MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Cost-Sensitive Decision Trees Applied To Medical Data

Alberto Freitas[a,b]

[a] *CINTESIS – Center for Research in Health Information Systems and Technologies Portugal*
[b] *Department of Biostatistics and Medical Informatics, Faculty of Medicine, University of Porto Portugal*

Abstract

Classification plays an important role in medicine, especially for medical diagnosis. Health applications often require classifiers that minimize the total cost, including misclassifications costs and test costs. In fact, there are many reasons for considering costs in medicine, as diagnostic tests are not free and health budgets are limited. Our aim with this work was to define, implement and test a strategy for cost-sensitive learning. We modified an algorithm for decision tree induction to consider costs, including test costs, delayed costs and costs associated with risk. Then we applied our strategy to train several cost-sensitive decision trees in medical data. Built trees were tested following some strategies, including group costs, common costs, and individual costs. Using the factor of "risk" it is possible to penalize invasive or delayed tests and obtain decision trees patient-friendly.

Keywords:

classification, costs and cost analysis, artificial intelligence, cost-sensitive learning

Introduction

In medical care, as in other areas, knowledge is crucial for decision making support, biomedical research and health management [1]. Data mining and machine learning can help in the process of knowledge discovery. Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [2]. Machine learning is concerned with the development of techniques which allow computers to "learn" [3].

Classification methods can be used to build models that describe classes or predict future data trends. It generic aim is to build models that allows predicting the value of one categorical variable from the known values of other variables. Classification is a common, pragmatic tool in clinical medicine. It is the basis for finding a diagnosis and, therefore, for the definition of distinct strategies of therapy. In addition, it plays an important role in Evidence-Based Medicine. Machine learning systems can be used to enhance the knowledge bases used by expert systems as it can produce a systematic description of clinical features that uniquely characterize clinical conditions. This knowledge can be expressed in the form of simple rules, or decision trees [4].

The majority of existing classification methods was designed to minimize the number of errors. Nevertheless, real-world applications often require classifiers that minimize the total cost, including misclassifications costs (each error has an associated cost) and test (attribute) costs. In medicine a false negative prediction, for instance failing to detect a disease, can have fatal consequences; while a false positive prediction may be less serious (e.g. giving a drug to a patient that does not have a certain disease). Each diagnostic test has also a cost and, for deciding if it is worthwhile pay the costs of tests, it is necessary to know both misclassification and tests costs. There are many reasons for considering costs in medicine. Diagnostic tests, as other health interventions, are not free and budgets are limited.

Misclassification and test costs are on the most important costs, but there are also other types of costs [5]. Cost-sensitive learning (also known as cost-sensitive classification) is the area of machine learning that deals with costs in inductive learning.

Our aim with this work was to study and implement a strategy for learning and testing cost-sensitive decision trees, considering several costs, and with application to medical data.

The rest of this paper is organized as follows. In the next section we expose the main types of costs. Then we review the related work. After that, we explain our cost-sensitive decision tree strategy. Next we present and compare some experimental results. And finally, we conclude and point out some future work.

Types of costs

Turney [5] presents a taxonomy for many possible types of costs that may occur in classification problems. From his enumeration, misclassification and test are on the most important costs. Costs can be measured in many distinct units as, for instance, money (euros, dollars), time (seconds, minutes) or other types of measures (e.g., quality of life).

Misclassification costs

A problem with n classes is normally associated with a matrix $n*n$, where the element in line i and column j represent the cost of classifying a case in class i being from class j. Usually the cost is zero when $i = j$. Typically, misclassification costs are constant, that is, the cost is the same for any instance classified in class i but belonging to class j. The traditional error rate measure occurs when the cost is 0 for $i = j$ and 1 for all other cells.

In some cases, the cost of misclassification errors could be conditional, that is, it could be dependent of specific features or dependent of the moment in time. The cost of prescribing a specific drug to an allergic patient may be different than prescribing that drug to a non allergic patient.

The cost of misclassification can be associated with the moment it occurs. A medical device can issue an alarm when a problem occurs and, in this situation, the cost is dependent simultaneously on the correctness of the classification and on the time the alarm is issued, that is, the alarm will only be useful if there is time for an adequate action [6]. Misclassification costs can also be dependent on the classification of other cases. In the previous example, if one alarm is correctly and consecutively issued for the same problem, then the benefit of the first alarm should be greater than the benefit of the others.

Cost of tests

In medicine, the majority of diagnostic tests have an associated cost (e.g., an echography or a blood test). These costs can be highly distinct between different tests (attributes).

The costs of tests may be constant for all patients or change according to specific patient features. A bronchodilatation test, for instance, have a higher cost for children under 6 years, which means that the feature age have influence in the test cost.

Medical tests can also be very distinct when considering their influence in the "quality of life". A range of tests are completely harmless for patients (e.g., obstetric echography), others can be dangerous and put patient life at risk (e.g., cardiac catheterism), and some can be (only) uncomfortable (e.g., digestive endoscopy).

Some tests can be cheaper (and faster) when ordered together (in group) than when ordered individually and sequentially (e.g., renal, digestive and gynecological echography). Some tests can also have common costs that can be priced only once. Blood tests, for instance, share a common cost of collecting the blood sample. There is not only an economic reduction but also a non-economical reduction in the cost of "worry" the patient.

A number of tests might depend of the results of other tests. The test "age", for instance, may influence the cost of the bronchodilatation test. Some tests can have an increased price as result of secondary effects. Other tests can have patient specific, time dependent or emergency dependent costs.

In general, tests should only be ordered if their costs are not superior to the costs of classification errors.

Related work

In inductive learning (learning by examples), the majority of the work is concerned with the error rate (or success rate).

Nevertheless, some work has been done considering non-uniform misclassification costs, that is, different costs for different types of errors [7, 8]. Other literature is concerned with the cost of tests, without taking into account misclassification costs [9, 10].

And there is also some work concerned simultaneously with more than one type of costs, including the work of Turney [11], Zubek and Dietterich [12], Greiner et al. [13], Arnt and Zilberstein [14], and Ling et al. [15-20]. At this level, the work of Turney [11] was the first to consider both test and misclassification costs. Next, we give a brief overview of this work considering both costs.

Turney [11] implemented a system, the ICET system, that uses a genetic algorithm for building a decision tree that minimizes test and misclassification costs. The ICET system was robust but very time consuming.

Several authors associated the cost problem to a Markov decision process that has the disadvantage of being computationally expensive. Zubek and Dietterich [12] used an optimal search strategy, while. Arnt and Zilberstein [14] included a utility cost for the time necessary to obtain the result of a test.

Greiner et al. [13] analyzed the problem of learning cost-sensitive optimal active classifiers, using a variant of the probably-approximately-correct (PAC) model.

Chai et al. [15] proposed a cost-sensitive naïve Bayes algorithm for reducing the total cost. Ling et al. [16] proposed a decision tree algorithm that uses a cost reduction splitting criteria during training, instead of minimum entropy. After that, Sheng et al. [17], presented another approach where a decision tree is built for each new test case. In another paper, Sheng et al. [18] proposed a hybrid model that result from the integration of a decision tree sensitive to costs with a naïve Bayes classifier. Zhang et al. [19] compared strategies for checking if missing values should be or not obtained and stated that, for tests with high costs or high risk, it should be more cost-effective to not obtain their values. Recently, Ling et al. [20], updated their strat-

egy for building cost-sensitive decision trees, with the inclusion of sequential test strategies.

Our cost-sensitive strategy

Next, we briefly describe our strategy for implementing a cost-sensitive decision tree. Our aim was to implement a tool for building decision models sensitive to costs, namely test costs, misclassification costs and other types of costs.

We opt to use decision trees because they have several interesting characteristics for health professionals. They are easy to understand and use, and present an intuitive and appealing structure. Their structure is congruent with decision making methods that physicians normally use in daily routine, when they try to understand which is the best diagnostic test or the best treatment for one patient [21].

We modified the C4.5 algorithm [22] to contemplate costs and, consequently, to generate cost-sensitive decision trees. Specifically we used de j48 class, implemented in the open source package for data mining Weka [3].

Cost function

We adapted the decision tree splitting criteria to contemplate costs, through the following cost function (1):

$$\frac{\Delta I_i}{\left(C_i \phi_i\right)^{\omega}} \tag{1}$$

Where ΔI_i is the information gain (or gain ratio) for attribute *i*, C_i is the cost of attribute (test) *i*, ϕ_i is the factor of "risk" associated with attribute (test) *i*, and ω is the factor of power.

When building a tree, for each node, the algorithm will select the attribute that maximizes the defined heuristic for the cost function. The cost function does not consider misclassification costs, as these costs do not have influence in the decision tree splitting criteria [22].

Attributes without cost, as age and gender, are assigned the value 1. Attributes with higher costs lead to lower results in the cost function and have, consequently, fewer chances to be selected.

The factor of "risk" is an influent piece in the cost function. This factor was introduced to penalize attributes that might be invasive, may cause discomfort and disturb, or could somehow contribute to low patient quality of life (e.g., the invasive test "coronary angiography"). The value 1 means absence of influence and is equivalent to a completely innocuous test, while values higher than 1 means that exists influence. Higher factors leads to lower results in the cost function and, therefore, to lower possibilities for and attribute to be selected.

If both test cost and factor of "risk" are equal to one, then their participation is neutral and, thus, the cost function is basically the traditional information gain function. The cost of the attribute can be adjusted in two ways. It can increase or decrease considering the inoffensiveness of the test (factor of "risk"), or it can be modified by the factor of power. The factor of power is a general parameter that is equally applied to all tests, reducing or increasing the influence of costs in attribute selection.

The factor of "risk" can also be used to penalize delayed tests. A longer test can have consequences in the patient quality of life and may increase other costs, as those related to staff, facilities and increased length of stay. Between two similar tests, but with one longer than the other, it makes sense that the faster test should be preferred. Hence, considering the average length of tests, an adjustment can be made through the factor of "risk".

The factor of power regulates the influence of costs, as it can make trees more (or less) sensitive to costs. This factor may also be used to adequate the cost function to the used scale. An inexpensive test, such as the "age" test, has the same cost for any scale. But, tests with real costs have clearly different values in different scales. If a test "A" cost 10 in the Euro scale then it costs 1000 in the centime scale, that is, the ratio between the tests "age" and "A" are 1 to 10 in the Euro scale and 1 to 1000 in the other. To avoid that a change in the scale could benefit some attributes, it is important to adjust it with the factor of power.

For a factor of power of 0 the costs are not considered as the denominator of the cost function became 1. This situation is equivalent to consider the original information gain. With an increase in the factor of power, the costs of tests will have more influence and less expensive tests will preferred. The factor of power typically will assume a value between 0 and 1.

For making our model sensitive to misclassification costs, we used a meta-learner implemented in Weka, CostSensitiveClassifier [3].

Decision tree test and usage strategies

After building cost-sensitive models, we want to test and use them. For that, we have a test strategy where we consider the cost of a test individually or in a group. Within groups of tests it is possible to distinguish between (i) tests ordered simultaneously or (ii) tests that have a common cost.

Sometimes (i), there is a big difference between the cost of a group of tests and the sum of individual tests costs. Many medical tests do not have an immediate result and there-

fore it also important to consider the length of time for the group against the total delay of the individual tests. When a medical doctor is at a node of the decision tree, and have the first of a group of tests, he must decide if he will ask for a group of tests or only for the node test. If he orders a group of tests, then the cost considered will be the group cost, even if some tests in the group are not used.

In the other situation (ii), it is possible to separate a common cost for a group of tests. In a group of blood tests, for instance, the cost of collecting blood is a common cost for all tests in the group. Only the first test of the group will be priced for that common cost.

Delayed tests are considered in the training phase of the decisions tree. Slower tests will, therefore, have tendency to be tested only after faster ones.

Individual costs

Our strategy also allows (a) to consider specific patient characteristics, (b) to modify test costs in situations where their values are already known (tested previously), and (c) to consider availability and slowness of certain tests.

The cost of each test can be conditional on the characteristics of the patient, that is, it is possible to have a variable cost associated with specific features of the patient. As seen before, the age can change the cost of the test "bronchodilatation" (higher for children less than 6 years). The comorbidity index of a patient is another example, as it can influence the cost of some tests and, therefore, could be used to adjust the cost of the test. For a specific test, different patients may require additional tests, consequently with an increase in costs.

In other circumstances (b), some tests might have been obtained previously and, logically, their original costs should not be considered again. Consequently, we adopt another approach for building trees where, for each new instance (patient), tests with known values are considered without cost and without "risk". For each new vector of distinct costs a new decision tree is built.

Resources are not infinite neither are always available. A specific test may be conditional on the availability of a medical device in a specific time period. In these cases, it is possible to exclude that test and build a new tree. Optionally, it is possible to increase de factor of "risk" of that test, decreasing it probabilities of being selected. As the availability is not constant over time, this problem should be analyzed for each case (patient); this is a circumstantial cost and not a cost intrinsic to the patient.

Experiments

We tested our cost-sensitive decision tree with several datasets, including the Pima Indians Diabetes. This dataset contains 768 instances and a class that assumes 2 values, "healthy" (500) or "diabetes" (268)[1]. In Table 1 we can see test costs (all attributes are numeric).

Table 1 - Attribute costs for Pima Indians diabetes

Test	Cost ($)	Group cost ($)
a. Number of times pregnant	1	
b. Glucose tolerance test	17.61	b+e=38.29
c. Diastolic blood pressure	1	
d. Triceps skin fold thickness	1	
e. Serum insulin test	22.78	b+e=38.29
f. Body mass index	1	
g. Diabetes pedigree function	1	
h. Age (years)	1	

Tests "glucose tolerance" and "serum insulin" are not immediate and have a distinct cost (other attributes have a symbolic cost of $1). Moreover they share a common cost, $2.1, from collecting blood. As these attributes are not immediate, when using the tree it is necessary to decide if both tests will be ordered together (in group) or not.

We ranged the factor of power from 0.0 to 1.0 and induced decision trees using 10-fold cross-validation. After that, we considered cost zero for cases classified correctly and varied misclassification costs, from $10 to $1000, with equal costs for false negative and false positives. In Table 2 we can see the results of the evaluation of five decision trees for that range of misclassification cots. Notice that for factor of power equal to zero, the model is equal to the obtained by the traditional C4.5.

1 UCI Repository Of Machine Learning Databases, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/

Table 2 - Average costs in the evaluation of 5 decision trees, for a range of misclassification costs and factors of power

fp:	**0.0**	**0.1**	**0.2**	**0.5**	**1.0**	**95% IC var.**
Accuracy (%):	73.8	73.7	72.3	68.5	68.2	
$10	23.4	21.9	19.5	6.4	5.7	±0.5
$20	26.0	24.5	22.2	9.5	8.8	±0.8
$50	33.8	32.4	30.6	19.0	18.4	±1.7
$100	46.9	45.5	44.4	34.7	34.2	±3.3
$200	73.1	71.8	72.2	66.2	66.0	±6.5
$500	152	151	155	161	161	±16
$1,000	283	282	294	318	320	±32

In this example, for low misclassification costs, the best results are obtained for a factor of power between 0.5 and 1.0. In these situations, average costs are much lower when compared with C4.5. For high misclassification costs, compared to test costs, best results occur with factor of power equal to 0.0 (C4.5) and 0.1, but without substantial differences for the other models.

As misclassification costs rise, test costs tend to be negligible. That is, for sufficiently higher misclassification costs and considering equal cost for false negatives and false positives, a higher accuracy rate corresponds to a lower average cost.

In this evaluation we distinguished between individual and group costs, and realized that results considering group costs always had higher average costs. This happened because the serum insulin test, which shares group costs with the glucose tolerance test, only intermittently appeared in the decision trees and, therefore, there was an extra imputed cost nothing or little used.

We also tested the factor of "risk" with other datasets and obtained interesting results. Sometimes, with minimum increase in the total cost it is possible to obtain models that are more patient-friendly.

Conclusions

In this paper we present approaches to build cost-sensitive decision trees, considering different aspects of test costs, where are included economic and non-economical costs. Our framework integrates also a cost-sensitive meta-learner to consider the situations where misclassifications costs are different. Results show that it outperforms the traditional, non cost-sensitive, C4.5.

As technologies became more expensive, it is even more rational to consider all the cost involved. A big challenge is to have better healthcare using less money. The factor of "risk" is an important item of the proposed framework as it can induce models patient-friendly. Decision trees represent a natural representation for classification problems (for diagnosis or prognosis) that includes costs.

In our future work we will continue to evaluate our strategies, with new experiments in other datasets, with real data and real costs. We will try to incorporate other costs associated with delayed tests, emergency situations, staff, facilities, and increased length of stay. The cost function will also be refined and tested for different scenarios. We will also study other methods for building and testing decision trees with the application of a cost-sensitive pruning strategy.

References

[1] Cios KJ (Ed.). Medical Data Mining and Knowledge Discovery. New York: Physica-Verlag, 2001.

[2] Fayyad UM, Piatetsky-Shapiro G, Smyth P, and Uthurusamy R (Eds.). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press 1996.

[3] Witten IH, and Frank E. Data mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] Coiera E. Guide to Health Informatics. A Hodder Arnold Publication; 2nd ed., 2003.

[5] Turney, P. Types of cost in inductive concept learning. Proc. Workshop on Cost-Sensitive Learning, 17th Int. Conf. Machine Learning. 2000; 15–21.

[6] Fawcett T, Provost F. Activity monitoring: Noticing interesting changes in behavior. Proc. 5th Int. Conf. Knowledge Discovery and Data Mining. 1999; 53–62.

[7] Breiman L, Freidman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont, California: Wadsworth. 1984.

[8] Elkan C. The Foundations of Cost-Sensitive Learning. Proc. 17th Int. Joint Conf. Artificial Intelligence. 2001; 973–978.

[9] Núñez M. The use of background knowledge in decision tree induction. Machine learning. 1991; 6:231–250.

[10] Melville P, Provost F, Saar-Tsechansky M, Mooney R. Economical active feature-value acquisition through Expected Utility estimation. Proc. 1st Int. Workshop on Utility-Based Data Mining. 2005; 10–16.

[11] Turney P. Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. J Artificial Intelligence Research. 1995; 2:369–409.

[12] Zubek VB, Dietterich T. Pruning improves heuristic search for cost-sensitive learning. Proc. 19th Int. Conf. Machine Learning. 2002; 27–35.

[13] Greiner R, Grove AJ, Roth D. Learning cost-sensitive active classifiers. Artificial Intelligence. 2002; 139(2):137–174.
[14] Arnt A, Zilberstein S. Attribute Measurement Policies for Cost-effective Classification. Workshop Data Mining in Resource Constrained Environments, 4th Int. Conf. Data Mining. 2004.
[15] Chai X, Deng L, Yang Q, Ling CX. Test-Cost Sensitive Naive Bayes Classification. Proc. 4th Int.Conf. Data Mining. 2004.
[16] Ling CX, Yang Q, Wang J, Zhang S. Decision Trees with Minimal Costs. Proc. 21st Int. Conf. Machine Learning, 2004.
[17] Sheng S, Ling CX, Yang Q. Simple Test Strategies for Cost-Sensitive Decision Trees. Proc. 16th European Conf. Machine Learning. 2005; 365–376.
[18] Sheng S, Ling CX. Hybrid Cost-sensitive Decision Tree. Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). 2005.
[19] Zhang S, Qin Z, Ling CX, Sheng S. "Missing Is Useful": Missing Values in Cost-Sensitive Decision Trees. IEEE Transactions on Knowledge and Data Engineering. 2005 17(12): 1689–1693.
[20] Ling CX, Sheng VS, Yang Q. Test Strategies for Cost-Sensitive Decision Trees. IEEE Transactions on Knowledge and Data Engineering. 2006; 18(8): 1055–1067.
[21] Grobman WA, Stamilio DM. Methods of clinical prediction. American Journal of Obstetrics and Gynecology. 2006; 194(3):888–94.
[22] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann, 1993.
[23] Drummond C, Holte RC. Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria. Proc. 17th Int. Conf. Machine Learning. 2000; 239–246.

Address for correspondence

Alberto Freitas; Dept Biostatistics and Medical Informatics, Faculty of Medicine, University of Porto; Al. Prof. Hernani Monteiro; 4200-319 Porto; Portugal; e-mail: alberto@med.up.pt