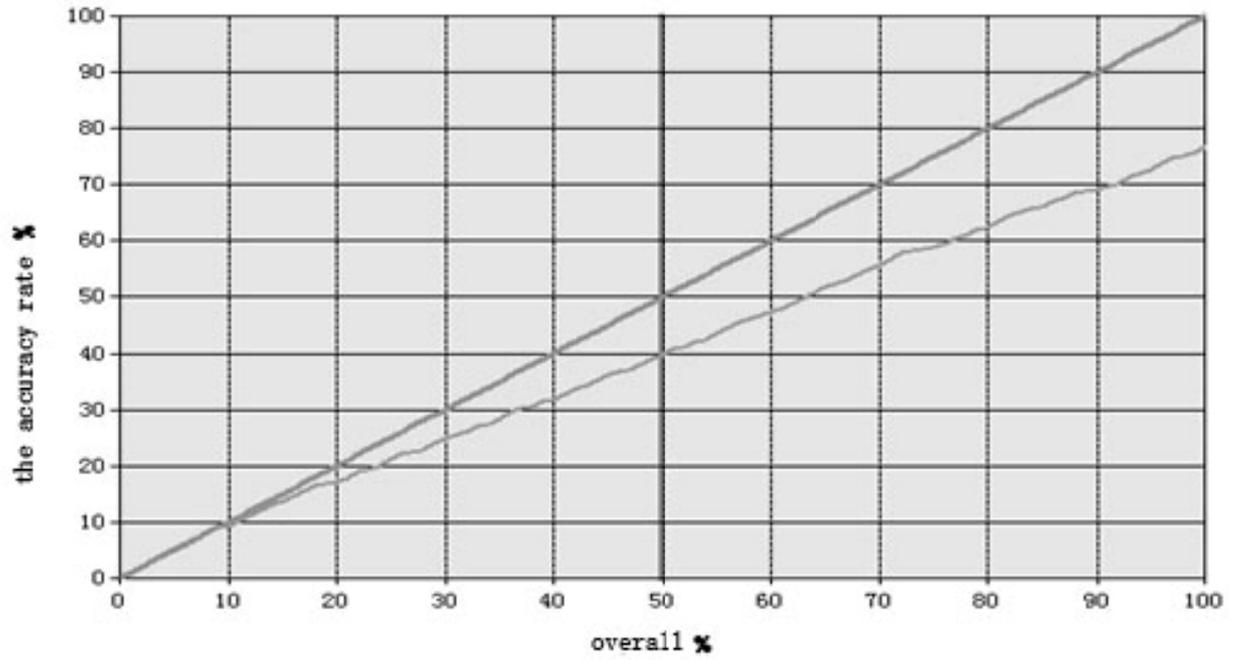


Figure 3. The Accuracy of Data Mining Results