Asian Journal of Information Technology 4 (11): 1080-1085, 2005

Learning Classification Rules under Multiple Costs

[1]Ni Ailing, [2]Shujie Yang, [1]Xiaofeng Zhu and [1,3]Shichao Zhang
[1]Computer Department, Guangxi Normal University, People's of China
[2]Capital Normal University, Beijing, People's of China
[3]Faculty of IT, University of Technology Sydney, Australia

Abstract: Fully taking into account the hints possibly hidden in the absent data, this paper proposes a new criterion when selecting attributes for splitting to build a decision tree for a given dataset. In our approach, it must pay a certain cost to obtain an attribute value. We also consider discounts in test costs when groups of attributes are tested together. When consumer offers finite resources, we can make the best use of the resources as well as optimal results obtained by the tree. In addition, we also put forward advice about whether it is worthy of increasing resources or not.

Key words: Classification rules, absent data, selecting attributes

INTRODUCTION

Existing machine learning and data mining algorithms depend strongly on the quality of the data. And the data quality has been very important yet challenging. However, the data is often incomplete in real-world applications. There are two major kinds of the incompleteness of the data: one is missing, that is to say, the data exist but it is missing now; the other is absent, namely, there was not any data originally. To the former, there are many methods to patch up the missing data; the latter, e.g., in the database of medical diagnosis to obtain the value of an attribute, resources is required in most examination. Those examination need cost and part of examination is expensive. So to those people who have not enough money to perform the expensive examination, the attribute value cannot be obtained but be absent. It is unmeaning and unreasonable if we patch the absent value. From the perspective of classical decision tree learning criterion, these incomplete data is disadvantageous for the classification right-ratio. But from the view of reality, the presence of these absent data is because of limited resources. And most of the resources are limited, so the absent data is inevitably. That is to say, the absent data is an important strategy. So we hope there is a method, which can select the most proper attributes to test according to the limited resources and let other attributes' value absent.

It is enjoyment that some experts has begun to research the test cost and the misclassification cost[1-8]. They considered not only the misclassification cost but also test cost instead of the right ratio of the classification and expect to achieve a balance between the test cost and the misclassification cost[2,5] proposed a new method to build a cost-sensitive decision tree to minimal the total cost (the test cost and the misclassification cost). Ling et al. also proposed several strategies to obtain the optimal total cost[5]. But in their work, the test cost and the misclassification cost have been defined on the same cost scale, such as the dollar cost incurred in a medical diagnosis. For example, the test cost of attribute A1 is 50 dollars, the A2 is 20 dollars, the cost of false positive is 600 dollars and the cost of false negative is 800 dollars.

But in fact, the same cost scale is not always reasonable. For sometimes we may meet difficulty to define the multiple costs on the same cost scale. It is not only a technology issue, but also a social issue. For example, in medical diagnosis, how much money you should assign for a misclassification cost? So we need to involve both of the two cost scales.

In this study, we use two different cost scales for test costs and misclassification costs. That is to say, the scale of test cost is dollars but the misclassification cost just is a relative value. For making the best use of limited resources, note that there are possibly hints about information of classification confining within different resources hidden in the database with absent values. So we propose a new method to obtain an optimal result confining within different resources in absent data. In addition, the same money has different effects at different stage. We also put forward advice about whether it is worthy of increasing resources or not.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we first introduce some simple conceptions that used in this paper. In Section 4, we present the process and the method of building decision tree. The strategy of dealing with the discount is introduced in Section 5. Then we

Corresponding Author: Ailing Ni, Computer Department, Guangxi Normal University, People's of China

show the results of the experiments in Section 6. Finally we conclude the work in Section 7.

REVIEW OF PREVIOUS WORK

More recently, researchers have begun to consider both test and misclassification costs[1-8]. The objective is to minimize the expected total cost of tests and misclassifications. Turney[6] analyzed a whole variety of costs, such as misclassification costs, test costs, active learning costs, computation cost, human-computer interaction cost, etc, in which, the first two types of costs are the misclassification costs and the test costs.

In[7], the cost-sensitive learning problem is cast as a Markov Decision Process (MDP) and an optimal solution is given as a search in a state space for optimal policies. For a given new case, depending on the values obtained so far, the optimal policy can suggest a best action to perform in order to both minimize the misclassification and the test costs.

Similar in the interest in constructing an optimal learner[8] studied the theoretical aspects of active learning with test costs using a PAC learning framework. Turney presented a genetic algorithm to build a decision tree to minimize the cost of tests and misclassification.

Ling *et al.*[5] propose a new decision tree learning program that uses minimal total cost of tests and misclassifications as the attribute split criterion. They also propose several test strategies to handle the missing value in the test data. But in their tree, it assumes both the test cost and the misclassification cost have been defined on the same scale, such as the dollar cost incurred in a medical diagnosis.

In our work, we address the problems above by building an any-cost sensitive decision tree by involving two kinds of cost scales, which make the best use of given specific resources and minimize the misclassification cost. In addition, we present a new strategy to deal with the discount.

CONCEPTION

Decision tree is one of the classical classifier. ID3 algorithm is one method to build decision tree proposed by Quinlan. It uses the Gain as a measurement to select attributes for splitting to build a decision tree. Later Quinlan use the Gain Ratio instead of Gain in the C4.5 for avoiding partial to an attribute with many values.

Gain ratio:

$$\text{Gain ratio}(A,T) = \text{Gain}(A,T)/\text{SplitInfo}(A,T) \quad (1)$$

where, Gain (A,T) is the information brought by condition attribution A in T, namely Gain Ratio. The information brought by the attribute will be larger when the Gain Ratio is larger. The Gain (A,T) obtained by the following formula:

$$\text{Gain}(A,T) = \text{Info}(T) - \text{Info}(A,T)$$

where, $\text{Info}(T) = -(p_1*\log_2(p_1) + p_2*\log(p_2) + L + p_n*\log 2(p_n))$, p_i is the percentage of the i-th value of decision attribute in all object T.

$$\text{Info}(A,T) \sum_{i=1} |Ti|/|T| * \text{Info}(Ti),$$

where, $|T|$ is the number of all object. $|T_i|$ is the number of the object that value of the attribute A is the i-th value.

The definition of SplitInfo(A) in formula(1) is:

$$\text{Split Info}(A,T) = I(|T_1|/|T|, |T_2|/|T|, \cdots |T_m|/|T|)$$

where, it assumes that the condition attribute A has m values, the number of the ith value is Ti, so

$$\text{Split Info}(A,T) = (|T_1|/|T| * \log_2 |T_1|/|T| + L + |T_m|/|T| * \log_2 |T_m|/|T|)$$

So when the Gain Ratio of an attribute is larger, it has more information.

Misclassification cost and test cost: If there is an error in a test, it must pay the misclassification cost. Further more, the costs are different when the errors are different. For example, if a case is positive (illness) but it is predicted to be negative (no illness), he will pay the cost of medicine, on the contrary, he may pay his life. The former error is False Negative (FN), the latter is False Positive (FP). In previous work, the cost of the two kinds of error and the test cost are presented with the same scale. But in fact, people own different sum of money may have different idea about the misclassification cost. For instance, haves may hope to reach the highest right ratio and consider nothing about money, while have-nots may hope to reach an acceptable right ratio because of his limited money. So the relation between the test cost and the misclassification cost is varying from people to people. In this paper, we present the cost of FP and FN in relative value, which can be given by experts. The relative value only implies that the relation between FP and FN. It has no relation with money.

Test cost is obtained by the knowledge of domains. For medical diagnosis, the test cost can be obtained from

the hospital and generally, it has the same scale with the money, such as dollar, etc.

The bias of experts: People who have domain knowledge (such as the doctor in the medical diagnosis) always have some bias about some special test because of their knowledge and experience. So here the expert of the domain can present the bias about attributes w_i, which is the bias of the i-th attribute. If he has no idea about the bias, the default bias is 1 to all attributes.

THE METHOD AND THE PROCESS OF BUILDING DECISION TREE

Selecting the attributes for splitting: In the early method of decision tree, the right ratio of classification is most important. It uses the fewest attributes to obtain the highest right ratio of the classification. So the Gain (in ID3) and the Gain Ratio (in C4.5) are the criteria for selecting attributes for splitting in building decision tree. But in our view, we hope to obtain optimal results when the cost of the false positive and the false negative are different and the resources are limited. So only using the Gain or the Gain Ratio is not properly. In our strategy, the Performance is equal to the return (the Gain Ratio multiply the total misclassification cost reduction) divided by the investment (test cost). That is to say, we select the attribute that it has larger Gain Ratio, lower test cost. In addition, it can decrease the misclassification cost sooner. The Performance defined as follows:

$$\text{Performance}(A_i)' = (2^{\text{Gain ratio}(A_i,T)} - 1) * \text{Redu\_Mc}(A_i) / (\text{TestCost}(A_i) + 1) \quad (2)$$

where, GainRatio(Ai,T) is the Gain Ratio of attribute Ai TestCost(Ai) is the test cost of attribute Ai. Redu_Mc(A_i) is the decrease of misclassification cost brought by the attribute A_i.

$$\text{Redu\_Mc}(A_i) = \text{Mc} - \sum_{i=0}^{n} \text{Mc}(A_i)$$

where, Mc is the misclassification cost before testing the attribute A_i. If an attribute A_i has n branches, $\sum_{i=0}^{n} \text{Mc}(A_i)$ is the total misclassification cost after splitting on A_i. Thinking of the bias of experts, we have the following formula:

$$\text{Performance}(A_i)' = (2^{\text{Gain ratio}(A_i,T)} - 1) * \text{Redu\_Mc}(A_i) / (\text{TestCost}(A_i) + 1) * W_i \quad (3)$$

So formula (3) is the criterion for selecting attribute for splitting to build a decision tree. We select the attribute A, when Performance(A) = max(Performance(A_i)), i is from 1 to m. m is the number of attributes.

Building tree: We will deal with the data, in which there are absent values. For example, in medical diagnosis database, some test is absent because of the restriction of money. It assumes that there will be the same situation, that is to say, there must be someone who has not enough money to finish all the tests in the future. So it is necessary that we make the best use of his limited resources to reduce the misclassification cost. Our goal is to build a decision tree using the data with absent values for finding the optimal attributes to test confined by the limited resources. Material process is as follows:

First, according to the formula (3), we select the optimal attribute. If some of the attributes' Performance are the same, the criterion to select test attribute should be follow in priority order: 1) the bigger Redu_Mc; 2) the bigger test cost. For our goal is to minimize the misclassification cost.

Second, we split a node into k+1 child-nodes, where, k is the value's number of the splitting attribute, 1 is a branch whose value is null. That is to say, there is a null branch to gather those cases whose value of splitting attribute is null. We use this null node strategy to achieve the goal that making the best use of the limited resources. In the process of building tree, we select the attribute that it has the highest ratio of performance to cost. So the splitting attribute in the parent node may has the highest test cost than the splitting attribute in the children node of the decision tree. When we test a case with limited resources, the left resources will decrease because of the test cost. If the leaving resource is not enough to test the next attribute, the test will not be done. But if we use the null embranchment, we can let the case enter the child node along the null embranchment. The splitting attribute in the children may be cheap enough to be test. So we can make the best use of the limited resources to obtain more information to decrease the misclassification cost.

For simplicity, we will illustrate the effect of the null branch with the following table and part of a decision tree (Fig. 1). There are seven attributes with different test costs in the Table 1. It assumes that we have built a decision tree with a given database.

It assumes that there is only 40 dollars left when it should test the attribute A_2. From table 1, we know that the test cost of A_2 needs 50 dollars. Obviously the resources are not enough to perform A_2. If there is no null branch, the test can only stop in the internal node A_2, the left resources will be wasted. But if we have null branch, we don't do the test A_2 and we along the null branch and enter the test of A_3. We also know from the Table 1 that the test cost of A_3 is 20 dollars. For we have 40 dollars left which is bigger than 20, the test of A_3 will be done. So we

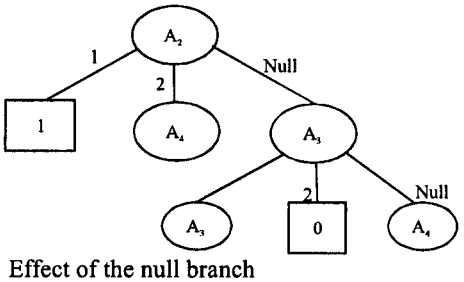

Fig. 1: Effect of the null branch

Table 1. Seven attribute and their test costs

attribute	A_1	A_2	A_3	A_4	A_5	A_6	A_7
Test cost	30	50	20	30	20	70	40

can obtain more information by the null branch. The misclassification cost will be decreased, too.

The condition of stopping building tree is similar to the C4.5. That is to say, when one of the following two conditions is satisfied, the process of building tree will be stopped. a. all the cases in one node are positive or negative; b. all the attributes are run out of. Because different people have different resources, so we build tree with all attributes. When the resources are enough and the case does not reach a leaf node, more tests will be done to decrease the misclassification cost. If the resource is used up and the case does not reach a leaf node, the case will stop in an internal node. And the criterion for judging whether an internal node is positive or not is as follows.

The criterion for judging the class of a node: For a node, P denotes the node is positive and N denotes the node is negative. The criterion is as follows:

$$P \text{ if } p*FN > n*FP$$
$$N \text{ if } p*FN < n*FP$$

where, p is the number of the positive case in the node, while n is the number of negative case. FN is the cost of false negative and FP is the cost of false positive. If a node including positive cases and negative cases, we must pay the cost of misclassification no matter what we conclude. But the two costs are different, so we will choice the smaller one. Here we consider that the cost of the right judgment is 0. For example, there is a node including 20 positive cases and 24 negative cases and if the cost of the false positive and false negative is the same, the node will be considered negative. But in our view, they are different, such as FN is 800 and FP is 600, the node will be considered positive because 16000 is larger than 14400.

In reality, if we consider a patient (positive) as a healthy man (false negative), maybe he will lost of his life. On the contrary (false positive), he may pay the cost of medicine, etc. So the cost is different and in general, we think that the cost of false negative (FN) is larger than the cost of False Positive (FP).

DEALING WITH THE DISCOUNT

In practice application, it maybe cheaper when we do some test together. For example, if we perform two kinds of blood test at the same time, the total cost may be smaller than we test them separately. This kind of cost was presented by P. Turney. He calls it Conditional Test Cost [6].

In our strategy, we group the tests together as a new attribute if the test cost of all these tests is discount when they are test at the same moment. The values of the new attribute are the combination of the values of the original attributes. Such as, there are two attributes A_1 (1,2), A_2 (1,2,3) grouped together. The values of the new attribute are {(1,1) (1,2) (1,3) (2,1) (2,2) (2,3)}, so we can mark them as (1,2,3,4,5,6). The test cost is group discount.

After dealing with the discount attributes, we can use the strategy of building tree illustrated in part 3 to build tree. But the difference is that all the attributes included in the new attribute should be deleted from the candidate attributes if the new attribute had been selected for splitting. When we think of the discount, the number of attributes is n+1, n is the true number of attributes and 1 is the new attribute. For example, if the attribute A_2 (cost: 50 dollars) and A_3 (20 dollars) are grouped together, the total test cost is 60 dollars. The new attribute is A_8. So, in a decision tree, the attributes A_2 and A_3 may appear in different nodes or in the same node, e.g. A_8.

Figure 2 is part of a decision tree. It assumes that attribute A1 has three values, A_2 and A_3 have 2 values. As a result, A_8 has 4 values. We can see that the attribute A_2 and A_3 may be grouped together or appear individually. Whether they are grouped together or not depends on the formula (3).

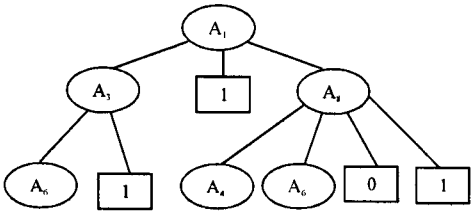

Fig. 2: A part of decision tree

EXPERIMENTS

We use the Mushroom dataset from the UCI Machine Learning Repository in the experiment. For test examples, a certain ratio of attributes are randomly selected and marked as unknown to simulate test cases with absent values. The test cost and the misclassification cost of the dataset are unknown. So we simply assign certain values as dollars for these test costs. This is reasonable because we compare the relative performance of the two strategies under the same costs. We choose randomly the test costs of all attributes to be some values between $10 and $100. The misclassification cost is set to 600/800 (600 for false positive and 800 for false negative). Note that here the misclassification cost is only a relative value. It has different scales with test costs.

For finding out the influence of the null branch in the decision tree, we compared the misclassification costs brought by the two kinds of decision tree, one has the null branch and the other has no null branch. The decision tree with no null branch keeps examples with missing values in internal nodes and does not build branches for them during tree building. When classifying a test example, if the tree encounters an attribute whose value is unknown, then the class probability of training examples falling at the internal node is used to classify it.

From the Fig. 3, we can conclude the following results:

First, with the increasing resources, at the most of the situations, the misclassification cost of the null branch strategy decreases sooner than the no null branch strategy. This is because the null branch can make the best use of the limited resources. As a result, the misclassification cost is smaller. But we also find from the results in the Figure 3 that the difference is not obvious on occasion. The primary reason is that the null branch has no use when the cost of the attribute in the children node is more expensive comparing to the parent node and the null branch has no effect.

Second, with the sum of the resources increasing, we conclude from the experiments that the misclassification cost is decreasing. But at the same time, we noted that the slope is different when the sum of resource is different. That is to say, the speed of the decreasing is not the same. And the speed is sooner when the sum of the resource is fewer. When the sum of the resource is larger than a certain value, the decrease of misclassification cost is very little. So in reality application, we can put up with some advices. For example, we will advise a patient to test the next attribute if the test can decrease the misclassification cost a lot by paying a little test cost. On the contrary, we will advise a patient not to do the next

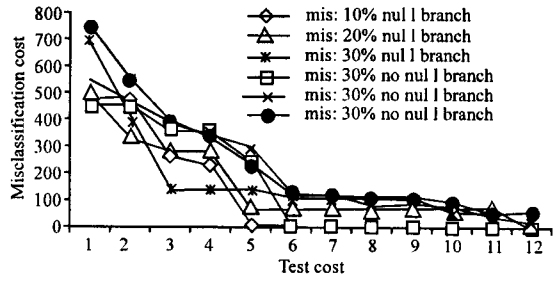

Fig. 3:

test. If we want to give a piece of advice to a patient, we must know the next test cost and the misclassification cost after the next test. The test cost is static, so we can get it directly. But the misclassification cost is different for the different values of the next attribute. We can count the average misclassification cost for all the child nodes of the next test. As a result, we can present the average misclassification cost, the misclassification cost now and the test cost to the patient then put forward the advice.

CONCLUSION

In this study, we have proposed a decision tree learning algorithm to minimize the misclassification cost with any given resources and built an any-cost sensitive decision tree by involving two kinds of cost scales. We have also considered possible discount on tests performed in groups of attributes. In addition, we have put forward a piece of advice according to the decision tree. We have experimentally evaluated the proposed approach and demonstrated it is efficient and promising.

In our future work, we plan to work with medical doctors to apply our algorithms to medical data with real costs. We also plan to incorporate other types of costs in our decision tree learning and test strategies.

REFERENCES

1. Qin, Z., S. Zhang and C. Zhang, 2004. Cost-sensitive Decision Trees with Multiple Cost Scales. Proceedings of AI 2004, pp 380-390.
2. Chai, X., L. Deng, Q. Yang and C.X. Ling, 2004. Test-cost sensitive naive bayes classification. Proceedings of IEEE International Conference on Data Mining (ICDM), 2004.
3. Zhang, S., Z. Qin, C.X. Ling and S. Sheng, 2006. Missing is useful: Missing Values in cost-sensitive decision trees. IEEE Transactions on Knowledge and Data Engineering, to appear in 2006.

4. Turney, P., 1995. Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm, J. Artificial Intelligence Res., 2: 369-409.
5. Charles, X.L., Q. Yang, J. Wang, S. Zhang, 2004. Decision trees with minimal costs. Proceedings of the 21st International Conference on Machine Learning (ICML), Banff, Canada.
6. Turney, P., 2000. Types of cost in inductive concept learning. Proceedings of the Cost-sensitive learning Workshop at the 17th ICML-2000 Conference, Standford, CA.
7. Zubek V., and T. Dietterich, 2002. Pruning improves heuristic search for cost-sensitive learning. In Proceedings of the Nineteenth International Conference of Machine Learning, Sydney, Australia, 2002: 20-35.
8. Russell Greiner, Adam J. Grove. Learning Cost-Sensitive Active Classifiers. Artificial Intelligence, Volume: 139, Issue: 2, August, 2002.