International Journal of Innovative
Computing, Information and Control
Volume **16**, Number **6**, December **2020**

ICIC International ©2020 ISSN 1349-4198
pp. **2047–2061**

VARIABLE COSTS-BASED MULTI-GRANULARITY FEATURE SELECTION WITH TEST COST CONSTRAINT

Shujiao Liao[1,*] and Yidong Lin[1,2]

[1]School of Mathematics and Statistics
Minnan Normal University
No. 36, Xianqianzhi Street, Xiangcheng District, Zhangzhou 363000, P. R. China
*Corresponding author: sjliao2011@163.com

[2]School of Mathematical Sciences
Xiamen University
No. 422, Siming South Road, Siming District, Xiamen 361005, P. R. China

Received March 2020; revised August 2020

Abstract. *In recent years, some researchers have studied the cost-sensitive feature selection problem based on the rough set theory. Especially, in view of the variability of test costs and misclassification costs as well as the diversity of feature-value granularities between different features, recently an effective cost-sensitive multi-granularity feature selection approach has been proposed. Nevertheless, the approach does not consider the case where a test cost constraint occurs because of limited resources. To tackle this problem, in this paper a variable costs-based multi-granularity feature selection approach is presented in consideration of test cost constraint. First, based on the theoretic framework called confidence-level-vector-based neighborhood rough set, the test cost-constrained multi-granularity feature selection problem is formally defined. Then a heuristic feature-granularity selection algorithm is designed, by which desirable features and their respective feature-value granularities can be simultaneously selected to minimize the total cost consumed in the test cost-constrained situation. Detailed experiments undertaken on six UCI (University of California – Irvine) datasets validate the effectiveness of the algorithm.*
Keywords: Feature-granularity selection, Multi-granularity, Neighborhood rough set, Test cost constraint, Variable costs

1. **Introduction.** Since feature selection can remove redundant or irrelevant features from high dimensional dataset so as to reduce the computational complexity of data processing, it is an important issue in data mining and machine learning [1, 2, 3, 4]. Besides, as an important branch of data mining and granular computing, rough set theory does well in dealing with the uncertainty and granulation of data [5, 6, 7, 8]. Their relation is that feature selection is a key task in the rough set domain. In the rough set community, feature selection is also called attribute reduction, and the selected feature subset is also called a reduct [9, 10].

Cost-sensitive learning is a challenging problem in data mining and machine learning [11, 12]. In practical applications there are various kinds of cost, among which test cost and misclassification cost are the most frequently considered [13]. Test cost, also named feature cost, is the money, time, or other resources consumed in obtaining a data item of an object, while misclassification cost refers to the penalty paid for the wrong classification of an object. In recent years, cost-sensitive learning methods have been studied in the rough set domain, among which the most common issue is cost-sensitive feature selection, also called cost-sensitive attribute reduction [14, 15, 16, 17, 18]. Cost-sensitive feature

DOI: 10.24507/ijicic.16.06.2047

selection aims at selecting a desirable feature subset to minimize some types of cost and meanwhile to keep certain information of the original data.

Although many cost-sensitive feature selection approaches have been proposed, most of them do not touch the feature-value granularity and the cost variability. In fact, on the one hand, measurement errors exist widely in real applications. For a given feature, a high error range of feature values, or equivalently a high feature-value granularity, will induce a low data precision. To obtain a rational data precision for a selected feature, the feature-value granularity should be carefully chosen. On the other hand, costs are not always constant. For example, acquiring fine-grained data items is often more expensive than acquiring coarse-grained ones; thus the test cost of a feature is often monotonically increasing with the diminution of the feature-value granularity. To address these challenges, in recent years some single-granularity feature selection approaches [19, 20, 21] and a multi-granularity feature selection approach [22], in which "granularity" refers to the feature-value granularity, have been presented based on measurement errors and variable costs. In comparison, the single-granularity approaches suppose all features have the same feature-value granularity; thus they are not feasible in some real-world scenarios; while the multi-granularity approach takes account of the granularity diversity between different features, namely different features may have different feature-value granularities, so it is more versatile and practical than the single-granularity approaches. However, the test cost constraint problem has not been studied in the multi-granularity approach.

In some data mining applications there is a test cost constraint on account of limited money, time, or other types of resources. To make the test costs under a budget, one usually has to sacrifice some properties that can be obtained in the test cost-unconstrained case; thus the results of feature selection are often different between the test cost-constrained case and the unconstrained case. A test cost-constrained feature selection approach has been presented based on variable costs [20], but it is single-granularity but not multi-granularity. To address this challenge, in this paper we study the variable costs-based multi-granularity feature selection approach with test cost constraint. It is notable that, to solve the test cost-unconstrained multi-granularity feature selection problem, in [22] a confidence-level-vector-based neighborhood rough set model was constructed, and multiple types of variable cost setting were discussed. In this study, we first give the definition of the test cost-constrained multi-granularity feature selection problem based on the existing theoretic framework. Then a weighted heuristic feature-granularity selection (the selection of features and their respective feature-value granularities) algorithm is designed to solve the new problem. The algorithm can select a desirable feature-granularity pair (the pair of features and feature-value granularities) to minimize the total cost consumed under a rational test cost constraint. And the algorithm is efficient because three accelerating techniques are used to speed up the procedure. Experiments undertaken on six UCI datasets validate the effectiveness of the proposed algorithm. Moreover, the influences of the test cost upper bound value to the feature-granularity selection result are also discussed through experiments. In general, since the proposed multi-granularity feature selection approach can simultaneously select features and their respective feature-value granularities in consideration of multiple realistic factors, including measurement errors, variable costs and test cost constraint, it is rather practical.

The rest of the paper is organized as follows. In Section 2, we first review the theory about the variable costs-based multi-granularity feature selection without test cost constraint. Based on this, then the new problem which takes account of the test cost constraint is defined formally. Section 3 proposes the heuristic feature-granularity selection algorithm and discusses the procedure of the algorithm. Experimental results are

introduced and analyzed in Section 4. Finally, the paper ends with conclusions in Section 5.

2. **Problem Definition.** This section starts from reviewing the theory with respect to the variable costs-based multi-granularity feature selection without test cost constraint, including the confidence-level-vector-based neighborhood rough set (CVRS), variable cost settings, and the calculation method of average total cost [22]. Based on these, we define formally the test cost-constrained multi-granularity feature selection problem.

2.1. **Confidence-level-vector-based neighborhood rough set.** In the existing multi-granularity feature selection approach [22], for a given feature, the feature-value granularity is evaluated by the confidence level of the feature values' measurement errors. The measurement errors are assumed to satisfy a normal distribution, and the confidence level refers to the frequency that an observed interval contains a specific error value. Based on these, the CVRS model was constructed in [22]. Some important concepts and properties in the model are reviewed below.

Definition 2.1. *A confidence-level-vector-based decision system (CVDS) S is the 6-tuple:*

$$S = \left(U, C, d, V = \{V_a | a \in C \cup \{d\}\}, I = \{I_a | a \in C \cup \{d\}\}, P = \left(p_{a_1}, p_{a_2}, \ldots, p_{a_{|C|}}\right)\right),$$

where U is a finite nonempty set of objects called the universe, C is the set of conditional attributes (also called as features), d is the decision attribute, V_a is the set of values for each $a \in C \cup \{d\}$, I_a: $U \to V_a$ is an information function for each $a \in C \cup \{d\}$, and $p_{a_i} \in (0, 0.997]$ is the error confidence level for the feature values of feature a_i.

An exemplary CVDS is composed of the decision system shown in Table 1 and the confidence level vector shown in Table 2. It is known from the two tables that $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $C = \{a_1, a_2, a_3\}$, $P = (0.5, 0.8, 0.6)$.

TABLE 1. An example of numeric decision system

	a_1	a_2	a_3	d
x_1	0.71	0.31	0.21	1
x_2	0.61	0.34	0.11	1
x_3	0.58	0.27	0.25	1
x_4	0.82	0.29	0.23	2
x_5	0.68	0.44	0.55	2
x_6	0.55	0.38	0.05	2

TABLE 2. An example of confidence level vector

a	a_1	a_2	a_3
p_a	0.5	0.8	0.6

Let $e(a, p_a)$ denote the upper error bound w.r.t. feature a and its corresponding confidence level p_a, and then according to the statistical knowledge [23, 24], one has that

$$e(a, p_a) = \sigma_a z_{p_a}, \quad p_a \in (0, 0.997], \tag{1}$$

where z_{p_a} is the quantile value of p_a w.r.t. the standard normal distribution, and σ_a is the standard deviation which is set to be

$$\sigma_a = k \cdot \max \left| a(x_i) - \overline{a(x)} \right|, \quad 1 \le i \le |U|, \tag{2}$$

where $k > 0$ is a constant, $a(x_i)$ is the feature value of object x_i w.r.t. feature a, and $\overline{a(x)} = \frac{1}{|U|}\sum_{i=1}^{|U|} a(x_i)$ is the average feature value of feature a for all objects in U.

Neighborhoods play an important role in the CVRS model. The neighborhood w.r.t. a single feature and its corresponding confidence level is defined as follows.

Definition 2.2. *Let $S = (U, C, d, V, I, P)$ be a CVDS, $x \in U$ and $a \in C$. The neighborhood of x with reference to feature a and confidence level p_a is defined as*

$$n_{(a,p_a)}(x) = \{x' \in U | \, |a(x') - a(x)| \leq 2e(a, p_a)\}, \quad p_a \in (0, 0.997].$$

The neighborhood of $x \in U$ induced by B and P_B is the intersection of the neighborhoods induced by each feature $a \in B$ and its corresponding confidence level p_a, i.e., the confidence-level-vector-based neighborhood is

$$n_{(B,P_B)}(x) = \bigcap_{a \in B} n_{(a,p_a)}(x).$$

Based on B and P_B, a neighborhood relation $R_{(B,P_B)}$ on the universe U can be induced. It can be written as a relation matrix $M\left(R_{(B,P_B)}\right) = (r_{ij})_{|U|\times|U|}$, where $r_{ij} = 1$ if $x_j \in n_{(B,P_B)}(x_i)$, or equivalently $x_i \in n_{(B,P_B)}(x_j)$, otherwise $r_{ij} = 0$. Then $< U, R_{(B,P_B)} >$ is called a neighborhood approximation space.

Lower approximation and positive region are fundamental issues in the rough set theory.

Definition 2.3. *Let $S = (U, C, d, V, I, P)$ be a CVDS, and let $U/\{d\}$ denote the partitions of the universe U induced by the decision attribute d. Suppose that $B \subseteq C$ and P_B is the corresponding confidence level subvector. Then for any $X \in U/\{d\}$, the lower approximation of X in the neighborhood approximation space $< U, R_{(B,P_B)} >$ is defined as*

$$\underline{N_{(B,P_B)}}(X) = \{x \in U | n_{(B,P_B)}(x) \subseteq X\}.$$

Obviously, $\underline{N_{(B,P_B)}}(X) \subseteq X$.

Definition 2.4. *Let $S = (U, C, d, V, I, P)$ be a CVDS, $B \subseteq C$, and P_B be the corresponding confidence level subvector. Suppose that $U/\{d\} = \{X_1, X_2, \ldots, X_K\}$, where X_i is the object subset with decision class i. Then the lower approximation of decision $\{d\}$ in the neighborhood approximation space $< U, R_{(B,P_B)} >$ is defined as*

$$\underline{N_{(B,P_B)}}(\{d\}) = \bigcup_{i=1}^{K} \underline{N_{(B,P_B)}}(X_i).$$

The lower approximation $\underline{N_{(B,P_B)}}(\{d\})$ is also called as the positive region and denoted by $POS_{(B,P_B)}(\{d\})$. It is easy to know that $POS_{(\emptyset,P_\emptyset)}(\{d\}) = \emptyset$. For each object in the positive region, its neighborhood granule consistently belongs to one of the decision classes, so the objects in the positive region can be certainly classified into one class. Many feature selection approaches aim to make the positive region as large as possible.

There are two main types of monotonicity for the fundamental concepts shown above. The first type of monotonicity relates to the addition of features.

Theorem 2.1. *(Type-1 monotonicity). Let $S = (U, C, d, V, I, P)$ be a CVDS, $B_1 \subseteq B_2 \subseteq C$, $P_{B_1} \sqsubseteq P_{B_2}$. We have*

1) $\forall x \in U$, $n_{(B_1,P_{B_1})}(x) \supseteq n_{(B_2,P_{B_2})}(x)$;

2) $R_{(B_1,P_{B_1})} \supseteq R_{(B_2,P_{B_2})}$;

3) $\forall X \in U/\{d\}$, $\underline{N_{(B_1,P_{B_1})}}(X) \subseteq \underline{N_{(B_2,P_{B_2})}}(X)$;

4) $\overline{POS_{(B_1,P_{B_1})}(\{d\})} \subseteq \overline{POS_{(B_2,P_{B_2})}(\{d\})}$.

The second type of monotonicity refers to the increase of confidence levels. To facilitate the discussion about the monotonicity, an order relation $\preceq$ is defined for the vectors with the same dimension.

Definition 2.5. *Given two vectors* $X = (x_1, x_2, \ldots, x_n)$ *and* $Y = (y_1, y_2, \ldots, y_n)$, $X \preceq Y$ *if* $x_i \leq y_i$, $i = 1, 2, \ldots, n$.

Based on the order relation, the second type of monotonicity is given as follows.

Theorem 2.2. *(Type-2 monotonicity). Let* $S = (U, C, d, V, I, P)$ *be a CVDS,* $B \subseteq C$, *and* P_1, P_2 *be two confidence level subvectors corresponding to* B, *which satisfies* $P_1 \preceq P_2$. *We have*

1) $\forall x \in U$, $n_{(B,P_1)}(x) \subseteq n_{(B,P_2)}(x)$;
2) $R_{(B,P_1)} \subseteq R_{(B,P_2)}$;
3) $\forall X \in U/\{d\}$, $\underline{N_{(B,P_1)}}(X) \supseteq \underline{N_{(B,P_2)}}(X)$;
4) $POS_{(B,P_1)}(\{d\}) \supseteq POS_{(B,P_2)}(\{d\})$.

Theorem 2.1 and Theorem 2.2 reveal the monotonicity of the fundamental concepts in the CVRS model w.r.t. the addition of features and the increase of confidence levels, respectively. The two theorems can significantly improve the efficiency of the heuristic feature-granularity selection algorithm designed in Section 3.

2.2. **Variable cost settings.** In real applications test costs and misclassification costs often occur simultaneously, and they are usually variable but not constant. Concretely, as discussed in [22], acquiring fine-grained data items usually costs more than acquiring coarse-grained ones, so the test cost of a feature often decreases monotonically with the increase of feature-value granularity, or equivalently the error confidence level of feature values. While the variability of the misclassification cost depends on the environment involved and the object considered. Taking the risk evaluation of granting credit as an example, if a customer is misclassified, both the cost (also called benefit if the cost is negative) of the customer and that of the finance company are usually constants. Taking the medical diagnosis as another example, for the misdiagnosis of a specific disease, the misclassification cost of the patient is often fixed, but that of the doctor is usually monotonically increasing with the total test cost paid by the patient. In view of these facts, some types of test cost functions and misclassification cost functions were given in [22], which are introduced as follows.

For a CVDS $S = (U, C, d, V, I, P)$, let tc denote the test cost function, and $tc(a)$ denote the highest test cost of feature a, namely the test cost paid for obtaining the highest data precision for feature a. Given feature a and its confidence level p_a, the test cost function can be represented in different forms according to the application backgrounds. For example, a linear-function-form test cost is

$$tc(a, p_a) = tc(a) \cdot (1 - \lambda_a p_a), \quad p_a \in (0, 0.997], \tag{3}$$

where $\lambda_a \in [0, 1]$ is the adjusting factor; a piecewise-constant-function-form test cost is

$$tc(a, p_a) = TC_i(a), \quad p_a \in [p_{i-1}, p_i) \quad (i = 1, 2, \ldots, m), \tag{4}$$

where m is the number of segments, $p_0 > 0$, $p_m < 1$, and $TC_i(a)$ are constant values satisfying $TC_1(a) > TC_2(a) > \cdots > TC_m(a) > 0$. Then, given a feature-granularity pair (B, P_B), the corresponding total test cost (TTC) is

$$tc(B, P_B) = \sum_{a \in B} tc(a, p_a). \tag{5}$$

Finally, let (k,l) denote the misclassification from class k to class l, which is called a misclassified class pair, and let $mc(B,P_B)_{(k,l)}$ denote the misclassification cost of (k,l) based on (B,P_B) pair. Obviously, if $k=l$, $mc(B,P_B)_{(k,l)}=0$. While if $k\neq l$, $mc(B,P_B)_{(k,l)}$ can be given in multiple forms according to reality. For example, a constant-form misclassification cost is

$$mc(B,P_B)_{(k,l)}=MC_{(k,l)}, \tag{6}$$

where $MC_{(k,l)}>0$ is a constant; a linear-function-form misclassification cost is

$$mc(B,P_B)_{(k,l)}=\gamma_{(k,l)}\cdot tc(B,P_B), \tag{7}$$

where $\gamma_{(k,l)}>0$ is a penalty factor; a piecewise-constant-function-form misclassification cost is

$$mc(B,P_B)_{(k,l)}=MC_j^{(k,l)},\quad tc(B,P_B)\in[TTC_{j-1},TTC_j)\quad (j=1,2,\ldots,n), \tag{8}$$

where n is the number of segments, all TTC_j and $MC_j^{(k,l)}$ are constant values, $TTC_0\geq 0$, and $0<MC_1^{(k,l)}<MC_2^{(k,l)}<\cdots<MC_n^{(k,l)}$.

2.3. **Calculation method of average total cost.** The objective of multi-granularity feature selection is to minimize the total cost consumed in the data processing. An effective calculation method of average total cost was introduced in [22]. Concretely, let $S=(U,C,d,V,I,P)$ be a CVDS, $x\in U$, $B\subseteq C$, and let $mc(x,B,P_B)$ denote the misclassification cost of $x\in U$ based on the feature-granularity pair (B,P_B). The calculation process of average total cost based on (B,P_B) is stated as follows:

1) For each object $x\in U$, classify it according to $n_{(B,P_B)}(x)$ and obtain $mc(x,B,P_B)$. There are two following cases.

A) If $\forall y\in n_{(B,P_B)}(x)$, $d(y)=d(x)$, x can be classified into the right class, so $mc(x,B,P_B)=0$.

B) If $\exists y\in n_{(B,P_B)}(x)$, $d(y)\neq d(x)$, x can be classified into the class which minimizes the total misclassification cost for all objects in $n_{(B,P_B)}(x)$, and the corresponding $mc(x,B,P_B)$ can be obtained according to Equations (3)-(8).

2) Compute the total misclassification cost (TMC) and average misclassification cost (AMC) for all objects in U.

$$TMC(U,B,P_B)=\sum_{x\in U}mc(x,B,P_B),$$

$$AMC(U,B,P_B)=\frac{TMC(U,B,P_B)}{|U|}.$$

3) Compute the average total cost (ATC) for all objects in U.

$$ATC(U,B,P_B)=tc(B,P_B)+AMC(U,B,P_B).$$

2.4. **Test cost-constrained multi-granularity feature selection problem.** In real-world applications, it might happen that the total test cost one can pay is limited, namely there is a test cost constraint. Based on the theoretical framework reviewed above, the test cost-constrained multi-granularity feature selection problem is formally defined below.

Problem 2.1. *The test cost-constrained multi-granularity feature selection problem.*

Input: a CVDS $S=(U,C,d,V,I,P)$, the test cost upper bound M, the test cost function for each feature, and the misclassification cost function for each misclassified class pair;

Output: the pair of selected feature subset B and confidence level vector P_B;

Constraint: $tc(B,P_B)\leq M$;

Optimization objective: $\min(ATC(U,B,P_B))$.

From Problem 2.1, it is known that the test cost-constrained multi-granularity feature selection aims at finding a desirable pair of feature subset and confidence level vector to minimize the average total cost under the test cost constraint. Moreover, the constraint formula will not work if $M = +\infty$, so it can be concluded that the test cost-unconstrained multi-granularity feature selection problem is a special case of the test cost-constrained counterpart.

In summary, based on some existing theory, this section defines formally the test cost-constrained multi-granularity feature selection problem. To solve the new problem, an efficient heuristic algorithm will be designed in the next section.

3. Algorithm Design.

Since not only features but also their respective feature-value granularities are chosen in the multi-granularity feature selection problems, the problems are more complicated than the traditional feature selection problems. In this section, a weighted heuristic feature-granularity selection algorithm is proposed to solve the test cost-constrained multi-granularity feature selection problem.

Algorithm 1 is the main framework of the heuristic feature-granularity selection algorithm, and Algorithm 2 is invoked by Algorithm 1. It is notable that, in the feature-granularity selection algorithm there are two vector-related operations, which have been defined in [22].

Definition 3.1. *Given a vector $X = (x_1, x_2, \ldots, x_n)$ and a number y, $X \sqcup (y) = (x_1, x_2, \ldots, x_n, y)$ denotes a new vector obtained by extending vector X and adding y as its last component.*

Definition 3.2. *Given a vector $X = (x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$, $X - (x_i) = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ denotes a new vector obtained by deleting component x_i from vector X.*

Besides, a so-called feature-granularity significance function, which has also been defined in [22], is used to evaluate the significance of an unselected feature and its feature-value granularity in the proposed algorithm. Concretely, suppose that B and P_B are the selected feature subset and confidence level vector respectively, then the incremental positive region (IPR) induced by feature $a \in C - B$ and confidence level p_a is denoted as

$$IPR_{(B,P_B)}(a, p_a) = POS_{(B\cup\{a\},P_B\sqcup(p_a))}(\{d\}) - POS_{(B,P_B)}(\{d\}).$$

And the weighted feature-granularity significance (FGS) function is defined as

$$FGS_{(B,P_B)}(a, p_a) = |IPR_{(B,P_B)}(a, p_a)| \cdot [tc(a, p_a)]^{\delta}, \tag{9}$$

where $\delta \leq 0$ is the weight. The feature-granularity significance function plays a crucial role in the proposed algorithm.

Algorithm 1 contains three main steps, in which both Step 2 and Step 3 are a while loop. Step 1 is shown in line 1 and it initializes several global variables. As listed in lines 2-23, Step 2 is an addition phase. For each iteration of the while loop in this step, the best pair of feature and confidence level which has the maximal FGS value is added into the feature-granularity pair (B, P_B) and the positive region is expanded gradually. Firstly, for each unselected feature a, the confidence level p_a is tried from the minimal value p_a^0 to the maximal value 0.997 with a step-size $s_a \in (0, 1)$. If the corresponding total test cost does not exceed the upper bound M, Algorithm 2 is invoked to obtain the IPR and the FGS, and the best value p_a^* which makes the FGS maximal is chosen. Then, for all unselected features, their maximal FGS values are compared to select the best feature a'. The addition process continues until the positive region cannot extend any more. Step 3 is shown in lines 24-35 and it is a deletion phase. For each iteration of the

Algorithm 1 A weighted heuristic feature-granularity selection algorithm.

Input: (1) the weight δ, the confidence-level-vector-based decision system (U, C, d, V, I, P), the test cost upper bound M, the test cost function for each feature, and the misclassification cost function for each misclassified class pair;
(2) for each feature $a \in C$, the confidence level's minimal value p_a^0 and the step-size s_a.
Output: the selected feature subset B and confidence level vector P_B.

Set $B = \emptyset$, $P_B = ()$, $POS_{(B,P_B)}(\{d\}) = \emptyset$, $CA = C$, and $S = U$, where B is the selected feature subset, P_B is the selected confidence level vector ("()" denotes an empty vector), $POS_{(B,P_B)}(\{d\})$ is the positive region, CA is the set of unselected features, and S is the set of objects outside the positive region.
while $(|S| > 0)$ **do**
  for (each $a \in CA$) **do**
    for (each $x \in S$) **do**
      sign_x = true;//sign_x is a global variable used in Algorithm 2
    end for
    for ($p_a = p_a^0$; $p_a \leq 0.997$; $p_a = p_a + s_a$) **do**
      if ($tc(a, p_a) <= M \wedge tc(B \cup \{a\}, P_B \sqcup (p_a)) <= M$) **then**
        Use actual parameters S, B, P_B, a, p_a to invoke Algorithm 2, and return $IPR_{(B,P_B)}(a, p_a)$ and $FGS_{(B,P_B)}(a, p_a)$;
        if ($\delta == 0$) **then**
          $p_a^* = p_a$; break;
        end if
      end if
    end for
    Select p_a^* satisfying $FGS_{(B,P_B)}(a, p_a^*) = \max(FGS_{(B,P_B)}(a, p_a))$;
  end for
  Select a' satisfying $FGS_{(B,P_B)}(a', p_{a'}^*) = \max(FGS_{(B,P_B)}(a, p_a^*))$;
  if ($FGS_{(B,P_B)}(a', p_{a'}^*) > 0$) **then**
    $B = B \cup \{a'\}$; $P_B = P_B \sqcup (p_{a'}^*)$; $POS_{(B,P_B)}(\{d\}) = POS_{(B,P_B)}(\{d\}) \cup IPR_{(B,P_B)}(a', p_{a'}^*)$; $CA = CA - \{a'\}$; $S = S - IPR_{(B,P_B)}(a', p_{a'}^*)$;//Update the five variables
  else
    exit while;
  end if
end while
while (ture) **do**
  for (each $a \in B$) **do**
    Compute $ATC(U, B - \{a\}, P_B - (p_a))$;
  end for
  Select a' satisfying $ATC(U, B-\{a'\}, P_B-(p_{a'})) = \min(ATC(U, B-\{a\}, P_B-(p_a)))$;
  if ($ATC(U, B, P_B) > ATC(U, B - \{a'\}, P_B - (p_{a'}))$) **then**
    $B = B - \{a'\}$; $P_B = P_B - (p_{a'})$;//Update the two variables
  else
    exit while;
  end if
end while
return B, P_B;

Algorithm 2 An algorithm to compute the incremental positive region and the feature-granularity significance.

Input: the set S including the objects outside the positive region; the selected feature subset B and confidence level vector P_B; an unselected feature a and its corresponding confidence level p_a.

Output: the incremental positive region (IPR) and the feature-granularity significance (FGS).

1: Set $IPR_{(B,P_B)}(a, p_a) = \emptyset$;
2: **for** (each $x \in S$) **do**
3: **if** (sign_x == true) **then**
4: Compute $n_{(B\cup\{a\},P_B\sqcup(p_a))}(x)$;//Compute the neighborhood of x with respect to the new feature-granularity pair
5: **if** ($\forall X \in U/\{d\}$, $n_{(B\cup\{a\},P_B\sqcup(p_a))}(x) \not\subseteq X$) **then**
6: sign_x = false;
7: **else**
8: $IPR_{(B,P_B)}(a, p_a) = IPR_{(B,P_B)}(a, p_a) \cup \{x\}$;//Update the incremental positive region
9: **end if**
10: **end if**
11: **end for**
12: $FGS_{(B,P_B)}(a, p_a) = |IPR_{(B,P_B)}(a, p_a)| \cdot [tc(a, p_a)]^{\delta}$;//Compute the feature-granularity significance
13: return $IPR_{(B,P_B)}(a, p_a)$, $FGS_{(B,P_B)}(a, p_a)$;

while loop in this step, the average total costs $ATC(U, B - \{a\}, P_B - (p_a))$ are computed for each feature-granularity element $(a, p_a) \in (B, P_B)$; and if their minimal value is less than the currently minimal average total cost $ATC(U, B, P_B)$, the corresponding feature-granularity element is considered to be redundant and is deleted from (B, P_B). The deletion process continues until the average total cost cannot decrease any more. The result of the feature-granularity selection algorithm is an optimal feature-granularity pair that has the minimal average total cost.

Particularly, three accelerating techniques are used in the algorithm according to the monotonicities with respect to the fundamental concepts in the CVRS model. Firstly, as shown in lines 4-6 of Algorithm 1 and lines 3-10 of Algorithm 2, global variable sign_x is employed to judge whether to continue calculating the neighborhood of x when p_a increases. Concretely, assuming that $n_{(B\cup\{a\},P_B\sqcup(p_a))}(x) \not\subseteq X$, $\forall X \in U/\{d\}$, "sign_x = false" is labelled to avoid calculating $n_{(B\cup\{a\},P_B\sqcup(p_a+s_a))}(x)$ because $n_{(B\cup\{a\},P_B\sqcup(p_a+s_a))}(x) \not\subseteq X$ according to Theorem 2.2, and naturally $x \notin IPR_{(B,P_B)}(a, p_a + s_a)$. Hence, this kind of objects does not need to be considered in the computation of $IPR_{(B,P_B)}(a, p_a + s_a)$. Secondly, as shown in lines 10-12 of Algorithm 1, if $\delta = 0$, then for each feature a, we let p_a^* be the minimal confidence level that satisfies the test cost constraint, and the computation of FGS values is avoided for larger confidence levels. That is because $FGS_{(B,P_B)}(a, p_a)$ is equal to $|IPR_{(B,P_B)}(a, p_a)|$ at this time and the latter will decrease with the increase of confidence level according to Theorem 2.2. Finally, as shown in line 19 of Algorithm 1, through using $S = S - IPR_{(B,P_B)}(a', p_{a'}^*)$, the objects needed to be judged whether they belong to the positive region get fewer and fewer as the feature-granularity selection goes on. The reason is that, given $B_1 \subseteq B_2 \subseteq C$ and $P_{B_1} \sqsubseteq P_{B_2}$, if $x \in POS_{(B_1,P_{B_1})}(\{d\})$, then

$x \in POS_{(B_2,P_{B_2})}(\{d\})$ according to Theorem 2.1. So we just need to discuss the objects in $U - POS_{(B_1,P_{B_1})}(\{d\})$ when computing $POS_{(B_2,P_{B_2})}(\{d\})$.

Compared with the existing feature-granularity selection algorithm in [22], the key characteristic of the proposed algorithm is considering the test cost constraint, which is shown in line 8 of Algorithm 1. In this way, some processing methods are different between the two feature-granularity selection algorithms. For example, if $\delta = 0$, then for each feature a, $p_a^* = p_a^0$ in the existing algorithm, while p_a^* is equal to the minimal confidence level satisfying the test cost constraint in the proposed algorithm. In fact, the proposed feature-granularity selection algorithm is a generalization of that in [22], and the latter is a special case of the former with the test cost upper bound $M = +\infty$.

Finally, it is worth mentioning that, for the same dataset and the same cost setting, different δ values will induce different feature-granularity selection results. One could use the competition strategy introduced in [22] to choose the minimal average total cost among multiple δ values. Naturally, the corresponding feature-granularity pair is optimal among these δ values.

4. **Experiments.** Experiments are carried out upon six UCI datasets, whose basic information is listed in Table 3. Before the experimentation, the datasets are preprocessed and some necessary parameters are set. Firstly, data items are normalized onto $[0, 1]$, and missing values are directly set to be 0.5. Secondly, since the UCI datasets have no intrinsic test costs and misclassification costs, we generate the two kinds of cost functions for each dataset according to its application background and Equations (3)-(8). For each dataset in Table 3, we generate 1000 different cost settings, in which the values of cost parameters are similar to those in [22]. Thirdly, the constant k in Equation (2) is set to be a small positive number so that the upper error bound computed by Equations (1) and (2) can lie within a reasonable range. Finally, to show the multi-granularity function of the proposed algorithm more clearly, both the confidence level's minimal value p_a^0 and the step-size s_a are respectively set to be the same among all features (when this assumption does not hold, it is known by experimentation that the results are similar). For example, the experiment results listed below are obtained by setting $k = 0.05$, $p_a^0 = s_a = 0.1$.

Table 3. Data information

Dataset	Domain	Samples	Features	Classes
Credit	finance	690	15	2
Image	graphics	2310	18	7
Liver	clinic	345	6	2
Sonar	physics	208	60	2
Wdbc	clinic	569	30	2
Wpbc	clinic	198	33	2

The weight δ in Equation (9) is set to be a series of values $-4, \ldots, -0.5, 0$. For each dataset and each cost setting, we run the proposed feature-granularity selection algorithm for all δ values under different values of test cost upper bound M. Especially, the case where $M = +\infty$, namely the test cost-unconstrained case is also taken into account for comparison. We first show a group of feature-granularity selection results of Liver dataset with respect to $M = +\infty$, 100, 80 in Tables 4-6 respectively, where the boldface numbers in the fourth columns of the tables are the minimal average total costs among multiple δ values, the integers in the fifth columns are the indexes of selected features, and the unit of run-time is 1ms. It is easy to know from the three tables that the proposed algorithm is effective for solving the test cost-constrained multi-granularity feature selection problem.

TABLE 4. A representative feature-granularity selection result for Liver dataset without test cost constraint, where TTC denotes the total test cost for each object, and AMC and ATC denote respectively the average misclassification cost and the average total cost for all objects

δ	TTC	AMC	ATC	Feature subset	Confidence level vector	Run-time
-4	99.28	206.6087	305.8887	$\{1,2,3,4,6\}$	$(0.997,0.9,0.9,0.997,0.997)$	3431
-3.5	102.0773	124.1739	226.2512	$\{1,2,3,4,6\}$	$(0.9,0.9,0.9,0.997,0.997)$	5258
-3	102.197	124.1739	226.3709	$\{1,2,3,4,6\}$	$(0.997,0.9,0.9,0.8,0.997)$	4332
-2.5	79.3442	64.1159	143.4602	$\{1,2,4,6\}$	$(0.997,0.9,0.7,0.9)$	6717
-2	115.5947	20	135.5947	$\{1,2,3,4,6\}$	$(0.997,0.9,0.6,0.7,0.9)$	5253
-1.5	85.2671	19.7101	104.9773	$\{1,2,4,6\}$	$(0.997,0.9,0.3,0.9)$	7136
-1	89.7119	25.7971	115.509	$\{1,2,4,6\}$	$(0.9,0.3,0.997,0.9)$	6766
-0.5	95.4795	5.5072	**100.9868**	$\{1,2,4,6\}$	$(0.7,0.3,0.997,0.9)$	3240
0	142.9966	0	142.9966	$\{1,2,3\}$	$(0.1,0.1,0.1)$	1600

TABLE 5. The feature-granularity selection result for Liver dataset with the same cost setting as that for Table 4 and test cost upper bound $M = 100$

δ	TTC	AMC	ATC	Feature subset	Confidence level vector	Run-time
-4	99.28	206.6087	305.8887	$\{1,2,3,4,6\}$	$(0.997,0.9,0.9,0.997,0.997)$	2042
-3.5	39.1538	437.6812	476.835	$\{1,4,6\}$	$(0.9,0.997,0.997)$	2832
-3	66.1492	382.6087	448.7579	$\{1,3,4,6\}$	$(0.997,0.9,0.8,0.997)$	2566
-2.5	79.3442	64.1159	143.4602	$\{1,2,4,6\}$	$(0.997,0.9,0.7,0.9)$	2078
-2	79.3442	64.1159	143.4602	$\{1,2,4,6\}$	$(0.997,0.9,0.7,0.9)$	2117
-1.5	85.2671	19.7101	104.9773	$\{1,2,4,6\}$	$(0.997,0.9,0.3,0.9)$	2102
-1	89.7119	25.7971	115.509	$\{1,2,4,6\}$	$(0.9,0.3,0.997,0.9)$	2223
-0.5	95.4795	5.5072	**100.9868**	$\{1,2,4,6\}$	$(0.7,0.3,0.997,0.9)$	3018
0	99.338	11.4783	110.8162	$\{1,2,4\}$	$(0.7,0.1,0.1)$	1774

TABLE 6. The feature-granularity selection result for Liver dataset with the same cost setting as that for Table 4 and test cost upper bound $M = 80$

δ	TTC	AMC	ATC	Feature subset	Confidence level vector	Run-time
-4	50.995	455.0725	506.0675	$\{1,3,6\}$	$(0.997,0.9,0.997)$	1612
-3.5	39.1538	437.6812	476.835	$\{1,4,6\}$	$(0.9,0.997,0.997)$	1907
-3	66.1492	382.6087	448.7579	$\{1,3,4,6\}$	$(0.997,0.9,0.8,0.997)$	1615
-2.5	79.3442	64.1159	143.4602	$\{1,2,4,6\}$	$(0.997,0.9,0.7,0.9)$	2005
-2	79.3442	64.1159	143.4602	$\{1,2,4,6\}$	$(0.997,0.9,0.7,0.9)$	2027
-1.5	49.2193	115.942	165.1614	$\{1,4,6\}$	$(0.997,0.3,0.9)$	2607
-1	77.4747	58.029	**135.5037**	$\{1,2,6\}$	$(0.9,0.3,0.9)$	1778
-0.5	77.4747	58.029	**135.5037**	$\{1,2,6\}$	$(0.9,0.3,0.9)$	1696
0	77.5246	84.8116	162.3362	$\{2,4\}$	$(0.1,0.1)$	1156

For each δ value, the algorithm can obtain a nice feature-granularity selection result. Besides, through comparing the results between Tables 4-6, we observe the following.

1) Even if both p_a^0 and s_a are respectively set to be the same among all features, the best confidence levels are not necessarily the same between different selected features, which verifies that the heuristic algorithm can effectively solve the test cost-constrained multi-granularity feature selection problem.

2) Generally speaking, with the decrease of test cost upper bound, three main kinds of change will occur upon the selected feature-granularity pair to reduce the total test cost: abandoning some feature-granularity elements, coarsening the granularities namely increasing the confidence levels of some features, or replacing some features with cheaper ones. With these changes, the average misclassification cost will increase, and the average total cost will also increase in most cases.

3) With the decrease of test cost upper bound, the run-time drops for most of δ values. This is mainly because the feature-granularity pairs which satisfy the test cost constraint turn less at this time; naturally, the computation of IPR and FGS values also turns less according to lines 8 and 9 of Algorithm 1.

We display the representative results of optimal feature-granularity selection among all δ values for Liver dataset under different values of test cost upper bound in Table 7, including the two ones shown in Table 4 and Table 6 respectively. Furthermore, we draw the representative changes of total test costs and average total costs corresponding to the optimal feature-granularity selection obtained by the proposed algorithm, and meanwhile draw the average total costs obtained by the existing test cost-constrained single-granularity feature selection approach in [20] under the same cost settings for the six datasets in Figures 1-3. For convenience, the three kinds of cost are abbreviated as TTC-MGFS, ATC-MGFS and ATC-SGFS respectively. From Table 7 and the three figures, it can be found that the optimal feature-granularity selection results obtained by the proposed algorithm among multiple δ values follow the rules observed above. Especially, in the extreme case where the test cost upper bound is low enough, there is no solution for the multi-granularity feature selection problem, which is in line with the reality. Moreover, in most cases the proposed heuristic algorithm performs better than the algorithm in [20] on minimizing the consumed total cost even if the latter is a backtracking algorithm. In particular, when the test cost upper bound is relatively low, the proposed algorithm can still find the solution, namely select a feature-granularity pair while the existing algorithm cannot. This is because the multi-granularity feature selection approaches allow different features have different feature-value granularities while the single-granularity ones do not, which is the fundamental advantage of the multi-granularity feature selection approaches.

In summary, it is known from the experimental results that the proposed heuristic feature-granularity selection algorithm performs well on solving the test cost-constrained multi-granularity feature selection problem. By using the algorithm, a desirable pair of feature subset and confidence level vector can be obtained to minimize the consumed total cost under a rational test cost constraint. Besides, the influences of the test cost upper bound value to the feature-granularity selection result are also discussed through experiments.

TABLE 7. The optimal feature-granularity selection results for Liver dataset with the same cost setting as that for Table 4 and different values of test cost upper bound M, where the pairs of feature subset and confidence level vector have the minimal average total costs among $\delta = -4, \ldots, -0.5, 0$

M	Optimal δ	TTC	AMC	ATC	Feature subset	Confidence level vector
$+\infty$	-0.5	95.4795	5.5072	100.9868	$\{1,2,4,6\}$	$(0.7, 0.3, 0.997, 0.9)$
80	$-1, -0.5$	77.4747	58.029	135.5037	$\{1,2,6\}$	$(0.9, 0.3, 0.9)$
50	-1.5	49.2193	115.942	165.16147	$\{1,4,6\}$	$(0.997, 0.3, 0.9)$
30	$-2.5, -2$	27.5058	484.058	511.5638	$\{4,6\}$	$(0.7, 0.997)$
20	$-2.5, -2, -1.5, -1, -0.5$	16.6349	542.029	558.6639	$\{4\}$	(0.7)
10	no solution					

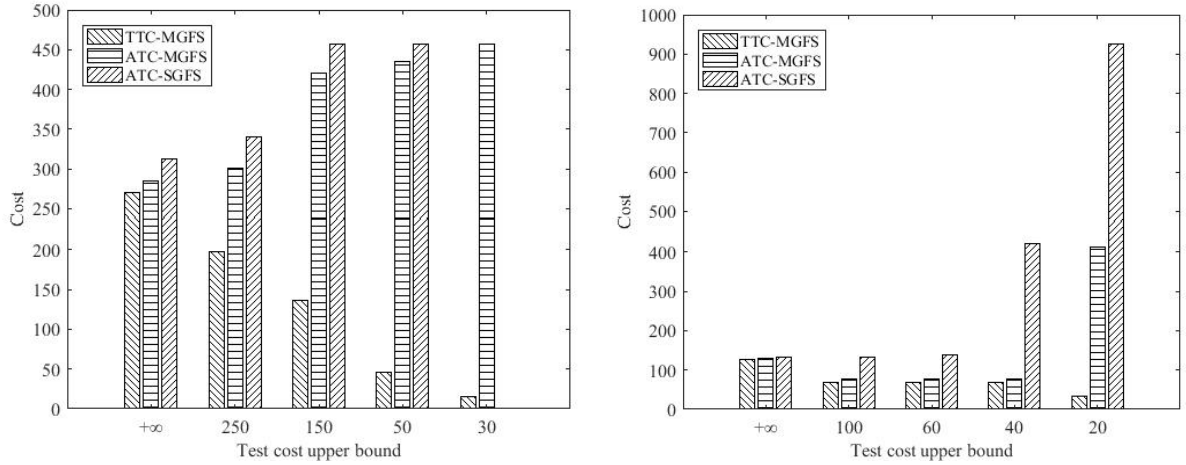

FIGURE 1. Changes of three kinds of cost with different values of test cost upper bound: (a) Credit, (b) Image, where TTC-MGFS and ATC-MGFS respectively denote the total test cost and the average total cost obtained by the proposed algorithm, while ATC-SGFS denotes the average total cost obtained by the existing single-granularity algorithm in [20]. Especially, as shown in Figure (a), when the test cost upper bound is 30, there is no solution for the algorithm in [20].

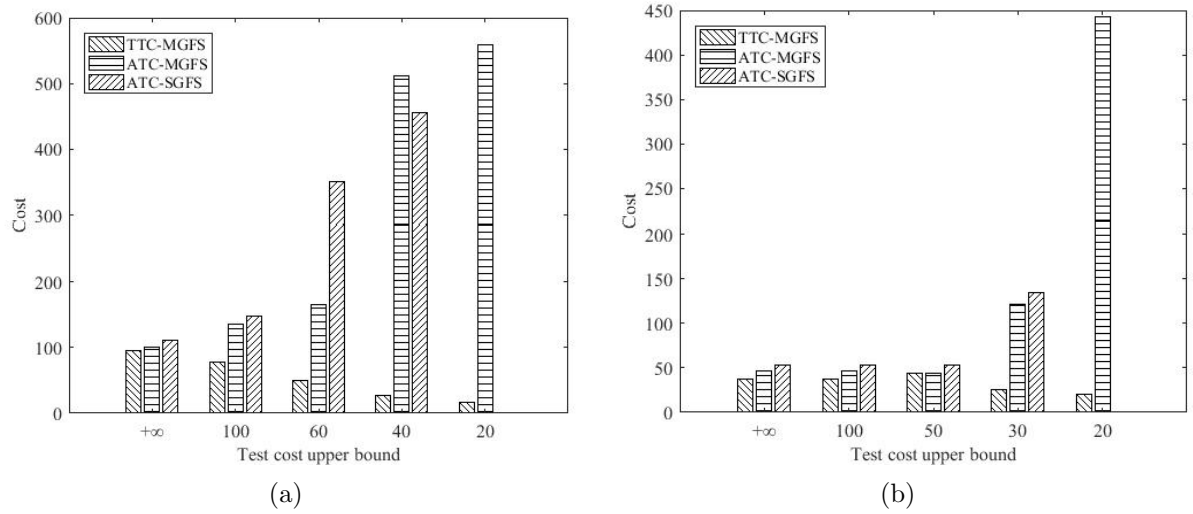

FIGURE 2. Changes of three kinds of cost with different values of test cost upper bound: (a) Liver, (b) Sonar

5. **Conclusions.** In this paper, based on measurement errors and variable costs, an effective approach is proposed to solve the test cost-constrained multi-granularity feature selection problem, so as to fill in the gap that the test cost constraint issue has not been touched in the existing cost-sensitive multi-granularity feature selection approach in [22]. Through using the designed heuristic feature-granularity selection algorithm, a desirable pair of feature subset and error confidence level vector can be selected to minimize the total cost consumed under a rational test cost upper bound. In other words, a satisfactory trade-off among feature dimension reduction, feature-value granularity selection and total

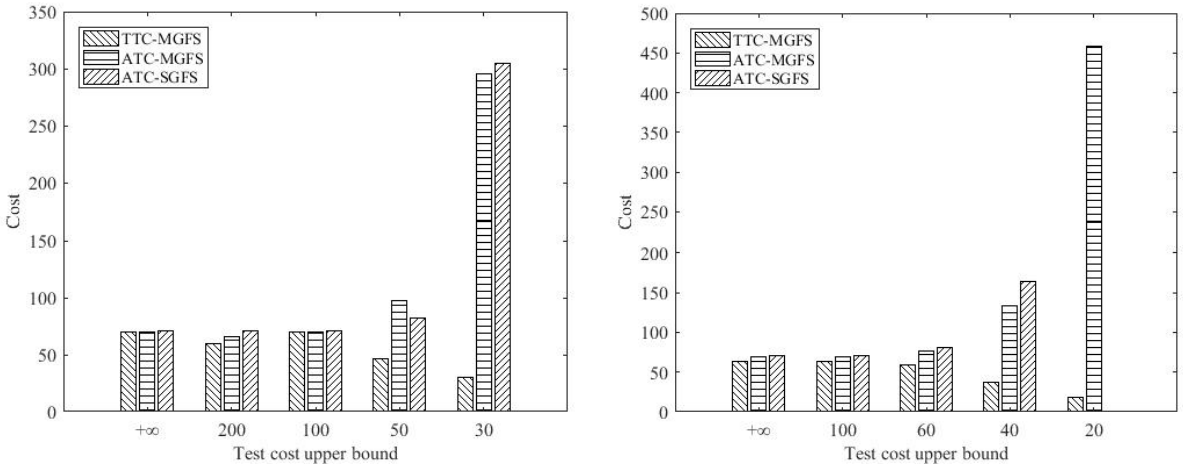

FIGURE 3. Changes of three kinds of cost with different values of test cost upper bound: (a) Wdbc, (b) Wpbc

cost minimization under test cost constraint can be obtained by using the proposed approach. The future work is to present the extended approaches to deal with more complex data, such as the composite data [25].

Acknowledgements. This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 11701258 and 11871259, the Institute of Meteorological Big Data-Digital Fujian and Fujian Key Laboratory of Data Science and Statistics. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] B. Ghaddar and J. Naoum-Sawaya, High dimensional data classification and feature selection using support vector machines, *European Journal of Operational Research*, vol.265, no.3, pp.993-1004, 2018.

[2] T. Ahmad and M. N. Aziz, Data preprocessing and feature selection for machine learning intrusion detection systems, *ICIC Express Letters*, vol.13, no.2, pp.93-101, 2019.

[3] P. Zhou, X. G. Hu, P. P. Li and X. D. Wu, Online streaming feature selection using adapted neighborhood rough set, *Information Sciences*, vol.481, pp.258-279, 2019.

[4] L. Sun, T. Y. Yin, W. P. Ding, Y. H. Qian and J. C. Xu, Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems, *Information Sciences*, vol.537, pp.401-424, 2020.

[5] Z. Pawlak and A. Skowron, Rudiments of rough sets, *Information Sciences*, vol.177, no.1, pp.3-27, 2007.

[6] Y. H. Qian, X. Y. Liang, Q. Wang, J. Y. Liang, B. Liu, A. Skowron, Y. Y. Yao, J. M. Ma and C. Y. Dang, Local rough set: A solution to rough data analysis in big data, *International Journal of Approximate Reasoning*, vol.97, pp.38-63, 2018.

[7] S. Liao, Q. Zhu and R. Liang, On the properties and applications of inconsistent neighborhood in neighborhood rough set models, *IEICE Trans. Information and Systems*, vol.E101-D, no.3, pp.709-718, 2018.

[8] J. M. Zhan, X. H. Zhang and Y. Y. Yao, Covering based multigranulation fuzzy rough sets and corresponding applications, *Artificial Intelligence Review*, vol.53, no.2, pp.1093-1126, 2020.

[9] J. H. Dai, Q. H. Hu, H. Hu and D. B. Huang, Neighbor inconsistent pair selection for attribute reduction by rough set approach, *IEEE Trans. Fuzzy Systems*, vol.26, no.2, pp.937-950, 2018.

[10] Y. B. Wang, X. J. Chen and K. Dong, Attribute reduction via local conditional entropy, *International Journal of Machine Learning and Cybernetics*, vol.10, pp.3619-3634, 2019.

[11] H. Y. Wan, G. Q. Wu, M. L. Yu and M. T. Yuan, Software defect prediction based on cost-sensitive dictionary learning, *International Journal of Software Engineering and Knowledge Engineering*, vol.29, no.9, pp.1219-1243, 2019.
[12] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel and R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, *IEEE Trans. Neural Networks and Learning Systems*, vol.29, no.8, pp.3573-3587, 2018.
[13] P. D. Turney, Types of cost in inductive concept learning, *Proc. of the Workshop on Cost-Sensitive Learning at the 17th ICML*, pp.1-7, 2000.
[14] Y. Fang, Z. H. Liu and F. Min, A PSO algorithm for multi-objective cost-sensitive attribute reduction on numeric data with error ranges, *Soft Computing*, vol.21, pp.7173-7189, 2017.
[15] S. Liao, Q. Zhu and F. Min, Cost-sensitive attribute reduction in decision-theoretic rough set models, *Mathematical Problems in Engineering*, vol.2014, Article ID 875918, pp.1-9, 2014.
[16] W. H. Shu and H. Shen, Multi-criteria feature selection on cost-sensitive data with missing values, *Pattern Recognition*, vol.51, pp.268-280, 2016.
[17] A. H. Tan, W. Z. Wu and Y. Z. Tao, A set-cover-based approach for the test-cost-sensitive attribute reduction problem, *Soft Computing*, vol.21, pp.6159-6173, 2017.
[18] S. Liao, Q. Zhu and R. Liang, An efficient approach of test-cost sensitive attribute reduction for numerical data, *International Journal of Innovative Computing, Information and Control*, vol.13, no.6, pp.2099-2111, 2017.
[19] H. Zhao and W. Zhu, Optimal cost-sensitive granularization based on rough sets for variable costs, *Knowledge-Based Systems*, vol.65, pp.72-82, 2014.
[20] S. Liao, Q. Zhu and R. Liang, Selecting attributes and granularity for data with test cost constraint, *Computer Engineering and Science*, vol.40, no.8, pp.1468-1474, 2018 (in Chinese).
[21] S. Liao, Q. Zhu and Y. Qian, Feature-granularity selection with variable costs for hybrid data, *Soft Computing*, vol.23, pp.13105-13126, 2019.
[22] S. Liao, Q. Zhu, Y. Qian and G. P. Lin, Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs, *Knowledge-Based Systems*, vol.158, pp.25-42, 2018.
[23] R. A. Fisher, On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol.222, pp.309-368, 1922.
[24] J. Neyman, Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol.236, no.767, pp.333-380, 1937.
[25] S. Y. Li, T. R. Li and J. Hu, Update of approximations in composite information systems, *Knowledge-Based Systems*, vol.83, pp.138-148, 2015.