Minimal cost feature selection of data with normal distribution measurement errors

Hong Zhao, Fan Min\(^*\), William Zhu

Lab of Granular Computing, Zhangzhou Normal University, Zhangzhou 363000, China

Abstract

Minimal cost feature selection is devoted to obtain a trade-off between test costs and misclassification costs. This issue has been addressed recently on nominal data. In this paper, we consider numerical data with measurement errors and study minimal cost feature selection in this model. First, we build a data model with normal distribution measurement errors. Second, the neighborhood of each data item is constructed through the confidence interval. Comparing with discretized intervals, neighborhoods are more reasonable to maintain the information of data. Third, we define a new minimal total cost feature selection problem through considering the trade-off between test costs and misclassification costs. Fourth, we proposed a backtracking algorithm with three effective pruning techniques to deal with this problem. The algorithm is tested on four UCI data sets. Experimental results indicate that the pruning techniques are effective, and the algorithm is efficient for data sets with nearly one thousand objects.

Keywords: feature selection; normal distribution measurement errors; test costs; misclassification costs.

1. Introduction

Minimal cost feature selection is devoted to obtain a trade-off between test costs and misclassification costs. Test costs are what we pay for collecting data items \(^1\). Test costs are often measured by time, money, and other resources. When test costs are only considered (see, e.g., \(^2\) \(^3\) \(^4\) \(^5\)), the minimal cost feature selection problem degrades to the minimal test cost reduct problem \(^1\). Therefore, the minimal cost feature selection problem is a generalization of the minimal test cost reduct problem.

In addition to the test costs, misclassification costs are necessary to be considered in cost-sensitive learning \(^6\). Misclassification cost is the penalty we
receive while deciding that an object belongs to class J when its real class is K [7, 8]. For example, in medical diagnosis, if a cancer (non-cancer) is regarded as the negative (positive) class, there will be punishment. When misclassification costs are only considered (see, e.g., [8, 9]), the average misclassification cost is a more general metric than the accuracy [10].

It is important to consider both test costs and misclassification costs in many applications [5]. Test costs are paid on each object. While misclassification costs are paid to misclassified objects. Therefore we should take into account total cost on a set of objects through considering the trade-off between test costs and misclassification costs. This issue has been addressed recently on nominal data. In real applications, however, data are often numeric and they always have some measurement errors. The measurement errors of the data have certain universality and are inescapability.

In this paper, we study minimal cost feature selection considering numerical data with measurement errors. First, a data model with measurement errors under the normal distribution is defined. Second, we construct the neighborhood of each data item through the confidence interval. Compared with discretized intervals, neighborhoods are more reasonable to maintain the information of data. Third, considering the trade-off between test costs and misclassification costs, we define a new minimal total cost feature selection problem. Fourth, a backtracking algorithm with three effective pruning techniques is proposed to deal with this problem.

Four open data sets from the UCI (University of California - Irvine) library are employed to study the efficiency and effect of our algorithm. Experiments are undertaken with open source software Coser [11] to validate the performance of this algorithm. Experimental results indicate that the pruning techniques are effective, and the algorithm is efficient for data sets with nearly one thousand objects.

This paper is organized as follows. Section 2 introduces a decision system with normal distribution measurement errors. Then we introduce test costs and misclassification costs to this decision system, and define a cost-sensitive decision system. In Section 3 we define a new minimal cost feature selection problem by considering test costs and misclassification costs. In order to deal with this problem, a full description of a backtracking algorithm is given in Section 4. Experimental settings and results are discussed in Section 5. Finally, Section 6 concludes and suggests further research trends.

2. Data models

In this section, the concept of decision systems with normal distribution measurement errors (NDME) is revisited. Then the neighborhood of each data item is constructed through the confidence interval. Finally decision systems based on NDME with test costs and misclassification costs is presented.
Definition 1. [12] A decision system with normal distribution measurement errors (NEDS) $S$ is the 6-tuple:

$$S = (U, C, d, V = \{V_a | a \in C \cup \{d\}\}, I = \{I_a | a \in C \cup \{d\}\}, n),$$

where $U$ is the nonempty set called a universe, $C$ is the nonempty set of conditional attributes, $d$ is the decision attribute, $V_a$ is the set of values for each $a \in C \cup \{d\}$, and $I_a : U \rightarrow V_a$ is an information function for each $a \in C \cup \{d\}$. $n : C \rightarrow \mathbb{R}^+ \cup \{0\}$ is the maximal measurement error of $a \in C$, and $\pm n(a)$ are confidence limits of $a$, respectively.

Confidence limits are the lower and upper boundaries of a confidence interval. In real applications, there are a number of measurement methods to obtain a numerical data item with different measurement errors. The measurement errors often satisfy normal distribution which is found to be applicable over almost the whole of science and engineering measurement. We introduce the confidence interval of normal distribution to our model. With Definition 1, a new neighborhood is defined as follows.

Definition 2. [12] Let $S = (U, C, d, V, I, n)$ be a NEDS. Given $x_i \in U$ and $B \subseteq C$, the neighborhood of $x_i$ with respect to normal distribution measurement errors on test set $B$ is defined as

$$n_B(x_i) = \{x \in U | \forall a \in B, |a(x) - a(x_i)| \leq 2n(a)\},$$

it represents the error value of $a$ in $[-n(a), +n(a)]$.

From Definition 2 we know that

$$n_B(x_i) = \bigcap_{a \in B} n_{\{a\}}(x_i).$$

That is, the neighborhood $n_B(x_i)$ is the intersection of a number of basic neighborhoods (see, e.g., [13, 14, 15, 16]). Given $\forall x \in U$, $\forall a \in B$, $x \in n_B(x)$. Therefore, for any $B \subseteq C$, $\bigcup_{x \in U} n_B(x) = U$. Hence the set $\{n_B(x_i) | x_i \in U\}$ is a covering (see, e.g., [17, 18, 19]) of $U$.

In cost-sensitive learning, test costs and misclassification costs are two most important types of costs. Now, we define a cost-sensitive decision system by considering both test and misclassification costs.

Definition 3. A decision system based on NDME with test costs and misclassification costs (NEDS-TM) $S$ is the 8-tuple:

$$S = (U, C, d, V, I, n, tc, mc),$$

where $U, C, d, V, I$ and $n$ have the same meanings as in Definition 1. $tc : C \rightarrow \mathbb{R}^+ \cup \{0\}$ is the test costs function and $mc : k \times k \rightarrow \mathbb{R}^+ \cup \{0\}$ is the misclassification costs function, where $k = |I_d|$. 

3
For any $B \subseteq C$, the sequence-independent test costs function $tc$ is defined as follows:

$$tc(B) = \sum_{a \in B} tc(a). \quad (5)$$

The misclassification costs function can be represented by a matrix $mc = \{mc_{k \times k}\}$. Misclassification costs [20] [21] [22] is the penalty we receive while deciding that an object belongs to class $m$ when its real class is $n$ [7] [23]. If classification is correct, the misclassification costs $mc[m, m] = 0$. The following example gives us intuitive understanding.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0.31</td>
<td>0.23</td>
<td>0.08</td>
<td>y</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.14</td>
<td>0.38</td>
<td>0.23</td>
<td>y</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.25</td>
<td>0.40</td>
<td>0.40</td>
<td>y</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.60</td>
<td>0.46</td>
<td>0.51</td>
<td>n</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.41</td>
<td>0.64</td>
<td>0.62</td>
<td>n</td>
</tr>
<tr>
<td>$x_6$</td>
<td>0.35</td>
<td>0.50</td>
<td>0.75</td>
<td>n</td>
</tr>
</tbody>
</table>

Table 2: A neighborhood boundaries vector and a test costs vector.

<table>
<thead>
<tr>
<th>$a$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n(a)$</td>
<td>0.0069</td>
<td>0.0087</td>
<td>0.0086</td>
</tr>
<tr>
<td>$c(a)$</td>
<td>$28$</td>
<td>$19$</td>
<td>$56$</td>
</tr>
</tbody>
</table>

Example 4. A NEDS-TM is given by Tables 1 and 2 and

$$mc = \begin{bmatrix} 0 & 800 \\ 200 & 0 \end{bmatrix}. \quad (6)$$

That is, the test costs are $28$, $19$ and $56$, respectively. In this data set, the decision attribute is used to split data set into two sets. In case that a person belongs to set “$y$”, and he is misclassified as set “$n$”, a penalty of $800$ is paid, contrarily $200$ is paid.

3. Minimal cost feature selection problem

In this work, we focus on selecting a subset of features to minimize the total cost, that is, cost-sensitive feature selection based on test costs and misclassification costs. The problem of finding such a subset of features is called the minimal cost feature selection problem.
Problem 5. The minimal cost feature selection problem (MCFS).

Input: $S = (U, C, d, V, I, n, tc, mc)$;
Output: $R \subseteq C$ and the average total cost (ATC);
Optimization objective: \( \min |ATC(R)| \).

Compared with the classical minimal reduction problem, there are several differences as follows. The first is the input, the external information are the test costs and misclassification costs as well as normal distribution measurement errors. The second is the optimization objective, which is to minimize the total cost instead of the number of features. In data mining applications, average total cost (ATC) is considered to be a more general metric than the accuracy [9]. The average total cost is computed as follows.

**Step 1. Compute the neighborhood of each data item.** Let $B$ be a selected feature set, and $N(B)$ be a set $\{n_B(x_i) | x_i \in U\}$.

**Step 2. Assign one class.** Let $U' = n_B(x_i)$ and $c_B(x_i)(d_m)$ be the number of class $d_m$ of $U'$, where $d_m \in \{I_d\}$. We adjust different classes of elements in $U'$ to one class for minimizing the misclassification cost of $U'$. $U'$ includes the following two types of cases.

1. $U' \subseteq POS_B(\{d\})$ if and only if $d(x) = d(y)$ for any $x, y \in U'$. In this case, the misclassification costs of $U'$ is $mc(U', B) = 0$. For any $x \in U'$, the assigning class $d'(x) = d(x)$.

2. However, if there exists $x, y \in U'$ st. $d(x) \neq d(y)$, we may adjust different classes of elements in $U'$ to one class. Let $d(x)$ be $m$-class and $d(y)$ be $n$-class. We select one case to minimize the misclassification costs of $U'$.

$$mc(U', B) = \min(mcm_{m,n} \times |U'_m|, mcm_{n,m} \times |U'_n|).$$

For any $x \in U'$,

$$d'(x) = \begin{cases} n\text{-class} & \text{if } mc(U', B) = mc_{m,n} \times |U'_m|, \\ m\text{-class} & \text{if } mc(U', B) = mc_{n,m} \times |U'_n|, \end{cases}$$

where $mc_{m,n}$ is the cost of classifying an object of the $m$-class into the $n$-class, and $|U'_m|$ is the number of $m$-class.

**Step 3. Compute average misclassification cost.** The class of $x_i$ is $d'(x_i) = d_m$ if and only if $\max\{c_B(x_i)(d_m) | d_m \in \{I_d\}\}$. The misclassification costs are

$$mc^*(d(x_i), d'(x_i)) = \begin{cases} 0 & \text{if } d(x_i) = d'(x_i), \\ mc_{m,n} & \text{if } d(x_i) = m \text{ and } d'(x_i) = n, \\ mc_{n,m} & \text{if } d(x_i) = n \text{ and } d'(x_i) = m. \end{cases}$$

In this way, the average misclassification cost is given by

$$\overline{mc}(U, B) = \frac{\sum_{x_i \in U} mc^*(d(x_i), d'(x_i))}{|U|}.$$  \hspace{1cm} (10)

**Step 4. Compute average total cost (ATC).** The average total cost is given by

$$ATC(U, B) = tc(B) + \overline{mc}(U, B).$$  \hspace{1cm} (11)
In this context, we select a feature subset in order to minimize the average total cost. The minimal average total cost is given by

\[ ATC(U, B) = \min \{ ATC(U, B') | B' \subseteq C \} \]  (12)

The MCFS problem has a similar motivation with the cheapest cost problem [24], or the minimal test cost reduct (MTR) problem (see, e.g., [1]). However, compared with the MTR problem, our MCFS problem is different from theirs in two aspects.

1. In addition to considering the test costs of each attribute, we take misclassification cost into account. When the misclassification costs are too large compared with test costs, the MCFS problem coincides with the MTR problem. Therefore the MCFS problem is a generalization of the MTR problem.

2. The attribute reduction needs to preserve a particular property of the decision system. The feature selection relies on the only cost information.

4. Algorithm

In this section, the backtracking algorithm to the minimal cost feature selection problem (MCFS) is illustrated in Algorithm 1. To invoke the algorithm, one should initialize the global variables as follows: \( R = \emptyset \) is a feature subset with minimal total cost; \( cmc = mc(U, R) \) is currently minimal cost; and use the following statement: backtrack\((R, 0)\). The result of a feature subset with minimal total cost will be stored in \( R \).

Generally, the search space of the feature selection algorithm is \( 2^{|C|} \). In this context, a backtracking algorithm with pruning techniques is used to select a feature subset to minimize the total cost. There are essentially three pruning techniques employed in Algorithm 1:

1. In Algorithm 1, Line 1 indicates that the variable \( i \) starts from \( il \) instead of 0. Whenever we move forward (see Line 14), the lower bound is increased. With this pruning technique, the solution space is \( 2^{|C|} \) instead of \(|C|!\).

2. In Algorithm 1, Lines 3 through 5 show the second pruning method. The misclassification costs are non-negative in the practical application. In this conditions, the feature subsets \( B \) will be discarded if the test costs of \( B \) is larger than the current minimal cost (\( cmc \)). This technique can prune most branches.

3. Lines 6 through 8 indicate that if the new feature subset produce high cost, the current branch will never produce the feature subset with minimal total cost.

5. Experiments

Experiments are undertaken on four data sets from the UCI Repository of Machine Learning Databases, as listed in Table 3. We undertake three groups of experiments from different viewpoints.
# Algorithm 1 A backtracking algorithm to the MCFS problem

**Input:** \((U, C, d, V, I, n, tc, mc)\), select tests \(R\), current level test index lower bound \(l\)

**Output:** A set of features \(R\) with minimal total cost and AMC, they are global variables

**Method:** backtracking

1. for \((i = l; i < |C|; i++)\) do
2. \(B = R \cup \{a_i\}\)
3. if \((tc(B) > cmc)\) then
4. continue; //Pruning for too expensive test costs
5. end if
6. if \((ATC(U, B) \geq ATC(U, R))\) then
7. continue; //Pruning for non-decreasing total cost
8. end if
9. if \((ATC(U, B) < cmc)\) then
10. \(cmc = ATC(U, B);\) //Update the minimal total cost
11. \(R = B;\) //Update the set of features with minimal total cost
12. end if
13. end for
14. backtrack\((R, i + 1)\);

---

**Table 3: Data sets information.**

| No. | Name     | Domain    | \(|U|\) | \(|C|\) | \(D = \{d\}\) |
|-----|----------|-----------|--------|--------|----------------|
| 1   | Liver    | clinic    | 345    | 6      | selector       |
| 2   | Credit   | commerce  | 690    | 15     | class          |
| 3   | Iono     | physics   | 351    | 34     | class          |
| 4   | Diab     | clinic    | 768    | 8      | class          |

From the viewpoint of different misclassification costs settings, we undertake two sets of experiments. First, we assume that the misclassification costs are different from each other. Table 4 is the optimal feature subset based on different misclassification costs for Diab data set. The ratio of two misclassification costs is set 10 in this experiment. As shown in this table, when the misclassification costs are too large compared with test costs, the test costs increase and even equal to the average total cost. In this case, the MCFS problem coincides with the MTR problem. Second, misclassification costs are identical for different misclassification. Table 5 shows the optimal feature subset based on unified misclassification cost. Since the misclassification costs are small enough, the algorithm chooses a feature subset with minimal average total cost.

From the efficiency of the algorithm, we design two sets of experiments to evaluate performance of pruning technique. In this set of experiments, misclassification costs are identical for different misclassification. In this case, we
propose Fast approach and Slow approach. The Fast approach is a backtracking algorithm with three pruning methods, which is given by the Algorithm 1. We also propose the Slow approach without the third pruning method. The first set studies the change of backtracking steps with the misclassification costs. Figure 1 shows the backtracking steps of the algorithm. The second set studies the change of run time with the misclassification costs. Figure 2 shows the run time. From these two figures we observe the effectiveness of the third pruning technique.

From the costs viewpoint, the changes of test costs and the average minimal total cost are shown in Figure 3. In real world, we could not select expensive tests when misclassification costs are low. Figure 3 shows this situation clearly.

6. Conclusion and further works

In this paper, we take data with normal distribution measurement errors into account and study feature selection with minimal cost. This new feature selection problem, called minimal cost feature selection (MCFS), has a wide
application area for two reasons. From the viewpoint of the data, measurement errors under considered are ubiquitous. From the viewpoint of the minimal cost problem, the resource one can afford is often limited. In order to obtain the optimal result, a backtracking algorithm with three effective pruning techniques is designed for MCFS problem. Experimental results show that the pruning techniques are effective. This work also serves as the benchmark for other heuristic algorithms which should be designed in our further works for large data sets.

Acknowledgment

This work is in part supported by the Fujian Province Foundation of Higher Education (No. JK2012028), the National Nature Science Foundation of China (No. 61170128), and the Natural Science Foundation of Fujian Province, China (Nos. 2011J01374, 2012J01294).
Figure 2: Run time: (a) Liver; (b) Credit; (c) Iono; (d) Diab

References


Figure 3: Test costs and average total cost: (a) Liver; (b) Credit; (c) Iono; (d) Diab


