Optimal Sub-Reducts with Test Cost Constraint

Fan Min and William Zhu

Lab of Granular Computing,
Zhangzhou Normal University, Zhangzhou 363000, China
minfanphd@163.com, williamfengzhu@gmail.com

Abstract. Cost-sensitive learning extends classical machine learning by considering various types of costs, such as test costs and misclassification costs, of the data. In many applications, there is a test cost constraint due to limited money, time, or other resources. It is necessary to deliberately choose a set of tests to preserve more useful information for classification. To cope with this issue, we define optimal sub-reducts with test cost constraint and a corresponding problem for finding them. The new problem is more general than two existing problems, namely the minimal test cost reduct problem and the 0-1 knapsack problem, therefore it is more challenging than both of them. We propose two exhaustive algorithms to deal with it. One is straightforward, and the other takes advantage of some properties of the problem. The efficiencies of these two algorithms are compared through experiments on the mushroom dataset. Some potential enhancements are also pointed out.

Keywords: Cost-sensitive learning, attribute reduction, test cost, constraint, exhaustive algorithm.

1 Introduction

Cost-sensitive learning has attracted much research interests in the past two decades. Two types of costs, namely test costs and misclassification costs [1], are more often addressed. The test cost is the measurement cost of determining the value of an attribute a exhibited by an object [1,2]. Hence in the context of cost-sensitive learning, an attribute is also called a test.

In some classification problems, there are many available tests, and we would like to remove some of them to save the test cost. An ideal solution is to minimize the test cost, and at the same time, preserve the information of the decision system. Then we can build classifiers which are as good as the ones built on the original decision system. This problem is called the minimal test cost reduct (MTR) problem and has been studied in [3,4]. Unfortunately, the test cost one can afford is limited in many applications; and one has to sacrifice necessary information to keep the test cost under budget. Our problem is: Given a test cost constraint, how to choose test set with which the information is preserved to the highest degree?

This paper proposes the concept of optimal sub-reducts with test cost constraint (OSRT). Since our problem is to find all these sub-reducts, we call it the

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 57–62, 2011.

OSRT problem. The new problem is more general than both the minimal test cost reduct problem [3] and the 0-1 knapsack problem. We propose two exhaustive algorithms to deal with it. The first is obtained from the problem definition directly. The second takes advantage of some properties of the problem, hence it is more efficient than the first one. By our open source software Coser [5], they can solve the OSRT problem of the mushroom dataset, which has 22 tests, in a number of seconds. Therefore they are applicable to datasets with rational sizes of tests. Another important use of them is to evaluate the result quality of heuristic algorithms, which can be employed in large datasets.

2 Preliminaries

Cost-sensitive decision systems are more general than decision systems. This paper considers the simplest though most widely used model, called *test-cost-independent decision systems* (TCI-DS) [4]. It is represented by a decision table and a test cost vector $c = [c(a_1), c(a_2), \ldots, c(a_{|C|})]$. Free tests are not considered.

Attribute reduction has been intensively studied by the rough set society. There are many extensions of the classical rough set model [6], such as covering-based [7,8], decision-theoretical [9], and dominance-based [10] rough set models. A number of definitions of relative reducts exist [11,12]. The definition based on the conditional information entropy is given below.

Definition 1. *[13] Let $S = (U, C, D, V, I)$ be a decision system, and $H(D|B)$ be the conditional entropy of $B \subseteq C$ w.r.t. D. Any $R \subseteq C$ is a Shannon's entropy reduct iff:*

1. $H(D|R) = H(D|C)$, *and*
2. $\forall a \in R, H(D|R - \{a\}) > H(D|C)$.

Sometimes we are interested in test sets without redundant test.

Definition 2. *Let $S = (U, C, D, V, I)$ be a decision system and $R \subseteq C$. R is a sub-reduct iff $\forall a \in R, H(D|R - \{a\}) > H(D|C)$.*

The aim of the classical reduct problem is to find a minimal reduct. When the test cost issue is involved, we are interested in reducts with minimal test costs. Let S be a TCI-DS and $Red(S)$ be the set of all reducts of S. Any $R \in Red(S)$ where $c(R) = \min\{c(R')|R' \in Red(S)\}$ is called a *minimal test cost reduct* [3]. The *minimal test cost reduct* (MTR) problem is more general than the classical reduct problem, which is NP-hard.

The 0-1 knapsack problem appears in textbooks such as data structure, algorithm design and implementation. It is NP-complete.

3 The Optimal Sub-Reducts Problem

Suppose that we are given limited amount of test cost in terms of time, money, etc. We have to sacrifice necessary information to meet the constraint. Naturally, we require that the selected test set has the minimal possible conditional entropy. This consideration brings us to the following definition.

Definition 3. *Let $S = (U, C, D, V, I, c)$ be a TCI-DS and m be the test cost upper bound. The set of all test sets subject to the constraint is*

$$T(S, m) = \{B \subseteq C | c(B) \leq m\}. \tag{1}$$

In $T(S, m)$, the set of all test sets with the minimal conditional entropy is

$$M_T(S, m) = \{B \in T(S, m) | H(D|B) = \min\{H(D|B') | B' \in T(S, m)\}\}. \tag{2}$$

In $M_T(S, m)$, the set of all optimal sub-reducts is

$$P_{M_T}(S, m) = \{B \in M_T(S, m) | c(B) = \min\{c(B') | B' \in M_T(S, m)\}\}. \tag{3}$$

Any element in $P_{M_T}(S, m)$ is called an optimal sub-reduct with test cost constraint, or an optimal sub-reduct for brevity.

In Definition 3, Equation (1) ensures that the test cost constraint is met. This is the basic requirement of our problem. Then Equation (2) ensures that the test set is optimal from the viewpoint of conditional information entropy. This is our primary optimization objective. Finally, Equation (3) ensures that the test cost is also optimized. This is our secondary objective. Without the secondary objective, redundant tests may exist when m is greater than the test cost of a minimal test cost reduct.

We have assumed that no free test exists, Equation (3) also ensures the that there is no redundant test. According to Definition 2, we know that any element in $P_{M_T}(S, m)$ must be a sub-reduct. This is why $P_{M_T}(S, m)$ is called the set of all optimal *sub-reducts*. The problem of constructing $P_{M_T}(S, m)$ is called the *optimal sub-reducts with test cost constraint* (OSRT) problem.

Now we analyze the relationships between the new problem and two problems mentioned in Section 2. When the test cost is enough for a reduct, the OSRT problem coincides with the minimal test cost reduct (MTR) problem [14]. On the other hand, the OSRT problem is very similar to the 0-1 knapsack problem. The key difference lies in that the value of each item is fixed, but the "value" of each test is variable; it depends on other selected tests. Therefore, the OSRT problem is more general, and more difficult than the 0-1 knapsack problem.

4 Exhaustive Algorithms

Due to the complexity of the new problem, exhaustive algorithms are inapplicable to large datasets. They are, however, important from the theoretical viewpoint. They also help to evaluate the performance of a heuristic algorithm, often on small datasets. In the following context we assume that $m < c(R)$ where $R \in MTR(S)$, so that our problem does not coincide with the MTR problem.

Definition 3 has indicated an exhaustive algorithm, called the straightforward exhaustive optimal sub-reduct algorithm (SESRA). It has three steps: Step 1, compute $T(S, m)$; Step 2, compute $M_T(S, m)$; and Step 3, compute $P_{M_T}(S, m)$.

The running time of Step 1 is exponential with respect to $|C|$. Step 2 is the most time consuming, and the running time for Step 3 is neglectful. This claim will be validated in Section 5 through experiments.

Next we discuss how to revise SESRA to provide better performance. The following propositions help to reduce some computation.

Proposition 1. *Let* $\emptyset \subset B \subset B' \subseteq C$, $H(D|B') \leq H(D|B)$.

Proposition 2. *Let* $B' \in T(S,m)$ *and* $\emptyset \subset B \subset B'$. $B \in T(S,m)$.

Proposition 1 indicates that to compute $\min\{H(D|B)|B \in T(S,m)\}$, we do not have to check every element in $T(S,m)$. Any element which is a subset of another element should be removed, and the subset to be checked is

$$T'(S,m) = T(S,m) - \{B \in T(S,m)|\exists B' \in T(S,m) st. B \subset B')\}. \quad (4)$$

Proposition 2 indicates that many elements of $T(S,m)$ may be removed, and $|T'(S,m)|$ may be significantly smaller than $|T(S,m)|$.

According to Proposition 1 and Equation (4),

$$\min\{H(D|B)|B \in T'(S,m)\} = \min\{H(D|B)|B \in T(S,m)\}. \quad (5)$$

Similar to Equation (2), let

$$M'_T(S,m) = \{B \in T'(S,m)|H(D|B) = \min\{H(D|B')|B' \in T'(S,m)\}\}. \quad (6)$$

We know that $M'_T(S,m) = M_T(S,m) \cap T'(S,m) \neq \emptyset$. Therefore $M'_T(S,m)$ always contains some test sets with the minimal conditional entropy. In most cases, however, not all test sets with the minimal conditional entropy are included in $M'_T(S,m)$. That is, $M_T(S,m) \not\subseteq M'_T(S,m)$. To make the matter worse, $P_{M_T}(S,m) \not\subseteq M'_T(S,m)$. Similar to Equation (3), let

$$P'_{M_T}(S,m) = \{B \in M'_T(S,m)|c(B) = \min\{c(B')|B' \in M'_T(S,m)\}\}. \quad (7)$$

We have $P'_{M_T}(S,m) \neq P_{M_T}(S,m)$, which indicates that we may miss optimal sub-reducts by discarding some test sets as indicated by Equation (4). The reason lies in that an element in $P_{M_T}(S,m)$ is included in $P'_{M_T}(S,m)$ only if no superset of it meets the constraint.

Fortunately, we have the following propositions to amend this flaw.

Proposition 3. $\forall B \in M_T(S,m)$, $\exists B' \in M'_T(S,m)$ *such that* $B \subseteq B'$.

Proof. Because $B \in M_T(S,m) \subseteq T(S,m)$, $\exists B' \in T'(S,m)$ such that $B \subseteq B'$. On one hand, according to Proposition 1, $H(D|B') \leq H(D|B)$. On the other hand, according to Equation (2) and $B' \in T'(S,m) \subseteq T(S,m)$, $H(D|B) \leq H(D|B')$. Therefore $H(D|B') = H(D|B)$. Equation (6) assures that $B' \in M'_T(S,m)$. This completes the proof.

The following proposition provides an approach to compute a superset of the set of all optimal sub-reducts.

Algorithm 1. The SESRA* algorithm

Input: $S = (U, C, D, V, I, c)$, m
Output: $P_{M_T}(S, m)$, the set of all optimal sub-reducts
Method: SESRA-star

1: Construct test sets and at the same time, obtain $T(S, m)$;
2: Remove elements from $T(S, m)$ and obtain $T'(S, m)$;
3: Select elements with the minimal conditional entropy and obtain $M'_T(S, m)$;
4: Compute $M''_T(S, m)$ using the exhaustive attribute reduction algorithm;
5: Select elements with the minimal test cost and obtain $P_{M_T}(S, m)$;

Proposition 4. *Let $Red_M(S)$ be the set of all minimal test cost reducts of S and*

$$M''_T(S, m) = \bigcup_{B' \in M'_T(S,m)} Red_M(U, B', D, V, I, c). \tag{8}$$

$$P_{M_T}(S, m) \subseteq M''_T(S, m). \tag{9}$$

Proof. For any $B \in P_{M_T}(S, m) \subseteq M_T(S, m)$, according to Proposition 3, $\exists B' \in M'_T(S, m)$ such that $B \subseteq B'$. On the other hand, B has the minimal test cost, therefore $B \in Red_M(U, B', D, V, I, c)$. This completes the proof.

According to above analysis, we obtain a new algorithm as listed in Algorithm 1. Step 2 through 4 of SESRA* correspond to Step 2 of SESRA.

5 Experiments

The main purpose of our experiments is to compare the performances of SESRA and SESRA*, which are implemented in Coser [5]. Experiments are undertaken on the mushroom dataset, where $|U| = 8124$ and $|C| = 22$. Parameter settings are as follows: Test costs are random numbers in [1..100]. $m = 0.8 \times c(R^*)$ where R^* is a minimal test cost reduct. Results are digested in Table 1.

Table 1. Results on the mushroom dataset (mean values for 100 test cost settings)

		SESRA	SESRA*
Set size	$T(S, m)$	961.01	961.01
	$T'(S, m)$	-	254.31
	$M_T(S, m)$	1.34	-
	$M'_T(S, m)$	-	1.04
	$M''_T(S, m)$	-	1.04
	$P_{M_T}(S, m)$	1.01	1.01
Run time (ms)	Candidates building	1685.73	1680.95
	Consistency computing	11425.12	2391.85
	Total	13110.85	4072.80

"-" stands for inapplicable.

Table 1 showed that the number of test sets with conditional entropy checked is reduced from 961.01 to 254.31, about 1/4 of the initial value. Consequently, the time for respective step is reduced from 11425.12 ms to 2391.85 ms, a little more than 1/5 of the initial value. Finally, the total time is reduced to about 1/3. In general, the improvement is significant.

6 Conclusions and Further Works

Exhaustive algorithms are undoubtedly the right choice for datasets with rational sizes. In this paper, we proposed the OSRT problem and two exhaustive algorithms to deal with it. SESRA* is about 2 times faster than SESRA in our experiments. In the future we will revise SESRA* to support bigger datasets. We will also develop heuristic algorithms for large datasets.

Acknowledgements. This work is in part supported by National Science Foundation of China under Grant No. 60873077/F020107.

References

1. Hunt, E.B., Marin, J., Stone, P.J. (eds.): Experiments in induction. Academic Press, New York (1966)
2. Turney, P.D.: Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of Artificial Intelligence Research 2, 369–409 (1995)
3. Min, F., He, H., Qian, Y., Zhu, W.: Test-cost-sensitive attribute reduction. In: To Appear in Information Sciences (2011)
4. Min, F., Liu, Q.: A hierarchical model for test-cost-sensitive decision systems. Information Sciences 179(14), 2442–2452 (2009)
5. Min, F., Zhu, W.: Coser: Cost-senstive rough sets (2011), `http://grc.fjzs.edu.cn/~fmin/coser/index.html`
6. Pawlak, Z.: Rough sets and intelligent data analysis. Information Sciences 147(12), 1–12 (2002)
7. Zhu, W.: Topological approaches to covering rough sets. Information Sciences 177(6), 1499–1508 (2007)
8. Zhu, W., Wang, F.: Reduction and axiomization of covering generalized rough sets. Information Sciences 152(1), 217–230 (2003)
9. Yao, Y., Zhao, Y.: Attribute reduction in decision-theoretic rough set models. Information Sciences 178(17), 3356–3373 (2008)
10. Greco, S., Matarazzo, B., Słowiński, R., Stefanowski, J.: Variable consistency model of dominance-based rough sets approach. In: Ziarko, W.P., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 170–181. Springer, Heidelberg (2001)
11. Hu, Q., Yu, D., Liu, J., Wu, C.: Neighborhood rough set based heterogeneous feature subset selection. Information Sciences 178(18), 3577–3594 (2008)
12. Qian, Y., Liang, J., Pedrycz, W., Dang, C.: Positive approximation: An accelerator for attribute reduction in rough set theory. Artificial Intelligence 174(9-10), 597–618 (2010)
13. Slezak, D.: Approximate entropy reducts. Fundamenta Informaticae 53(3-4), 365–390 (2002)
14. Min, F., Zhu, W.: Attribute reduction with test cost constraint. Journal of Electronic Science and Technology of China 9(2) (June 2011)