

## A Strategy of Constructing Heterogeneous Cost-sensitive Decision Tree

Yuan Dingrong

College of Computer Science and Information Technology,  
Guangxi Normal University, Guilin 541004, China  
e-mail: dryuan@mailbox.gxnu.edu.cn

Duan Qiaoling

College of Computer Science and Information  
Technology, Guangxi Normal University, Guilin  
541004, China, e-mail: qiaolingduan@163.com

Huang Xiaomeng

College of Computer Science and Information Technology,  
Guangxi Normal University, Guilin 541004, China  
e-mail: ljbhl01@163.com;

Huiwen Fu

College of Computer Science and Information  
Technology, Guangxi Normal University, Guilin,  
541004, china, e-mail: huiwen\_fu@163.com

**Abstract**—Usually, the algorithm of constructing cost-sensitive decision tree assume that all types of cost can be converted into a unified units of the same price, apparently how to construct an cost conversion function is an challenge. In this paper, a strategy of constructing heterogeneous cost-sensitive decision tree is designed and the different cost are take into account together in split attribute selection. What's more, an attribute selection model based on heterogeneous cost-sensitive is constructed and the pruning strategy based on cost-sensitive is designed. The experimental results show that the proposed method is correct and more efficient than the present other methods

**Keywords**-Decision-tree Learning, Cost-sensitive Decision tree, Heterogeneous Cost-sensitive.

### I. INTRODUCTION

Decision tree learning is an active and important research topic in machine learning and data mining. Typical algorithms on decision tree such as ID3, CART and C4.5 have widely and success fully been used in real applications. This algorithms is to for maximally improving the accuracy of classification, however, it makes the kind of classifier bias to the main classes and ignores the minor classes, which has less influence on the accuracy but have great influence on results. That is to say, it neglects the cost problem aroused by misclassification, Based on the application, Elkan(2001) proposed Cost-sensitive Decision Tree, however, it mainly focuses on the misclassification holding the same cost, but practical application is not as this, for example, the cost of misclassifying a sheep into a group of wolves is only losing a sheep, but the cost of misclassifying a wolf into a flock of sheep is losing a flock of sheep. This lead to the research of the CSL(Cost-sensitive Learning) widely and thoroughly.

The split attribute selection is a key and basic step of decision tree construction. The most popular attribute selection method focuses on the measurement of attribute information gain (Gini Coefficient), this will lead to bias to attributes holding much value. Therefore, Mitchell replaced the information gain with average information gain[1]. But, when the cost of misclassification can not be ignored, it is a natural to take into account the cost mechanism and attribute

information together as criteria of split attribute selection, This is to minimize the total misclassification cost. Usually, the cost does not only come from misclassification, it could be also aroused by testing and computing etc. So the present single cost mechanism is powerless for the heterogeneous cost, Zhang(2010) [2] proposed a method for split attribute selection based on the heterogeneous cost. But in his later works, Zhang thinks the method in 2010 can clearly amplify the influence of test cost on attribute selection, and then improved it in 2012, but there are still limits in the method. In this paper, we proposed a method to solve the heterogeneous cost balance problem.

The rest of paper is organized as follows. Section 2 briefly recalls the related work. The related foundation is in section 3, and section 4 is for the proposed strategy. Experimental results are presented in Section 5 and our conclusion in Section 5.

### II. RELATED WORK

Since Quinlan (1986) proposed decision tree learning algorithm ID3, it has drawn the great attention and has been applied wildly. Due to the ID3 inherent limitations, Quinlan (1997) designed algorithm C4.5 based on ID3, which adopt the information gain to select split attribute, expanded ID3 to deal with continuous value attributes and training samples containing missing value. However, ID3 and C4.5 still cannot solve the cost problem resulting from misclassification, thus Elkan(2001) put forward the cost sensitive decision tree learning [4]. The method proposed by Elkan which only considers misclassification cost sensitive and ignores other cost such as test cost and compute cost. As a result, heterogeneous cost sensitive learning become the important topic in recent years. Zhang in 2010[2] proposed a strategy of split attribute selection based on heterogeneous cost-sensitive, which is as follows:

$$SAS_{CTS}(A) = \frac{(2^{AverageGain(A,T)} - 1)}{TC(A) + 1} * ReduceMc(A) \quad (1)$$

In later works, Zhang in 2012[3] found that information may be ignored because the molecular in formula (1) is too small, and the test costs is likely to be expanded limitlessly

for its denominator too large, therefore, he standardizes as follows:

$$SAS_{CTS}(A) = \left( \frac{2^{AverageGain(A,T)} - 1}{TC(A) + 1} \right)_{normal} * ReduceMC(A) \quad (2)$$

But formula (2) avoids the possibility that the test costs may be amplified, but split attribute information still may be ignored because of its value too small, and misclassification cost may be amplified for its value too big. Thus we put forward a new method to avoid the probability of the attribute information being ignored and misclassification cost being amplified. The relevant concepts are as follows.

### III. FOUNDATION OF COST SENSITIVE DECISION LEARNING

CSL is the extension of traditional decision tree learning, Elkan (2001) defined it as follows:

Definition 1 cost matrix. Let training dataset  $T$  have  $n$  attributes  $a_1, a_2, \dots, a_n$ , the split attribute  $a_i \in \{a_1, a_2, \dots, a_n\}$  has  $m$  separate class identifier  $l_1, l_2, \dots, l_m$ ,  $c_{ij}$  is the cost of classifying the data in class  $j$  into class  $i$ . If  $i=j$  is right classification  $c_{ij} = 0$ . We call the matrix  $C = (c_{ij})$  as cost matrix.

Definition 2 Predict cost. For any sample  $x$ , if it is classified into  $i$ -th class, the possibly total cost is  $L(x, i) = \sum_j p(j|x)C(i, j)$ , We call  $L(x, i)$  as total predict cost of classifying  $x$  into  $i$ -th class, for short, call it as predict cost. If  $x$  is sure to be classified into the  $i$ -th class,  $L(x, i)$  is classification cost.

Definition 3 Misclassification cost sensitive decision. In the process of decision tree, the split attribute is decided by the attribute information and predict cost. We call it as misclassification cost sensitive decision.

Definition 4 Heterogeneous cost sensitive decision. In the process of decision, besides misclassification cost, there are other type costs such as test cost, education cost, involvement cost. We consider all these sorts of costs together at the same time in split attribute selection. The decision is called Heterogeneous cost sensitive decision.

Definition 5 Information gain. Split attribute is defined according to the Gini index[1]. The Gini index is expressed as

$$G(T) = 1 - \sum_j p(l_j)^2$$

where  $p(l_j)^2$  is the probability of a sample belonging to class  $l_j$ . The information gain of attribute  $A$  could be represented as follows:

$$Gain(A) = G(T) - G(A, T) \quad (3)$$

Where,  $G(A, T)$  is the total Gini index after  $T$  being split by attribute  $A$ .

Property 1. The bigger the information gain of attribute  $A$  is, the more likely it will be the split attribute in heterogeneous cost sensitive decision learning.

Definition 6 misclassification cost reduce. If a data  $T$  is classified by attribute  $A$ , the decrease of misclassification

before and after splitting is called misclassification cost reduce, represented as:

$$Reduce(A) = \frac{Mc - \sum_{i \in ClassSet(A)} Mc(A_i)}{Mc} \quad (4)$$

Where,  $Mc$  is the cost before splitting,  $\sum_{i \in ClassSet(A)} Mc(A_i)$  is the total cost after splitting. And  $ClassSet(A)$  is the set of classes after splitting by attribute  $A$ .

Property 2. The more the misclassification cost reduce, the more likely it will be the split attribute

Definition 7 Test cost. The test cost of attribute  $A$ , could be represented as:

$$TC(A_i)_{normal} = \frac{Max(1, TC(A_i))}{Max(1, TC(A_1), TC(A_2), \dots, TC(A_n))} \quad (5)$$

Property 3 The less the test cost of attribute  $A$ , the more likely it will be the split attribute

### IV. THE STRUCTURE STRATEGY OF HETEROGENEOUS COST SENSITIVE DECISION

The significant distinction between cost sensitive decision and the traditional decision lie in the strategy of attribution selection and pruning, the topics discussion as follows:

#### A. The split attribute selection of heterogeneous sensitive cost

Define We design the cost sensitive split attribute selection factor ASF(Attribute Selection Factor) as:

$$ASF(A_i) = \frac{(2^{AverageGain(A_i)} - 1)}{TC(A_i)_{normal}} * Reduce(A_i) \quad (6)$$

Where,  $AverageGain(A_i)$  represents average information gain,  $TC(A_i)_{normal}$  stands for standard testing cost,  $Reduce(A)$  is misclassification cost reduce. The proposed attribute selection factor is influenced by the attribute information, test cost and misclassification loss together.

To strengthen the classification ability of information gain, we use average information gain to replace the traditional information gain. In order to avoid ignoring the influences of information because of the large denominator, we standardize the test cost. Misclassification cost reduce is intended to decrease cost between before and after splitting. Therefore, we have relevant theorem as:

Theorem 1 Let the attribute tags of data set are  $A_1, A_2, \dots, A_n$ , if  $A_i, i \in \{1, 2, \dots, n\}$  is  $Max(ASF(A_i))$  then  $A_i$  is the split attribute.

Proof: According to the attribute selection factor represented as (6), we know that the greater the attribute information gain, the greater the misclassification cost reduce, and the smaller the testing cost, then the greater the  $ASF(A_i)$  is. According to the property (1), (2) and (3), it is obviously. So if  $splitA$  is split attribute, then we have:

$$splitA = \text{Max} (ASF (A_i)) \quad (7)$$

**B. Structure heterogeneous cost sensitive decision tree**

First, according to (6), calculate each candidate split attribute selection factor, and then select the attribute satisfies (7). If a ASF have the same value, the values of misclassification reduce and test cost should be consider according to the application.

If one of the following two conditions is satisfied, the process of attribute selection be stopped.

- 1) In the same node, all the examples belong to the same class;
- 2) All properties are used up.

The structure algorithm of heterogeneous cost sensitive decision tree is described in Algorithm.1

**Algorithm.1 heterogeneous cost sensitive decision tree**

1 Create a root node N
2 If the training set is empty, Then return node N labeled Failure;
3 If all records in the training set belong to the same class, the class is marked node N;
4 If the candidate attribute is empty, return N as a leaf node and label it as the most common class
5 for each candidate attribute_list
5.1 caculate the ASF
5.2 Select the attribute satisfies (7) as split attribute
5.3 Label note as the attribut SplitA ;
6 for each known value SplitA <sub>i</sub> in attribute SplitA
6.1 build a branch from SplitA and label it as SplitA <sub>i</sub>
6.2 Let S <sub>i</sub> is a training sample set of training set, in which SplitA = SplitA <sub>i</sub>
6.3 if S <sub>i</sub> is empty then add a leaf, and label it as the commonest training set else recursive call the algorithm
7 end

Step 5 is the core of this algorithm, it is the step that make heterogeneous cost-sensitive decision tree algorithm different from the traditional one, because multi-cost are considered in the slipt attribute selection together. Other steps are similar to IC4.5.

**C. Pruning strategy**

The The purpose of pruning is to prevent decision tree learning over fitting, the main strategies include (1) Reduce the misclassification, such as pessimistic error pruning and minimum error pruning. In the pessimistic error pruning strategy, we judge whether to prune or not by the wrong sample number before and after pruning, Minimum error pruning is to get the smallest expected error rate of a decision tree. (2) Reduce the computational cost, such as

minimum complexity pruning and minimum description length pruning. In the pruning process of the complexity pruning, the total increased cost of misclassification samples arised from the replacement with leaf node is called computational cost, and the reduced number of leaf node in subtree is called complexity. Minimum description length pruning is to prune according to the value of coding cost of decision tree. The goal is to make the most of the training sample in line with the tree, the coding length in decision tree minimum, the sample data inconsistent with tree node is regarded exceptional codes, and make the coding length in exceptional node minimum also.

Our strategy gives the priority to the misclassification reduce, and then considers test costs; i.e. for any positive numbers  $\alpha, \beta$ , given by users satisfied:

- (1) Reduce(A) <  $\alpha$
- (2) TC(A<sub>i</sub>) >  $\beta$ .

The first, we prun all branches satisfied(1), and then prun the branches satisfied (2)

**V. EXPERIMENT RESEARCH**

In order to evaluate the performance of the algorithm of heterogeneous cost sensitive decision tree (recorded as "HCSDT"), we chose the related methods to compare. Such as "SAS" (zhang (2010,2012) on the heterogeneous cost sensitive learning), "GR" (split attribute selection, prescind cost), "MTC" (take the misclassification cost and test cost with different units into account, but do not consider the accuracy of splitting attribute classification). GR is similar to C4.5 algorithm, but it has no cost sensitive function. The infinite test cost together with limited misclassification cost are taken to evaluate total cost in all this the testing methods. All the experiments are achieved based on the UCI data sets.

**A. Experiments in conditions of different missing value**

From Figure1, we can see that all the misclassification cost increases with the increase of the missing values. The higher the rate of missing data is, the worse the construction of the model is. Comparing HCSDT method and other methods under different missing rate, we know that HCSDT has lowest misclassification cost. Comparing HCSDT algorithms and SAS algorithm, we easily conclude that HCSDT is feasible and reasonable in heterogeneous cost sensitive learning.

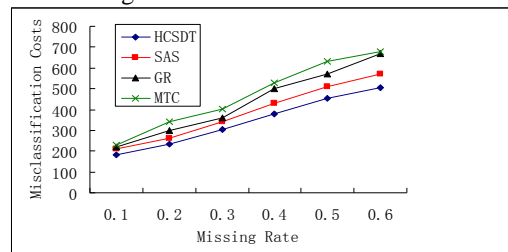


Figure 1. Relation between missing value and cost.

*B. Experiments under different test costs*

In our experiments, the domain of test costs are between 100 to 600 and the y-axis represents misclassification cost. The results also show that the algorithm HCSDT is better than the other algorithms in limited test cost. From Figure2, we can see that HCSDT algorithm has the best total cost in the maximum gain rate and all the resources. In addition, the GR algorithm (the traditional method) obtained the worst results.

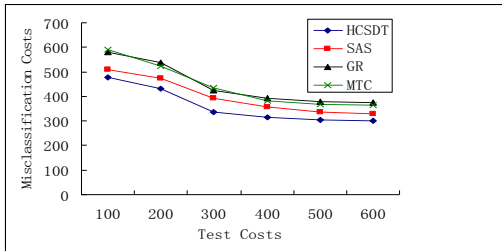


Figure 2. Relation between test and misclassification cost

*C. The comparison of the algorithms' accuracy in experiment*

We can see the accuracy of experiment in TABLE1 clearly. In the cost sensitive decision tree, the accuracy of GR is the lowest and the misclassification is the biggest. However, the classification cost of HCSDT is the best, and the SAS's is a bit worse

TABLE I. THE TABLE OF CLASSIFICATION ACCURACY AND MISCLASSIFICATION IN DIFFERENT ALGORITHMS

Algorithm' name	classification accuracy	misclassification rate
GR	80.07±2.78	19.93±2.78
MTC	83.23±1.36	16.77±1.36
SAS	89.67±1.82	10.33±1.82
HCSDT	93.89±1.91	6.11±1.91

VI. Conclusion

This paper proposes a split attribute selection strategy based on heterogeneous cost sensitive. In the process of attribute selection, to avoid the risk of the attribute

information being ignored for its value too small, misclassification cost reduce and test cost being amplified for its value too big in split attribute selection, we combine attribute information with misclassification cost and test cost together, in order to balance the influence coming from different cost in split attribute selection. The experiments have evaluated the performance of the proposed method and compared with the existing methods in 6 UCI data sets, the result shows that our method is more robust and effective when dealing with heterogeneous cost.

ACKNOWLEDGMENTS

This work was supported in part by the Nature Science Foundation of China (NSFC) under grant 61170131, the Guangxi innovation team project under grant GA060004 and Found of Guanxi Normal University

REFERENCES

[1] Mitchell, T.M., 1997. Machine Learning. McGraw Hill.  
 [2] Zhang, S.C., Cost-sensitive classification with respect to waiting cost. Knowledge-Based Systems 23, 369–378. (2010)  
 [3] Shichao Zhanga. Decision tree classifiers sensitive to heterogeneous costs. The Journal of Systems and Software 85(2012), 771-779.  
 [4] Elkan, C., 2001. The foundations of cost-sensitive learning. In: Proceeding of the Seventeenth International Joint Conference of Artificial Intelligence, Morgan Kaufmann, Seattle, August 4–10, 2001, pp. 973–978.  
 [5] Nunez, M., The use of background knowledge in decision tree induction. Machine Learning 1991.6, 231–250.  
 [6] Tan, M., Schimmer, J., 1989. Cost-sensitive concept learning of sensor use in approach and recognition. In: Proceedings of the 6th International Workshop on Machine Learning, Ithaca, New York, pp. 392–395.  
 [7] Freitas, A., Costa-Pereira, A., Brazdil, P., 2007. Cost-sensitive decision trees applied to medical data. In: Proceedings of DaWak-2007, LNCS 4654, pp. 303–312.  
 [8] Davis, J.V., Ha, J., Rossbach, C.J., Ramadan, H.E., Witchel, E., 2006. Cost-sensitive decision tree learning for forensic classification. In: Proceedings of the 17th European Conference on Machine Learning (ECML), pp. 622–629.  
 [9] Zhang, S.C., Jin, Z., Zhu, X.F., 2011. Missing Data Imputation by Utilizing Information within Incomplete Instances. Journal of Systems & Software 84, 452–459.