Soft Computing
https://doi.org/10.1007/s00500-019-03854-2

METHODOLOGIES AND APPLICATION

Feature–granularity selection with variable costs for hybrid data

Shujiao Liao[1] · Qingxin Zhu[2] · Yuhua Qian[3]

Abstract

In recent years, cost-sensitive feature selection has drawn much attention. However, some issues still remain to be investigated. Particularly, most existing work deals with single-typed data, while only a few studies deal with hybrid data; moreover, both the test cost of a feature and the misclassification cost of an object are often assumed to be fixed, but in fact they are usually variable with the error range of the data, or equivalently the data granularity. In view of these facts, a feature–granularity selection approach is proposed to select the optimal feature subset and the optimal data granularity simultaneously to minimize the total cost for processing hybrid data. In the approach, firstly an adaptive neighborhood model is constructed, in which the neighborhood granules are generated adaptively according to the types of features. Then, multiple kinds of variable cost setting are discussed according to reality, and finally, an optimal feature–granularity selection algorithm is designed. Experimental results on sixteen UCI datasets show that a good trade-off among feature dimension reduction, data granularity selection and total cost minimization could be achieved by the proposed algorithm. In particular, the influences of different cost settings to the feature–granularity selection are also discussed thoroughly in the paper, which would provide some feasible schemes for decision making.

Keywords Adaptive neighborhood · Feature–granularity selection · Hybrid data · Measurement errors · Variable costs

1 Introduction

Cost-sensitive learning is one of the key issues in data mining and machine learning. It takes cost information into consideration and thus is close to real applications (Chai et al. 2004; Domingos 1999; Du et al. 2007; Greiner et al. 2002; Zhou and Zhou 2016). Test cost and misclassification cost are two main types of cost addressed in cost-sensitive learning (Turney 2000). Test cost, also called feature cost or acquisition cost, is the money, time or other resources consumed in collecting a data item for an object. For example, it takes both money and time to obtain the medical data of a patient. Misclassification cost is the extra consumption caused by categorizing an object into a class that it does not belong to. Different types of misclassification often incur different costs. For instance, the cost of misdiagnosing a patient as healthy may be much larger than that of misdiagnosing a healthy person as sick.

Cost-sensitive feature selection, as a powerful mechanism, is generated by introducing cost-sensitive learning into the domain of feature selection. As is well known, feature selection is an important data preprocessing technique in data mining, machine learning and pattern recognition. For a dataset, necessary and discriminative features could be chosen, and at the same time irrelevant or redundant features could be removed by using the feature selection technique (Boussouf and Quafafou 2000; Dash and Liu 2003; Guyon and Elisseeff 2003; Hu et al. 2010; Kannan and Ramaraj 2010; Liang et al. 2012). Consequently, the data dimensionality could be reduced effectively, and the subsequent data processing will become more efficient. Cost-sensitive feature selection aims to select a suitable feature subset which

Communicated by V. Loia.

✉ Shujiao Liao
sjliao2011@163.com

Qingxin Zhu
qxzhu@uestc.edu.cn

Yuhua Qian
jinchengqyh@126.com

[1] School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, China

[2] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

[3] Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China

Published online: 21 February 2019

can minimize one type of cost or the summation of several types of cost and meanwhile keep the ability of original decision system as much as possible. Generally speaking, through considering cost information in the feature selection process, cost-sensitive feature selection is more close to real applications than cost-insensitive feature selection.

Combined with the reality, there are two major challenges in the study of cost-sensitive feature selection. One challenge is that measurement errors are ubiquitous in the process of data collecting. For a quantity, its measure error is the difference between the measured value and the true value. It is notable that measurement errors are only considered for numeric data but not nominal data. The reason is that there are usually just several distinct feature values for a nominal feature, i.e., the feature values could be regarded as being roughly sorted into several categories. It is not significant to take the measurement errors into account for nominal features. For a numeric quantity, its measurement errors often satisfy a normal (or near normal) distribution in real applications. Obviously, for a numeric or mixed dataset, the larger the error range is, the bigger the data granularity is, and the lower the data precision becomes. The other challenge is that both the cost paid for testing a numeric feature and the cost induced by a misclassification are often not fixed but variable in practical applications. Concretely, fine-grained data items are usually more costly to collect than the coarse-grained ones; thus, for a numeric feature, the test cost is monotone decreasing with the feature values' granularity (note that, for any nominal feature, the test cost is supposed to be constant because measurement errors are not considered for nominal features; namely, all nominal features are regarded to be fixed granularity and fixed test cost). Meanwhile, for an object, the misclassification cost is often monotone increasing with the total test cost. Taking the same misdiagnosis of a particular disease as an example, if high test costs have been paid, the patient may be very angry and ask for high compensation, so the misclassification cost is higher than that in the case of low test costs while the total test cost is determined by the selected features and the data granularity. Hence, it is meaningful to obtain a trade-off among data granularity, feature subset and variable costs.

To address the challenges discussed above, in this paper a cost-sensitive feature–granularity selection (the selection of feature subset and data granularity, namely, the feature values' granularity) approach is proposed for hybrid data. Hybrid data exist widely in practical applications (Chen and Yang 2014; Hu et al. 2010). It is notable that some other researchers have also presented feature–granularity selection approaches in recent years, but they have not touched cost factors (Ansorge and Schmidt 2015). The proposed cost-sensitive feature–granularity selection approach aims at finding an optimal feature–granularity pair (the pair of feature subset and data granularity) to minimize the total cost (the summation of consumed test costs and misclassification costs). The relationship between the above-mentioned challenges and the aim of the proposed approach is shown in Fig. 1, which is meanwhile the derivation of the approach.

In the proposed feature–granularity selection approach, for numeric features, their feature values' measurement errors are assumed to satisfy a normal distribution, and the data granularity is evaluated by the confidence level of the measurement errors. Accordingly, the error confidence level

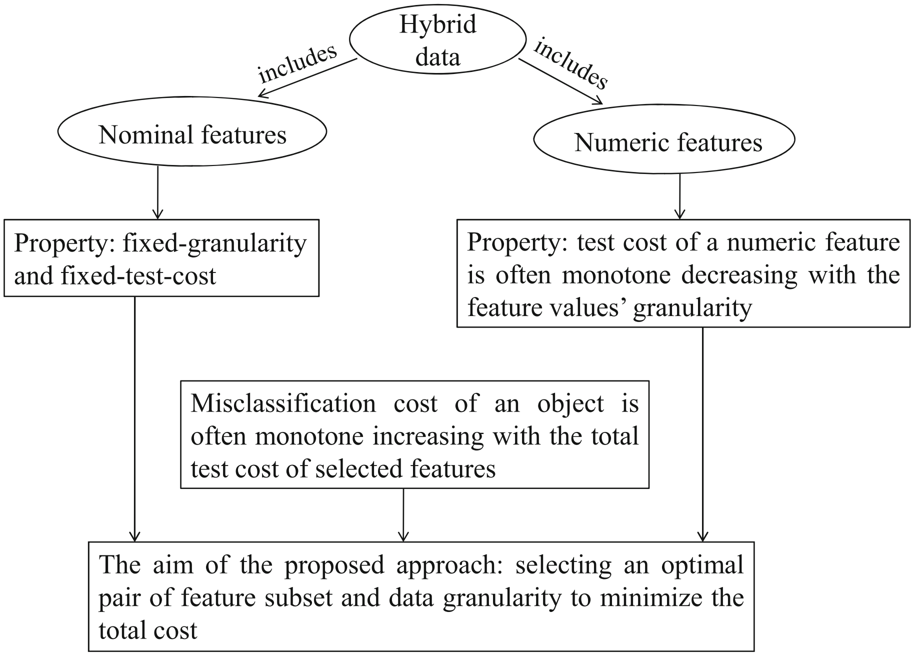

Hybrid data
includes
includes
Nominal features
Numeric features
Property: fixed-granularity and fixed-test-cost
Property: test cost of a numeric feature is often monotone decreasing with the feature values' granularity
Misclassification cost of an object is often monotone increasing with the total test cost of selected features
The aim of the proposed approach: selecting an optimal pair of feature subset and data granularity to minimize the total cost

Fig. 1 Derivation of the proposed feature–granularity selection approach

is closely related to the data precision. In this context, an adaptive neighborhood model is constructed, in which the neighborhood granules of objects on a given feature subset are adaptively computed according to the types of the features. If a feature is numeric, the neighborhoods are constructed according to the error confidence level; while if the feature is nominal, the neighborhoods are generated according to the equivalence relations. The properties with respect to the built model are discussed thoroughly. Then, several types of variable cost setting are introduced according to reality, in which the relationship among error confidence level, test costs and misclassification costs is taken into consideration. The calculation method of average total cost is also developed for any given pair of feature subset and confidence level. Furthermore, the influences of cost setting changes to the feature–granularity selection results are studied. Finally, an optimal feature–granularity selection algorithm is designed, by which not only the optimal feature subset but also the optimal confidence level can be selected to minimize the average total cost. Three pruning techniques are employed in the algorithm to improve the computational efficiency. Experiments are undertaken on sixteen datasets from the University of California—Irvine (UCI) Library (Blake and Merz 1998) with multiple different cost settings. Experimental results demonstrate the effectiveness of the feature–granularity selection algorithm. A satisfactory trade-off among feature dimension reduction, data granularity selection and total cost minimization can be obtained by the algorithm. The algorithm outperforms multiple existing feature selection algorithms on minimizing the total cost consumed in data processing. In particular, through in-depth experimental analyses concerning the influences of different cost settings, some feasible suggestions are given for decision making.

The rest of the paper is organized as follows. The related work is introduced in Sect. 2. Section 3 builds the adaptive neighborhood model and discusses the notions and properties in the model. Section 4 first constructs several types of variable cost setting, then develops the calculation method of average total cost and finally investigates the influences of cost setting changes. Section 5 proposes the optimal feature–granularity selection algorithm. Experiment settings and results are discussed in depth in Sect. 6. Finally, Sect. 7 concludes the paper and suggests further research ideas.

2 Related work

In recent years, cost-sensitive feature selection, as an effective incorporation of cost-sensitive learning and feature selection, has drawn much attention due to its wide application backgrounds. Some related literatures are reviewed in this section.

Zhou et al. (2016) proposed a random forest-based feature selection algorithm that incorporates test costs into the base decision tree construction process to produce low-cost feature subsets. Iswandy and Koenig (2006) studied a multi-objective extension of feature selection which considers test costs for optimizing sensor system design. Wang et al. (2010) handled the issue of data over-fitting in test cost-sensitive decision tree learning by combining feature selection, smoothing and threshold pruning. Cao et al. (2013) improved the classification performance of cost-sensitive support vector machine by simultaneously optimizing the pair of feature subset, intrinsic parameters and misclassification cost parameters. Pendharkar (2013) proposed a two-stage solution approach for solving the misclassification cost minimization feature selection problem. Weiss et al. (2013) developed a feature selection approach which takes feature costs and misclassification costs into consideration based on histogram comparisons and a genetic search strategy. Bian et al. (2016) presented a cost-sensitive feature selection approach that adds the test cost- and misclassification cost-based evaluation function of a filter feature selection using a chaos genetic algorithm. Zhang et al. (2008) put forward an attribute selection strategy, which is a trade-off method between attribute information and cost information that includes test costs and misclassification costs with different units, for selecting splitting attributes in decision trees induction. Zhang (2017) presented the first study of multi-objective particle swarm optimization for test cost-sensitive feature selection problems. Huang and Zhu (2017) defined the cost distance among samples and developed a misclassification cost-based feature selection approach via manifold learning. Zhao and Yu (2019) made use of the $l_{2,1}$-norm to propose an embedded feature selection algorithm based on test costs and misclassification costs.

Particularly, there is a series of research work about cost-sensitive feature selection in rough set domain, in which feature selection is also termed attribute reduction. Min et al. defined the minimal test cost attribute reduction problem in Min et al. (2011) and studied the feature selection with test cost constraint in Min et al. (2014). Liao et al. (2017) designed a fast forward test cost-sensitive attribute reduction algorithm for numerical data by using the properties of inconsistent neighborhoods in neighborhood rough set models. Yao and Zhao (2008) addressed the attribute reduction problem regarding different classification properties, such as decision monotonicity, coverage, decision cost which mainly includes misclassification costs, and so on. Jia et al. (2013) proposed a heuristic algorithm, a genetic algorithm and a simulated annealing algorithm to solve the minimum decision cost attribute reduction problem. Liao et al. (2014) designed a backtracking algorithm and a heuristic algorithm to solve the attribute reduction involving both decision cost and test costs. Shu and Shen (2016) presented a multi-criteria

Table 1 Some quantile values of standard normal distribution

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.997
z_p	0.1257	0.2533	0.3853	0.5244	0.6745	0.8416	1.0364	1.2816	1.6449	3.0

evaluation function to address the test cost- and misclassification cost-based feature selection problem for data with missing values. Zhao et al. (2013) studied the test cost- and misclassification cost-sensitive feature selection problem for numeric data with normal distribution measurement errors. Liao et al. (2018) proposed a multi-granularity feature selection approach which takes measurement errors as well as variable test costs and misclassification costs into consideration. Yu and Zhao (2018) presented a test cost- and misclassification cost-based feature selection approach, in which the importance of each feature is evaluated by both rough sets and Laplacian score.

In general, although cost-sensitive feature selection has been studied from different perspectives, the issue of feature selection with measurement errors and variable costs for hybrid data remains to be investigated and it is the main topic which this paper concerns.

3 Adaptive neighborhood model for hybrid data

In this section, an adaptive neighborhood model is built for hybrid data. The section starts from reviewing some preliminaries about confidence level and confidence interval. Then, measurement errors are introduced into hybrid decision systems, especially in terms of error confidence level. Finally, adaptive neighborhoods and corresponding coverings are developed.

3.1 Some preliminaries about confidence interval and confidence level

Confidence interval, confidence level and confidence limit are three commonly used concepts in statistics (Fisher 1922). Confidence interval is a kind of interval estimation for a population parameter, and the confidence level determines how frequently an observed interval contains a specific parameter value. The left endpoint and right endpoint of confidence interval are called the lower confidence limit and upper confidence limit, respectively. For a normal distribution, the confidence interval and the confidence level follow a so-called 3-sigma rule, which refers to that 99.7% of the data lie within 3 standard deviations of the mean.

For a normal distribution, the confidence interval can be computed through the quantile function of the confidence level (Fisher 1922). Concretely, for a standard normal distribution, the quantile function is

$$z_p = \sqrt{2}\mathrm{erf}^{-1}(2p-1),\ p \in (0,1), \tag{1}$$

where p is a confidence level and $\mathrm{erf}^{-1}(2p-1)$ is the inverse error function. Given a normal distribution with mean μ and variance σ^2, the quantile function is

$$F^{-1}(p) = \mu + \sigma z_p,\ p \in (0,1). \tag{2}$$

According to Eqs. (1–2), the confidence interval $[\mu - \sigma z_p, \mu + \sigma z_p]$ can be obtained. Assuming that x is a normal random variable with mean μ and variance σ^2, it will lie inside $[\mu - \sigma z_p, \mu + \sigma z_p]$ with probability $2p - 1$ and lie outside the interval with probability $2(1-p)$. Some typical quantile values of standard normal distribution are listed in Table 1.

3.2 Error-related hybrid decision systems

Decision system is a fundamental concept in data mining and machine learning. Traditional decision system was defined as follows:

Definition 1 (Yao 2004) A decision system (DS) S is the 5-tuple:

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

where U is a finite nonempty set of objects called the universe, C is the set of conditional attributes (also called features), D is the set of decision attributes with only discrete values, 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 $a \in C \cup D$.

Hybrid data exist in many real-world applications, but the hybrid characteristic cannot be found from Definition 1. To deal with hybrid datasets, the hybrid decision system is defined as follows:

Definition 2 A *hybrid decision system* (HDS) S is the 5-tuple:

$$S = (U, C = C_o \bigcup C_u, D, V, I),$$

where U, D, V, I have the same meanings as in Definition 1 and C is a hybrid feature set which is composed of a nominal feature subset C_o and a numeric feature subset C_u, where $C_o \cap C_u = \emptyset$.

Table 2 A hybrid decision system, which is a sub-table of Credit dataset

	a_1	a_2	a_3	a_4	D
x_1	a	0.6755	0.1593	u	+
x_2	b	0.4323	0.1755	y	+
x_3	b	0.0564	0.7857	l	+
x_4	a	0.2932	0.1071	y	−
x_5	b	0.4123	0.0477	u	−
x_6	a	0.1229	0.4166	u	−

An example of hybrid decision system is shown in Table 2, which is a sub-table of Credit dataset from the UCI library. Obviously, $U = \{x_1, x_2, \ldots, x_6\}, C_o = \{a_1, a_4\}, C_u = \{a_2, a_3\}$. The feature values for numeric features a_2, a_3 have been normalized, and the decision values "+" and "−" represent positive and negative instances of people who are and are not granted credit, respectively.

As discussed earlier, errors exist widely in real applications. For a quantity, the wider its error interval is, the bigger its granularity is. In this paper, the feature values' measurement errors are taken into account for numeric features. If the error ranges are the same for all numeric features, an error range-related hybrid decision system can be defined as follows:

Definition 3 An *error range-related hybrid decision system* (ERHDS) S is the 6-tuple:

$$S = (U, C = C_o \bigcup C_u, D, V, I, e),$$

where U, C, C_o, C_u, D, V, I have the same meanings as in Definition 2 and $e > 0$ is the error range for any numeric feature $a \in C_u$.

However, because of the diversity among different features, their error ranges are not necessarily the same. To address this situation, the error confidence level is used to measure the data granularity for different features uniformly. An error confidence level-related hybrid decision system is defined as follows:

Definition 4 An *error confidence level-related hybrid decision system* (ECLHDS) S is the 6-tuple:

$$S = (U, C = C_o \bigcup C_u, D, V, I, p), \tag{3}$$

where U, C, C_o, C_u, D, V, I have the same meanings as in Definition 2 and $p \in (0, 1)$ is the error confidence level for any numeric feature $a \in C_u$.

Naturally, when the error range or the error confidence level is not taken into consideration, namely, $e = 0$ in Definition 3 or $p = 0$ in Definition 4, the ERHDS or the ECLHDS degenerates to a HDS. Hence, both ERHDS and ECLHDS are a generalization of the HDS. This paper focuses on the ECLHDS but not the ERHDS because the former is more close to the real applications.

For a quantity, assuming that its measurement errors follow a normal distribution with mean 0 and variance σ^2, then the error intervals can be written as $[-\sigma z_p, \sigma z_p]$, $p \in (0, 0.997]$ according to Eq. (2). The reason why 0.997 rather than 1 is chosen as the maximal confidence level is that the error interval is $(-\infty, \infty)$ when $p = 1$, which is not practical in real applications. Let $e(a, p)$ denote the upper error bound with respect to (w.r.t.) a numeric feature a and a confidence level p, then one has that

$$e(a, p) = \sigma_a z_p, \; p \in (0, 0.997]. \tag{4}$$

Obviously,

$$\max(e(a, p)) = e(a, 0.997) = 3\sigma_a. \tag{5}$$

Let

$$\sigma_a = k_a \cdot \max|a(x_i) - \overline{a(x)}|, \; 1 \leq i \leq |U|, \tag{6}$$

where $\overline{a(x)} = \frac{1}{|U|}\sum_{i=1}^{|U|} a(x_i)$ and the constants $k_a > 0$ could be applied to adjust the maximal upper error bounds according to users' preference. Then, for each numeric feature a and feature values' error confidence level p, the upper error bound can be computed according to Eqs. (4) and (6), and it is adaptive to the feature and the confidence level instead of being totally set by the users.

3.3 Adaptive neighborhoods and the coverings

Neighborhood granule is a key notion for dealing with numeric or hybrid data (Hu et al. 2008). A type of adaptive neighborhood is defined based on the error confidence level-related hybrid decision system as follows:

Definition 5 Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS, $a \in C$ and $x \in U$. The adaptive neighborhood of x w.r.t. feature a and error confidence level p is denoted as $n_{(a,p)}(x)$, and
if $a \in C_o$,

$$n_{(a,p)}(x) = \{x' \in U | a(x') = a(x)\}; \tag{7}$$

if $a \in C_u$,

$$n_{(a,p)}(x) = \{x' \in U || a(x') - a(x)| \leq 2e(a, p)\}. \tag{8}$$

From Definition 5, it is known that the neighborhoods are computed adaptively according to the types of the features. In particular, if $a \in C_o$, the neighborhoods do not change with

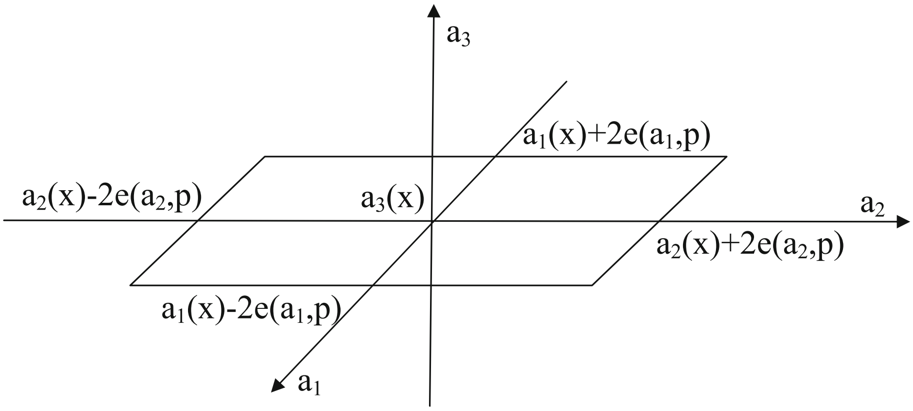


Fig. 2 An exemplary three-dimensional neighborhood, in which a_1, a_2 are numeric features and a_3 is a nominal feature

the confidence level. Note that, given a feature $a \in C_u$, the error interval is $[-e(a, p), e(a, p)]$, while the neighborhood interval of x w.r.t. a is $[a(x) - 2e(a, p), a(x) + 2e(a, p)]$. The reason why $2e(a, p)$ rather than $e(a, p)$ is chosen as the maximal distance in Eq. (8) is analyzed as follows. Suppose that the true value of $x \in U$ is $a'(x)$ for $a \in C_u$, then $a'(x) - e(a, p) \leq a(x) \leq a'(x) + e(a, p)$. In extreme cases, $a'(x) - e(a, p)$ and $a'(x) + e(a, p)$ can be the measured values of the same object. At this time, $|(a'(x) - e(a, p)) - (a'(x) + e(a, p))| = 2e(a, p)$. Hence, for feature $a \in C_u$, the objects with measured value differing from $a(x)$ by no more than $2e(a, p)$ should be drawn into the neighborhood $n_{(a,p)}(x)$ together.

Naturally, for an attribute subset $B \subseteq C$ and a confidence level p, the neighborhood of sample x induced by B is the intersection of the basic neighborhoods induced by each single attribute in B, i.e.,

$$n_{(B,p)}(x) = \bigcap_{a \in B} n_{(a,p)}(x). \tag{9}$$

An exemplary three-dimensional neighborhood is shown in Fig. 2, where a_1, a_2 are numeric features, and a_3 is a nominal feature.

It is easy to obtain the following proposition:

Proposition 1 *Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS, $B \subseteq C$. Then, for any object in U, its neighborhood based on B and p satisfies*

(1) reflexivity: $\forall x \in U, x \in n_{(B,p)}(x)$;
(2) symmetry: $\forall x_i, x_j \in U$, if $x_j \in n_{(B,p)}(x_i)$, $x_i \in n_{(B,p)}(x_j)$.

For each sample in an ECLHDS, the size of its neighborhood is influenced by the given feature subset and error confidence level. In the following two propositions, the monotonicity of neighborhoods is discussed in terms of the two factors, respectively.

Proposition 2 *Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS and $B_1 \subseteq B_2 \subseteq C$. For any $x \in U$, one has that*

$$n_{(B_1,p)}(x) \supseteq n_{(B_2,p)}(x). \tag{10}$$

Proof Because $B_1 \subseteq B_2$, there are more features in B_2 than in B_1. It is easy to know that $n_{(B_1,p)}(x) \supseteq n_{(B_2,p)}(x)$ according to Definition 5 and Eq. (9). □

Proposition 3 *Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS, $B \subseteq C$ and $p_1 \leq p_2$. For any $x \in U$, one has that*

$$n_{(B,p_1)}(x) \subseteq n_{(B,p_2)}(x). \tag{11}$$

Proof $\forall a \in B$, if $a \in C_o$; the neighborhoods do not change with the confidence level according to Eq. (7), so $n_{(a,p_1)}(x) = n_{(a,p_2)}(x)$; if $a \in C_u$, $\forall x' \in n_{(a,p_1)}(x)$, $|a(x') - a(x)| \leq 2e(a, p_1) \leq 2e(a, p_2)$ in that $p_1 \leq p_2$, so $x' \in n_{(a,p_2)}(x)$, $n_{(a,p_1)}(x) \subseteq n_{(a,p_2)}(x)$. According to Eq. (9), one has that $n_{(B,p_1)}(x) \subseteq n_{(B,p_2)}(x)$. □

From Propositions 2–3, it is known that in an ECLHDS, the neighborhoods shrink with the addition of features, while they expand with the increase in error confidence level.

Covering, as a common concept in set theory (Liu and Sai 2009), is discussed in the new environment. Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS, $B \subseteq C$. It is easy to know that $\forall x \in U, x \in n_{(B,p)}(x)$, so $U \subseteq \{n_{(B,p)}(x) | x \in U\}$. Hence, $\{n_{(B,p)}(x) | x \in U\}$ is a covering of U. It is denoted as $\mathrm{Cov}(B, p)$ for brevity, i.e.,

$$\mathrm{Cov}(B, p) = \{n_{(B,p)}(x) | x \in U\}. \tag{12}$$

Obviously, there are $|U|$ elements in $\mathrm{Cov}(B, p)$, and each element is an object subset of U. An order relation is defined for coverings as follows:

Definition 6 Suppose that $\mathrm{Cov}_1, \mathrm{Cov}_2$ are two coverings on the same universe. If for any $K \in \mathrm{Cov}_1$, there exists $L \in$

Table 3 Some neighborhoods with error confidence level $p = 0.6$

U	$n_{(\{a_1\},0.6)}(x)$	$n_{(\{a_2\},0.6)}(x)$	$n_{(\{a_3\},0.6)}(x)$	$n_{(\{a_4\},0.6)}(x)$	$n_{(\{a_1,a_2\},0.6)}(x)$	$n_{(\{a_1,a_2,a_3\},0.6)}(x)$	$n_{(C,0.6)}(x)$
x_1	$\{x_1, x_4, x_6\}$	$\{x_1\}$	$\{x_1, x_2, x_4, x_5\}$	$\{x_1, x_5, x_6\}$	$\{x_1\}$	$\{x_1\}$	$\{x_1\}$
x_2	$\{x_2, x_3, x_5\}$	$\{x_2, x_5\}$	$\{x_1, x_2, x_4\}$	$\{x_2, x_4\}$	$\{x_2, x_5\}$	$\{x_2\}$	$\{x_2\}$
x_3	$\{x_2, x_3, x_5\}$	$\{x_3, x_6\}$	$\{x_3\}$	$\{x_3\}$	$\{x_3\}$	$\{x_3\}$	$\{x_3\}$
x_4	$\{x_1, x_4, x_6\}$	$\{x_4\}$	$\{x_1, x_2, x_4, x_5\}$	$\{x_2, x_4\}$	$\{x_4\}$	$\{x_4\}$	$\{x_4\}$
x_5	$\{x_2, x_3, x_5\}$	$\{x_2, x_5\}$	$\{x_1, x_4, x_5\}$	$\{x_1, x_5, x_6\}$	$\{x_2, x_5\}$	$\{x_5\}$	$\{x_5\}$
x_6	$\{x_1, x_4, x_6\}$	$\{x_3, x_6\}$	$\{x_6\}$	$\{x_1, x_5, x_6\}$	$\{x_6\}$	$\{x_6\}$	$\{x_6\}$

Table 4 Some neighborhoods with error confidence level $p = 0.9$

U	$n_{(\{a_1\},0.9)}(x)$	$n_{(\{a_2\},0.9)}(x)$	$n_{(\{a_3\},0.9)}(x)$	$n_{(\{a_4\},0.9)}(x)$	$n_{(\{a_1,a_2\},0.9)}(x)$	$n_{(\{a_1,a_2,a_3\},0.9)}(x)$	$n_{(C,0.9)}(x)$
x_1	$\{x_1, x_4, x_6\}$	$\{x_1\}$	$\{x_1, x_2, x_4, x_5\}$	$\{x_1, x_5, x_6\}$	$\{x_1\}$	$\{x_1\}$	$\{x_1\}$
x_2	$\{x_2, x_3, x_5\}$	$\{x_2, x_4, x_5\}$	$\{x_1, x_2, x_4, x_5, x_6\}$	$\{x_2, x_4\}$	$\{x_2, x_5\}$	$\{x_2, x_5\}$	$\{x_2\}$
x_3	$\{x_2, x_3, x_5\}$	$\{x_3, x_6\}$	$\{x_3\}$	$\{x_3\}$	$\{x_3\}$	$\{x_3\}$	$\{x_3\}$
x_4	$\{x_1, x_4, x_6\}$	$\{x_2, x_4, x_5, x_6\}$	$\{x_1, x_2, x_4, x_5\}$	$\{x_2, x_4\}$	$\{x_4, x_6\}$	$\{x_4\}$	$\{x_4\}$
x_5	$\{x_2, x_3, x_5\}$	$\{x_2, x_4, x_5\}$	$\{x_1, x_2, x_4, x_5\}$	$\{x_1, x_5, x_6\}$	$\{x_2, x_5\}$	$\{x_2, x_5\}$	$\{x_5\}$
x_6	$\{x_1, x_4, x_6\}$	$\{x_3, x_4, x_6\}$	$\{x_2, x_6\}$	$\{x_1, x_5, x_6\}$	$\{x_4, x_6\}$	$\{x_6\}$	$\{x_6\}$

Cov_2 satisfying $K \subseteq L$; we say that Cov_1 is finer than Cov_2 or equivalently Cov_2 is coarser than Cov_1 which is denoted as $\mathrm{Cov}_1 \preceq \mathrm{Cov}_2$ or equivalently $\mathrm{Cov}_2 \succeq \mathrm{Cov}_1$.

According to Propositions 2–3, two propositions about the relations of different coverings are obtained as follows:

Proposition 4 *Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS and $B_1 \subseteq B_2 \subseteq C$. One has that*

$$\mathrm{Cov}(B_1, p) \succeq \mathrm{Cov}(B_2, p). \tag{13}$$

Proof $\forall n_{(B_2,p)}(x) \in \mathrm{Cov}(B_2, p), \exists n_{(B_1,p)}(x) \in \mathrm{Cov}(B_1, p)$ s.t. $n_{(B_1,p)}(x) \supseteq n_{(B_2,p)}(x)$ according to Proposition 2, so $\mathrm{Cov}(B_1, p) \succeq \mathrm{Cov}(B_2, p)$. □

Proposition 5 *Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS, $B \subseteq C$ and $p_1 \leq p_2$. One has that*

$$\mathrm{Cov}(B, p_1) \preceq \mathrm{Cov}(B, p_2). \tag{14}$$

The proof of Proposition 5 is similar to that of Proposition 4 and is omitted for brevity. It is known from Propositions 4–5 that, in an ECLHDS, the coverings get finer with the addition of features or the decrease in confidence level.

An example is given to illustrate the computation of neighborhoods and coverings as follows:

Example 1 A hybrid decision system is given in Table 2, where $U = \{x_1, x_2, \ldots, x_6\}, C_o = \{a_1, a_4\}$, and $C_u = \{a_2, a_3\}$. Let $k_2 = 0.2, k_3 = 0.15$ in Eq. (6), and let $p_1 = 0.6, p_2 = 0.9$. The upper error bounds for the two numeric features are computed according to Eqs. (4) and (6), which are $e(a_2, 0.6) = 0.0578, e(a_3, 0.6) = 0.0636, e(a_2, 0.9) = 0.113$ and $e(a_3, 0.9) = 0.1243$. Then, the neighborhoods w.r.t. single features are computed according to Definition 5. Accordingly, the neighborhoods w.r.t. multiple features can be obtained by using Eq. (9). Some representative results are shown in Tables 3 and 4.

From Tables 3 and 4, it is easy to know that the neighborhoods for a nominal feature do not change with the data granularity. Besides, the coverings on U can be obtained for the given feature subsets and error confidence levels. For example, $\mathrm{Cov}(\{a_2\}, 0.6) = \{\{x_1\}, \{x_4\}, \{x_2, x_5\}, \{x_3, x_6\}\}$, $\mathrm{Cov}(\{a_2\}, 0.9) = \{\{x_1\}, \{x_3, x_6\}, \{x_2, x_4, x_5\}, \{x_3, x_4, x_6\}, \{x_2, x_4, x_5, x_6\}\}$, $\mathrm{Cov}(\{a_1, a_2\}, 0.6) = \{\{x_1\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_2, x_5\}\}$, $\mathrm{Cov}(\{a_1, a_2\}, 0.9) = \{\{x_1\}, \{x_3\}, \{x_2, x_5\}, \{x_4, x_6\}\}$, and so on. It can be found that $\mathrm{Cov}(\{a_1, a_2\}, 0.6) \preceq \mathrm{Cov}(\{a_1, a_2\}, 0.9) \preceq \mathrm{Cov}(\{a_2\}, 0.9)$; meanwhile, except $\mathrm{Cov}(\{a_2\}, 0.9)$, other three coverings are all the partitions of the universe U.

4 Variable cost-based feature–granularity selection problem for hybrid data

In this section, the optimal feature–granularity selection problem is discussed for hybrid data with measurement errors and variable costs. At the beginning, several kinds of variable cost setting are designed according to reality, in which the relationship among data granularity, test costs and misclassification costs is taken into account. In particular, the data granularity is evaluated by the confidence level of the feature values' measurement errors for numeric features. Then, the computation method of total misclassification cost and average total cost is introduced for any given feature subset and error confidence level. Finally, a formal problem statement of the optimal feature–granularity selection is presented, and the influences of cost setting changes to the selection results are investigated.

4.1 Variable cost settings

As discussed earlier, except that the test costs for nominal features could be seen as constant, both the test costs for numeric features and the misclassification costs are often not fixed but variable in many real applications. On the one hand, collecting fine-grained data is usually more costly than collecting coarse-grained data. So for a numeric feature, the test cost is monotone decreasing with the data granularity, namely, the confidence level of the feature values' measurement errors. On the other hand, for an object, the misclassification cost often increases monotonically with the total test cost. In consideration of these facts, the test cost functions and the misclassification cost functions are designed in what follows.

Let $S = (U, C = C_o \bigcup C_u, D, V, I, p)$ be an ECLHDS, $p \in (0, 0.997]$, $a \in C$. And let tc and TC denote the test cost function and a constant test cost value, respectively. If $a \in C_o$, no matter how the confidence level p changes, the test cost for feature a is immutable, which is denoted as

$$\mathrm{tc}(a, p) = \mathrm{TC}^o(a), \tag{15}$$

where $\mathrm{TC}^o(a) > 0$. If $a \in C_u$, the test cost for feature a is monotone decreasing with the increase in p. A linear test cost function can be given as follows:

$$\mathrm{tc}(a, p) = \mathrm{TC}^u(a) \cdot (1 - \lambda_a p), \tag{16}$$

where $\mathrm{TC}^u(a) > 0$ is the highest test cost for numeric feature a, namely, the test cost paid for obtaining the highest data precision for a, and $\lambda_a \in [0, 1]$ is the test cost adjusting factor. Other forms of test cost function can also be developed for numeric features according to reality. For example, a piecewise constant function of test cost is given as follows:

$$\mathrm{tc}(a, p) = \mathrm{TC}_i^u(a), \ p \in [p_{i-1}, p_i](i = 1, 2, \ldots, m), \tag{17}$$

where m is the number of segments, $p_0 > 0$, $p_m = 0.997$, and $\mathrm{TC}_1^u(a) > \mathrm{TC}_2^u(a) > \cdots > \mathrm{TC}_m^u(a) > 0$. The test cost functions in Eqs. (16) and (17) are both monotone decreasing with the confidence level. The difference between them is that the former is strictly monotone decreasing while the latter is not.

In this paper, it is supposed that the test costs among different features are independent of one another, and each object in the universe has the same total test cost. Then, given a feature subset $B \subseteq C$, the total test cost for each object is

$$\mathrm{tc}(B, p) = \sum_{a \in B} \mathrm{tc}(a, p). \tag{18}$$

Let mc and MC denote the misclassification cost function and a constant misclassification cost value, respectively. Given decision classes k and l, let (k, l) denote the misclassified class pair, namely, misclassifying from class k to class l, and let $\mathrm{mc}(B, p)_{(k,l)}$ denote the corresponding misclassification cost based on the feature–granularity pair (B, p). Obviously, if $k = l$, $\mathrm{mc}(B, p)_{(k,l)} = 0$. If $k \neq l$, $\mathrm{mc}(B, p)_{(k,l)}$ is monotone increasing with $\mathrm{tc}(B, p)$. Analogously to test costs, more than one type of misclassification cost function can be presented according to reality. For instance, the misclassification cost can be given in a form of linear function

$$\mathrm{mc}(B, p)_{(k,l)} = \gamma_{(k,l)} \cdot \mathrm{tc}(B, p), k \neq l, \tag{19}$$

where $\gamma_{(k,l)} > 0$ is the misclassification cost adjusting factor or in a form of piecewise constant function

$$\begin{aligned} \mathrm{mc}(B, p)_{(k,l)} = \mathrm{MC}_j^{(k,l)}, \ &\mathrm{tc}(B, p) \\ &\in [\mathrm{TC}_{j-1}, \mathrm{TC}_j](j = 1, 2, \ldots, n, k \neq l), \end{aligned} \tag{20}$$

where n is the number of segments and $0 < \mathrm{MC}_1^{(k,l)} < \mathrm{MC}_2^{(k,l)} < \cdots < \mathrm{MC}_n^{(k,l)}$.

Continuing with Example 1, the following example is given to illustrate the construction of multiple variable cost settings:

Example 2 According to Example 1, it is known that the hybrid decision system is given in Table 2, where $U = \{x_1, x_2, \ldots, x_6\}$, $C_o = \{a_1, a_4\}$, and $C_u = \{a_2, a_3\}$. Let the confidence level be $p = 0.9$. Suppose that the constant test cost for nominal feature a_1 is $\mathrm{TC}^o(a_1) = 62$, then $\mathrm{tc}(a_1, 0.9) = 62$. As to the numeric feature a_2, the test cost is discussed in two cases as follows:

(1) In the form of linear function:
Assuming that $\mathrm{TC}^u(a_2) = 108$ and $\lambda_{a_2} = 0.3$, one has that $\mathrm{tc}(a_2, 0.9) = 108 \times (1 - 0.3 \times 0.9) = 78.84$; then, $\mathrm{tc}(\{a_1, a_2\}, 0.9) = 62 + 78.84 = 140.84$.

(2) In the form of piecewise constant function: Assuming that

$$\mathrm{tc}(a_2, p) = \begin{cases} 120, & p \in (0, 0.3] \\ 100, & p \in (0.3, 0.7] \\ 80, & p \in (0.7, 0.997] \end{cases},$$

one has that $\mathrm{tc}(a_2, 0.9) = 80$, then $\mathrm{tc}(\{a_1, a_2\}, 0.9) = 62 + 80 = 142$.

Similarly, there are also multiple types of misclassification cost function, but here only the linear form is discussed for brevity. According to the application context of Credit dataset, it is assumed that $\gamma_{(+,-)} = 10$, $\gamma_{(-,+)} = 100$. Then, the misclassification costs based on the piecewise constant-form test costs are $\mathrm{mc}(\{a_2\}, 0.9)_{(+,-)} = 80 \times 10 = 800$, $\mathrm{mc}(\{a_2\}, 0.9)_{(-,+)} = 80 \times 100 = 8000$, $\mathrm{mc}(\{a_1, a_2\}, 0.9)_{(+,-)} = 142 \times 10 = 1420$ and $\mathrm{mc}(\{a_1, a_2\}, 0.9)_{(-,+)} = 142 \times 100 = 14200$.

For an ECLHDS, multiple variable cost settings can be generated by using the above-mentioned cost functions. Note that, in order to save space, only linear functions and piecewise constant functions are discussed above for test costs and misclassification costs. There are also other forms of function for the two kinds of cost in real applications.

4.2 Computation method of average total cost

As pointed out earlier, total cost minimization is the optimization objective of the feature–granularity selection approach. In order to achieve this goal, one needs to know how to compute the total cost. In this subsection, the computation method of average total cost is introduced. As mentioned in Sect. 4.1, for each object in the universe, the total test cost is supposed to be the same and it is equal to $\mathrm{tc}(B, p)$ for the given feature–granularity pair (B, p), so the average total cost is composed of the total test cost for each object and the average misclassification cost for all objects. To obtain the average misclassification cost, it needs to compute the total misclassification cost at first.

Let $S = (U, C = C_o \bigcup C_u, d, V, I, p)$ be an ECLHDS, $x \in U$ and $B \subseteq C$, and let $\mathrm{mc}(x, B, p)$ denote the misclassification cost of x based on B and p. The process of computing the total misclassification cost and then the average total cost is presented as follows. In the process, the objects are categorized by following a rule that minimizes the total misclassification cost.

(1) Classify each object $x \in U$ and compute its misclassification cost $\mathrm{mc}(x, B, p)$. There are two cases according to the neighborhood $n_{(B,p)}(x)$.

A) If $\forall x' \in n_{(B,p)}(x), d(x') = d(x)$ (namely, the neighborhood $n_{(B,p)}(x)$ is consistent), the object x can be categorized into the right class, so $\mathrm{mc}(x, B, p) = 0$.

B) If $\exists x' \in n_{(B,p)}(x), d(x') \neq d(x)$ (namely, the neighborhood $n_{(B,p)}(x)$ is not consistent), since the objects in a neighborhood granule are indistinguishable, they are supposed to have the same decision value in the classification. Then, $n_{(B,p)}(x)$ is categorized into the class which can minimize the total misclassification cost of the objects in it. Accordingly, the misclassification cost of x, namely, $\mathrm{mc}(x, B, p)$, can be obtained.

(2) Calculate the total misclassification cost (TMC) and the average misclassification cost (AMC) for all objects in U, which are

$$\mathrm{TMC}(U, B, p) = \sum_{x \in U} \mathrm{mc}(x, B, p), \tag{21}$$

$$\mathrm{AMC}(U, B, p) = \frac{\mathrm{TMC}(U, B, p)}{|U|}. \tag{22}$$

(3) Calculate the average total cost for all objects in U. As mentioned above, the total test cost for each object is supposed to be the same and is equal to $\mathrm{tc}(B, p)$, so the average total cost (ATC) is

$$\mathrm{ATC}(U, B, p) = \mathrm{tc}(B, p) + \mathrm{AMC}(U, B, p). \tag{23}$$

Let MR denote the corresponding misclassification rate, one has that

$$\mathrm{MR}(U, B, p) = \frac{|\{x_i | \mathrm{mc}(x_i, B, p) > 0\}|}{|U|}. \tag{24}$$

Based on Example 1 and Example 2, the process of computing the average total cost is displayed in what follows.

Example 3 Let $p = 0.9$, the average total costs are computed for $B_1 = \{a_2\}$ and $B_2 = \{a_1, a_2\}$, respectively. First, it is known from Table 4 that $n_{(B_1,p)}(x_1) = \{x_1\}$, $n_{(B_1,p)}(x_2) = n_{(B_1,p)}(x_5) = \{x_2, x_4, x_5\}$, $n_{(B_1,p)}(x_3) = \{x_3, x_6\}$, $n_{(B_1,p)}(x_4) = \{x_2, x_4, x_5, x_6\}$, $n_{(B_1,p)}(x_6) = \{x_3, x_4, x_6\}$. Now the misclassification costs are computed for the six objects according to their neighborhoods. Since $n_{(B_1,p)}(x_1)$ is consistent, $\mathrm{mc}(x_1, B_1, p) = 0$. For x_2 and x_5, if their neighborhood $\{x_2, x_4, x_5\}$ is categorized into class "+," x_4 and x_5 will be misclassified, and the misclassification cost for $\{x_2, x_4, x_5\}$ is 8000×2; conversely, if $\{x_2, x_4, x_5\}$ is categorized into class "−," x_2 will be misclassified and the misclassification cost for $\{x_2, x_4, x_5\}$ is 800. Thus, class "−" is chosen for the neighborhood granule $\{x_2, x_4, x_5\}$ to obtain a lesser misclassification cost. In this case, x_2 is classified incorrectly, $\mathrm{mc}(x_2, B_1, p) = 800$; x_5 is classified correctly, $\mathrm{mc}(x_5, B_1, p) = 0$. Similarly, x_3, x_4 and x_6 are also categorized into class "−," and one has that $\mathrm{mc}(x_3, B_1, p) = 800$, $\mathrm{mc}(x_4, B_1, p) = \mathrm{mc}(x_6, B_1, p) = 0$. Then, according to Eqs. (21)-(24) and Example 2, one has that $\mathrm{TMC}(U, B_1, p) = 0 + 800 + 800 + 0 + 0 + 0 = 1600$, $\mathrm{AMC}(U, B_1, p) = \frac{1600}{6} \approx 266.67$, $\mathrm{ATC}(U, B_1, p) = \mathrm{tc}(B_1, p) + \mathrm{AMC}(U, B_1, p) = 80 + 266.67 = 346.67$ and $\mathrm{MR}(U, B_1, p) = \frac{1}{3}$.

Next, it is known from Table 4 that $n_{(B_2,p)}(x_1) = \{x_1\}, n_{(B_2,p)}(x_2) = n_{(B_2,p)}(x_5) = \{x_2, x_5\}, n_{(B_2,p)}(x_3) = \{x_3\}, n_{(B_2,p)}(x_4) = n_{(B_2,p)}(x_6) = \{x_4, x_6\}$. Analogously, it could be obtained that $\mathrm{mc}(x_1, B_2, p) = \mathrm{mc}(x_3, B_2, p) = \mathrm{mc}(x_4, B_2, p) = \mathrm{mc}(x_5, B_2, p) = \mathrm{mc}(x_6, B_2, p) = 0$ and $\mathrm{mc}(x_2, B_2, p) = 1420$. Then, according to Eqs. (21)–(24) and Example 2, one has that $\mathrm{TMC}(U, B_2, p) = 0 + 1420 + 0 + 0 + 0 + 0 = 1420$, $\mathrm{AMC}(U, B_2, p) = \frac{1420}{6} \approx 236.67$, $\mathrm{ATC}(U, B_2, p) = \mathrm{tc}(B_2, p) + \mathrm{AMC}(U, B_2, p) = 142 + 236.67 = 378.67$ and $\mathrm{MR}(U, B_2, p) = \frac{1}{6}$.

It is found From Example 3 that, with the addition of attributes, the total test cost often increases while the misclassification rate and the average misclassification cost often decrease. Besides, in fact a computation method of average total cost was introduced in Zhao and Zhu (2014), in which $\mathrm{TMC}(U, B, p) = \sum_{X \in \mathrm{Cov}(B,p)} \mathrm{mc}(X, B, p)$, i.e., the total misclassification cost is equal to the sum of the misclassification costs of each neighborhood granule in the covering. If using the existing method, the obtained $\mathrm{TMC}(U, B_2, p)$ and $\mathrm{ATC}(U, B_2, p)$ are the same as the above ones; however, $\mathrm{TMC}(U, B_1, p) = 800 \times 4 = 3200$ and $\mathrm{ATC}(U, B_1, p) = 80 + \frac{3200}{6} \approx 613.33$, both of which are more than those obtained by the proposed method. The reason of the excess is that $\mathrm{Cov}(B_1, p)$ is not a partition of the universe; the misclassification costs of x_2 and x_3 are repetitively computed in the previous method because both of x_2 and x_3 belong to two neighborhood granules. In comparison, the proposed method of computing average total costs is appropriate for any kind of coverings, while the method in Zhao and Zhu (2014) is only suitable for the partition case. Hence, the proposed computation method is more general.

4.3 Influences of cost setting changes to the optimal feature–granularity selection

As discussed earlier, this paper aims to find a pair of optimal feature subset and optimal data granularity to minimize the total cost. In real applications, both test costs and misclassification costs could be given in multiple different forms. And the data granularity for the hybrid data is measured with the error confidence level of the feature values of the numerical features. Based on these considerations, the optimal feature–granularity selection problem could be defined in the following optimization form:

Problem 1 The optimal feature–granularity selection problem.
Input: an ECLHDS $S = (U, C = C_o \bigcup C_u, D, V, I, p)$; the test cost function for each feature and the misclassification cost function for each misclassified class pair;
Output: the optimal feature subset R^* and the optimal confidence level p^*;
Optimization objective: $\min(\mathrm{ATC}(U, R, p))$.

From Problem 1, it is found that the optimal feature-granularity selection problem could also be seen as a cost-sensitive variable granularity–feature selection problem.

Generally speaking, for a given setting of test costs and misclassification costs, people cannot know which data granularity and feature subset will be optimal by intuition. However, it can be found that there are some interesting rules about the relation of costs, feature selection and the misclassification rate formulated in Eq. (24). It is notable that, for simplicity, sometimes the test cost for each feature and the misclassification cost for each misclassified class pair are called as individual test cost and individual misclassification cost, respectively. On the one hand, if the test cost functions remain unchanged while the individual misclassification costs increase, the misclassification costs play a larger role in feature–granularity selection than before. Usually, to avoid a substantial increase in total misclassification cost, more necessary features are chosen. Hence, the total test cost and the average total cost often become large, while the misclassification rate gets small. On the other hand, if the misclassification cost functions stay the same while the individual test costs increase, the test costs play a larger role than before. Generally, to restrict the enlargement of total test cost, less features are selected. Therefore, the misclassification rate and the average misclassification cost usually become large, which poses the increase in average total cost. An illustrative example is given as follows, in which the process of computing neighborhoods, total test costs and average misclassification costs is omitted for brevity.

Example 4 An ECLHDS is constituted by the hybrid decision system shown in Table 2 and the error confidence level $p = 0.9$. Let $B \subseteq C$.

(1) Let test cost functions be fixed. For instance, let $\mathrm{TC}^o(a_1) = 50, \mathrm{TC}^u(a_2) = 70, \mathrm{TC}^u(a_3) = 80, \mathrm{TC}^o(a_4) = 60, \lambda_{a_2} = 0.2$ and $\lambda_{a_3} = 0.3$, then it can be obtained that $\mathrm{TC}(a_1, 0.9) = 50, \mathrm{tc}(a_2, 0.9) = 57.4, \mathrm{tc}(a_3, 0.9) = 58.4$ and $\mathrm{tc}(a_4, 0.9) = 60$ according to Eqs. (15)–(16). As shown below, two different misclassification cost functions are used, respectively,

(A) If $\mathrm{tc}(B, P) < 100$, let $\mathrm{mc}(B, p)_{(+,-)} = 200$, $\mathrm{mc}(B, p)_{(-,+)} = 2000$; and if $\mathrm{tc}(B, P) \geq 100$, let $\mathrm{mc}(B, p)_{(+,-)} = 2 \cdot \mathrm{tc}(B, P), \mathrm{mc}(B, p)_{(-,+)} = 20 \cdot \mathrm{tc}(B, P)$. Through computation and comparison, one has that $\min_{B \subseteq C}(\mathrm{ATC}(U, B, 0.9)) = \mathrm{ATC}(U, \{a_2\}, 0.9) \approx 124.07$, and the corresponding misclassification rate $\mathrm{MR} = \frac{1}{3}$.

(B) If $\mathrm{tc}(B, P) < 100$, let $\mathrm{mc}(B, p)_{(+,-)} = 500$, $\mathrm{mc}(B, p)_{(-,+)} = 5000$; and if $\mathrm{tc}(B, P) \geq 100$, let $\mathrm{mc}(B, p)_{(+,-)} = 5 \cdot \mathrm{tc}(B, P), \mathrm{mc}(B, p)_{(-,+)} = 50 \cdot \mathrm{tc}(B, P)$. Through computation and comparison, one has that $\min_{B \subseteq C}(\mathrm{ATC}(U, B, 0.9)) = \mathrm{ATC}(U, \{a_1, a_2, a_4\}, 0.9) \approx 167.4$, $\mathrm{MR} = 0$.

It is known from the results that, with the increase in individual misclassification costs, the number of selected features and the minimal average total cost often grow; meanwhile, the corresponding misclassification rate often drops.

(2) Let misclassification cost functions be the same as those in (1-B), while the individual test costs be double of those in (1), namely $\mathrm{tc}(a_1, 0.9) = 100, \mathrm{tc}(a_2, 0.9) = 114.8, \mathrm{tc}(a_3, 0.9) = 116.8, \mathrm{tc}(a_4, 0.9) = 120$ by setting $\mathrm{TC}^o(a_1) = 100, \mathrm{TC}^u(a_2) = 140, \mathrm{TC}^u(a_3) = 160, \mathrm{TC}^o(a_4) = 120$. Through computation and comparison, it can be obtained that $\min_{B \subseteq C}(\mathrm{ATC}(U, B, 0.9)) = \mathrm{ATC}(U, \{a_2\}, 0.9) \approx 306.13$, $\mathrm{MR} = \frac{1}{3}$. Compared with the results in (1-B), it is known that with the increase in individual test costs, the number of selected features often decreases, while the minimal average total cost and the corresponding misclassification rate often enlarge.

(3) In (1)–(2), the individual misclassification costs are set to be variable. Here the constant case is investigated further. For example, let $\mathrm{mc}(B, p)_{(+,-)} = 500, \mathrm{mc}(B, p)_{(-,+)} = 5000$. If the individual test costs in (1) are adopted, the results are the same as those in (1-B), and if the individual test costs in (2) are adopted, one has that $\min_{B \subseteq C}(\mathrm{ATC}(U, B, 0.9)) = \mathrm{ATC}(U, \{a_2\}, 0.9) \approx 281.47$, $\mathrm{MR} = \frac{1}{3}$. The results validate the observations in (2) again.

Note that although Example 4 only displays the influences of cost setting changes for the confidence level $p = 0.9$ to save the space, the obtained rules are also valid for other confidence levels. Hence, the rules are suitable for the optimal feature–granularity selection. In addition, the rules are applicable not only for the binary classification (as shown in Example 4), but also for the multiple classification, which is verified in Sect. 6.2.

5 Algorithm design

In this section, an algorithm is proposed to solve the optimal feature–granularity selection problem, so the algorithm is called the optimal feature–granularity selection (OFGS) algorithm. The algorithm is composed of Algorithm 1 and Algorithm 2, in which Algorithm 2 is invoked by Algorithm 1. Note that $D = \{d\}$ in the input of Algorithm 1, namely, the OFGS algorithm deals with the hybrid decision systems which have only one decision attribute. This kind of decision systems is widespread in applications. If there is more than one decision attribute in a decision system, one could construct multiple new decision systems, with each having exactly one decision attribute.

Algorithm 1 The optimal feature–granularity selection (OFGS) algorithm.

Input: An ECLHDS $S = (U, C = C_o \cup C_u, D = \{d\}, V, I, p)$; the confidence level's minimal value p_0 and the step size s; the test cost function for each feature and the misclassification cost function for each misclassified class pair.

Output: The globally optimal feature subset R^* and optimal confidence level p^* with minimal average total cost $gmtc$. The three variables are all global variables.

```
1: $gmtc = +\infty$; //$gmtc$ is the globally minimal average total cost
2: for ($p = p_0$; $p \le 0.997$; $p = p + s$) do
3:     for ($i = 1$; $i \le |C|$; $i++$) do
4:         Compute $tc(a_i, p)$ according to the test cost function of feature $a_i$;
5:     end for
6:     $cmtc = +\infty$; //$cmtc$ is currently minimal average total cost
7:     $B = \emptyset$; //$B$ is currently selected feature subset
8:     $cttc = 0$; //$cttc$ is current total test cost
9:     backtracking($cttc$, $B$, 1); //The output of the backtracking is $R$ and $cmtc$
10:    if ($cmtc < gmtc$) then
11:        $gmtc = cmtc$; //Update the globally minimal average total cost
12:        $p^* = p$; //Update the optimal confidence level
13:        $R^* = R$; //Update the globally optimal feature subset
14:    end if
15: end for
```

In the optimal feature–granularity selection algorithm, the error confidence level p is tried from the minimal value $p_0 > 0$ to the maximal value 0.997 with a step size $s > 0$, which is shown in line 2 of Algorithm 1. The users can choose p_0 and s according to their preference. For each tried confidence level, the minimal average total cost and the corresponding feature subset are obtained by running Algorithm 2 which is essentially a backtracking algorithm. Then, the obtained average total costs are compared among different confidence levels to choose the minimal one. Naturally, the corresponding confidence level is the optimal data granularity, and the optimal feature subset is obtained accordingly. In the process, the individual test costs and the individual misclassification costs are computed according to Sect. 4.1, and the total misclassification costs and the average total costs are computed according to Sect. 4.2. In particular, three pruning techniques are used in Algorithm 2 to improve the efficiency. Firstly, as shown in line 1, the backtracking algorithm begins with current level test index lower bound l rather than 1. As the backtracking algorithm goes on, the lower bound raises. Then, as shown in lines 2–4 and lines 7–9, the other two pruning techniques are used to discard the feature subsets whose test costs are too large.

Algorithm 2 The backtracking algorithm.

Input: The current total test cost $cttc$, currently selected tests B, and current level test index lower bound l.
Output: The currently optimal feature subset R and minimal average total cost $cmtc$, both of which are global variables.
Method: Backtracking

```
for ($i = l; i \le |C|; i++$) do
    if ($tc(a_i, p) \ge cmtc$) then
        continue; //Prune for too large test cost
    end if
    $B = R \cup \{a_i\}$;
    $tc(B, p) = cttc + tc(a_i, p)$; //Compute the temporary total test cost
    if ($tc(B, p) \ge cmtc$) then
        continue; //Prune for too large total test cost
    end if
    for ($j = 1; j \le |V_d|; j++$) do
        for ($k = 1; k \le |V_d|; k++$) do
            Compute $mc(B, p)_{(j,k)}$ according to the misclassification cost function for misclassified class pair $(j, k)$;
        end for
    end for
    Compute the total misclassification cost $TMC(U, B, p)$ and the average total cost $ATC(U, B, p)$;
    if ($ATC(U, B, p) < cmtc$) then
        $cttc = tc(B, p)$; //Update the current total test cost
        $cmtc = ATC(U, B, p)$; //Update the currently minimal average total cost
        $R = B$; //Update the currently optimal feature subset
    end if
    backtracking($cttc, B, i + 1$); //Next level backtracking
end for
```

6 Experiments

In this section, the following questions are investigated by experimentation.

(1) Why is an error confidence level rather than an error range employed to measure the data granularity?
(2) Is the optimal feature–granularity selection algorithm effective?
(3) How do the feature–granularity selection results change with different cost settings?
(4) Is there an optimal value or a rational value range for the data granularity?

6.1 Data generation

The experiments are conducted on sixteen datasets from the UCI library. The basic information of these datasets is shown in Table 5, where "Nominal" and "Numeric" represent the numbers of nominal features and numeric features, respectively. For brevity, in the following context CAD-diagnosis, Cylinder-bands, Diabetic-retinopathy, Heart and Mice-protein are abbreviated as CAD, Cylinder, Diabetic, Heart and Mice, respectively. The data items of numeric features are normalized into [0, 1], and those of nominal features remain unchanged. If an object has multiple missing values, it will be deleted from the dataset; otherwise, the missing values of a nominal feature are set to be random feature values of the feature, and those of a numeric feature are directly set to be 0.5.

Since there are not test costs and misclassification costs in these UCI datasets, the experiments start from generating the two types of cost for the datasets according to Sect. 4.1. First, there are two cases for generating the parameters for test cost functions. One is regarding datasets Bridges, Diabetes, Heart and Wine, each having less than 15 features. For these datasets, the constant test costs $\mathrm{TC}^o(a)$ (shown in Eq. (15))

Table 5 Data information

Dataset	Domain	Samples	Features	Nominal	Numeric	Classes
Bridges	Engineering	108	11	7	4	7
CAD-diagnosis	Clinic	303	58	24	34	2
Credit	Finance	690	15	9	6	2
Cylinder-bands	Physics	430	36	16	20	2
Diabetes	Clinic	768	8	0	8	2
Diabetic-retinopathy	Clinic	1151	19	3	16	2
German	Finance	1000	20	13	7	2
Heart	Clinic	303	13	8	5	5
Hepatitis	Clinic	155	19	13	6	2
Image	Graphics	2310	18	0	18	7
Ionosphere	Physics	351	34	0	34	2
Mice-protein	Biology	1080	80	3	77	8
Sonar	Physics	208	60	0	60	2
Wdbc	Clinic	569	30	0	30	2
Wine	Agriculture	178	13	0	13	3
Wpbc	Clinic	198	33	0	33	2

Table 6 Some results of average error bounds for numeric features, where p is the error confidence level

p	0.2	0.4	0.6	0.8	0.997
Bridges	0.0025	0.0052	0.0084	0.0128	0.03
CAD	0.003	0.0062	0.0099	0.0151	0.0353
Credit	0.0038	0.0078	0.0125	0.019	0.0445
Cylinder	0.0029	0.006	0.0096	0.0146	0.0342
Diabetic	0.0035	0.0073	0.0117	0.0177	0.0415
Diabetes	0.0031	0.0063	0.0102	0.0155	0.0362
German	0.0032	0.0066	0.0106	0.0161	0.0377
Heart	0.0028	0.0058	0.0093	0.0142	0.0333

for nominal features, as well as the highest test costs $\mathrm{TC}^u(a)$ (in Eq. (16)) and the piecewise constant test costs $\mathrm{TC}_i^u(a)$ (in Eq. (17)) for numeric features, are all set to be uniformly distributed random integers lying within [20, 100]. The other is corresponding to other twelve datasets, each having no less than 15 features. For these datasets, $\mathrm{TC}^o(a), \mathrm{TC}^u(a)$ and $\mathrm{TC}_i^u(a)$ are all set to be lying within [20, 200]. Note that, in the two cases, $\mathrm{TC}_m^u(a) > \mathrm{TC}_n^u(a)$ if $m < n$, and the test cost adjusting factors λ_a (in Eq. (16)) are set to be random decimals lying within [0, 1]. Then, as for the parameters for misclassification cost functions, the adjusting factors $\gamma_{(m,n)}$ (in Eq. (19)) and the piecewise constant misclassification costs $\mathrm{MC}_j^{(k,l)}$ (in Eq. (20)) are, respectively, set to be integers locating in [10, 200] and [500, 25, 000], where $\mathrm{MC}_m^{(k,l)} < \mathrm{MC}_n^{(k,l)}$ if $m < n$. It is notable that, in order to be close to reality, the misclassification cost parameters are set carefully. For example, people are described by a set of attributes as good or bad credit risks in German dataset. It is worse to classify a customer as good when he/she is bad than to classify a customer as bad when he/she is good, so the misclassification cost of the former is higher. Finally, the error bound adjusting factors k_a (in Eq. (6)) are uniformly set to be 0.05 for convenience. If they are not the same among different numeric features, it can be found that the results are similar by experimentation.

6.2 Experimental results and the analyses

This subsection studies the performance of the feature-granularity selection algorithm and the influence of different cost settings to the selection results from multiple perspectives.

1. Error confidence level versus error range

For the datasets, the average error bounds of numeric features are computed based on different error confidence

Table 7 An exemplary feature–granularity selection result of Heart dataset with $(p_0, s) = (0.1, 0.1)$ and a l–p-type cost setting, where p_0 and s are the minimal value and the step size of error confidence level, respectively

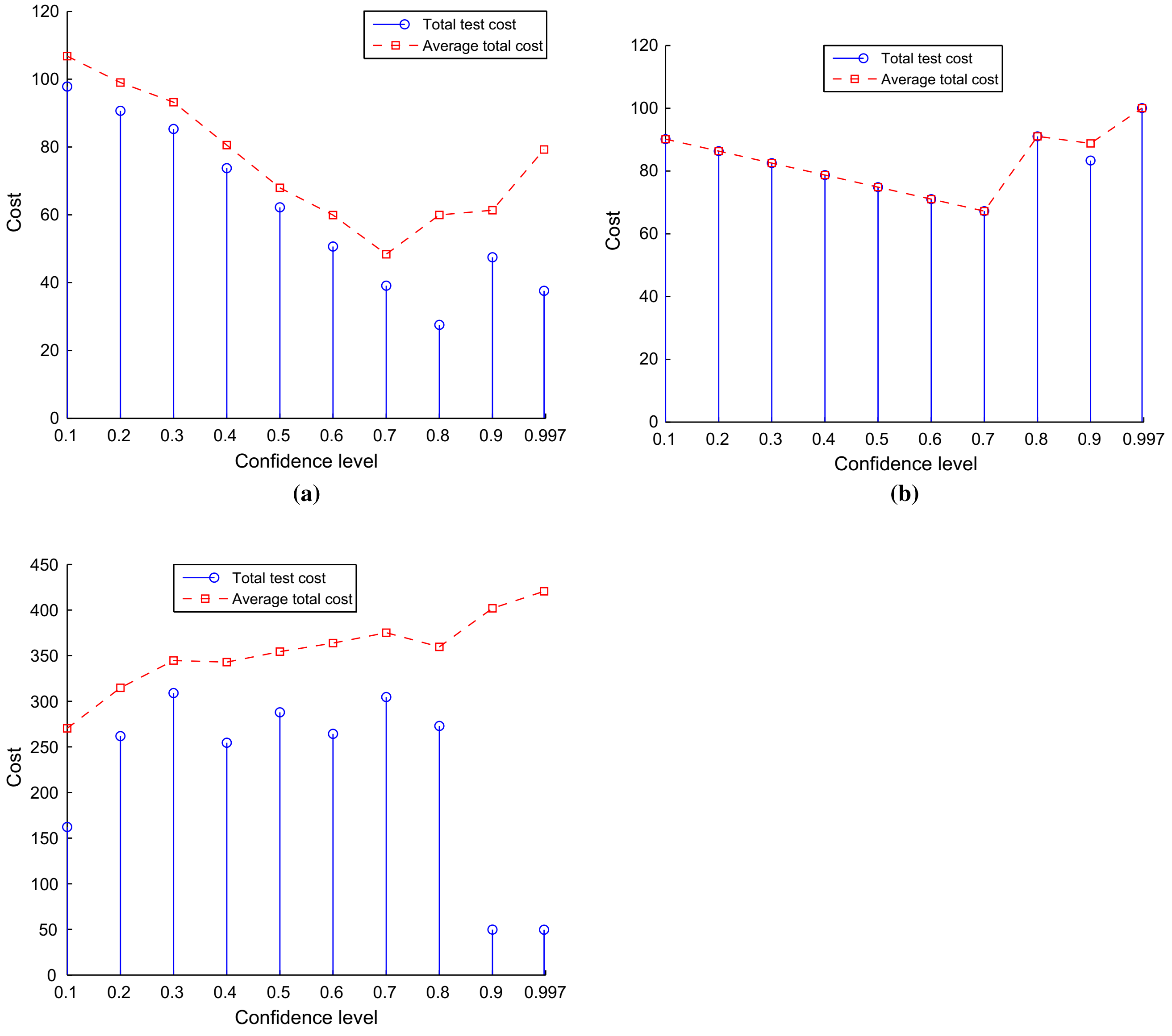
p	AEB	TTC	AMC	ATC	Feature subset	Feature types
0.1	0.0014	156.1648	0	156.1648	{4,8,11,12}	[u,u,o,o]
0.2	0.0028	95.0792	31.3531	126.4323	{8,10}	[u,u]
0.3	0.0043	93.9695	52.8053	146.7748	{4,8}	[u,u]
0.4	0.0058	98.8584	24.7525	123.6108	{8,10,12}	[u,u,o]
0.5	0.0075	97.6492	14.8515	112.5007	{4,8,12}	[u,u,o]
0.6	0.0093	92.8841	14.8515	107.7356	{4,8,10}	[u,u,u]
0.7	0.0115	97.4039	9.901	107.3049	{1,8,10}	[u,u,u]
0.8	0.0142	88.8788	14.8515	103.7302	{4,8,10,12}	[u,u,u,o]
0.9	0.0183	89.2034	8.2508	**97.4542**	{1,4,8,10}	[u,u,u,u]
0.997	0.0333	92.4953	18.1518	110.6471	{1,4,8,10,12}	[u,u,u,u,o]

Table 8 An exemplary feature–granularity selection result of Heart with $(p_0, s) = (0.1, 0.1)$ and a p–l-type cost setting

p	AEB	TTC	AMC	ATC	Feature subset	Feature types
0.1	0.0014	139.2904	4.5875	**143.8779**	{5,8,13}	[u,u,o]
0.2	0.0028	139.2904	18.3498	157.6403	{5,8,13}	[u,u,o]
0.3	0.0043	154.4463	15.2475	169.6938	{5,8,12,13}	[u,u,o,o]
0.4	0.0058	154.4463	15.2475	169.6938	{5,8,12,13}	[u,u,o,o]
0.5	0.0075	133.4027	21.9472	155.3499	{5,8,10}	[u,u,u]
0.6	0.0093	187.6527	0	187.6527	{3,5,8,10,12}	[o,u,u,u,o]
0.7	0.0115	166.6138	5.4785	172.0924	{3,5,8,10,12}	[o,u,u,u,o]
0.8	0.0142	184.4388	6.0726	190.5114	{3,5,8,10,12,13}	[o,u,u,u,o,o]
0.9	0.0183	145.5749	52.6403	198.2152	{3,5,8,10,12}	[o,u,u,u,o]
0.997	0.0333	48.5888	188.1188	236.7076	{8,13}	[u,o]

Table 9 An exemplary feature–granularity selection result of Heart with $(p_0, s) = (0.08, 0.13)$ and a l–l-type cost setting

p	AEB	TTC	AMC	ATC	Feature subset	Feature types
0.08	0.0011	116.7197	22.9703	139.69	{1,5,9}	[u,u,o]
0.21	0.003	109.3174	35.9736	145.291	{1,5,9}	[u,u,o]
0.34	0.0049	101.915	46.6667	148.5817	{1,5,9}	[u,u,o]
0.47	0.007	129.4989	8.5149	138.0137	{1,4,5}	[u,u,u]
0.6	0.0093	124.7225	8.1848	132.9074	{1,5,9,10}	[u,u,o,u]
0.73	0.0122	115.6243	3.7954	**119.4197**	{1,4,5,9}	[u,u,u,o]
0.86	0.0164	127.5937	25.1485	152.7422	{1,4,5,9,10}	[u,u,u,o,u]
0.99	0.0286	143.6598	56.6337	200.2935	{1,4,5,9,10,11}	[u,u,u,o,u,o]



(a)

(b)

(c)

Fig. 3 Total test costs and average total costs: **a** Bridges, **b** CAD, **c** Credit

levels according to Eqs. (4) and (6). Some results are listed in Table 6, from which it is known that the average error bounds are monotonically increasing with the increase in confidence level. More importantly, the error bounds with the same confidence level usually vary between different datasets, so the error confidence level, rather than the error

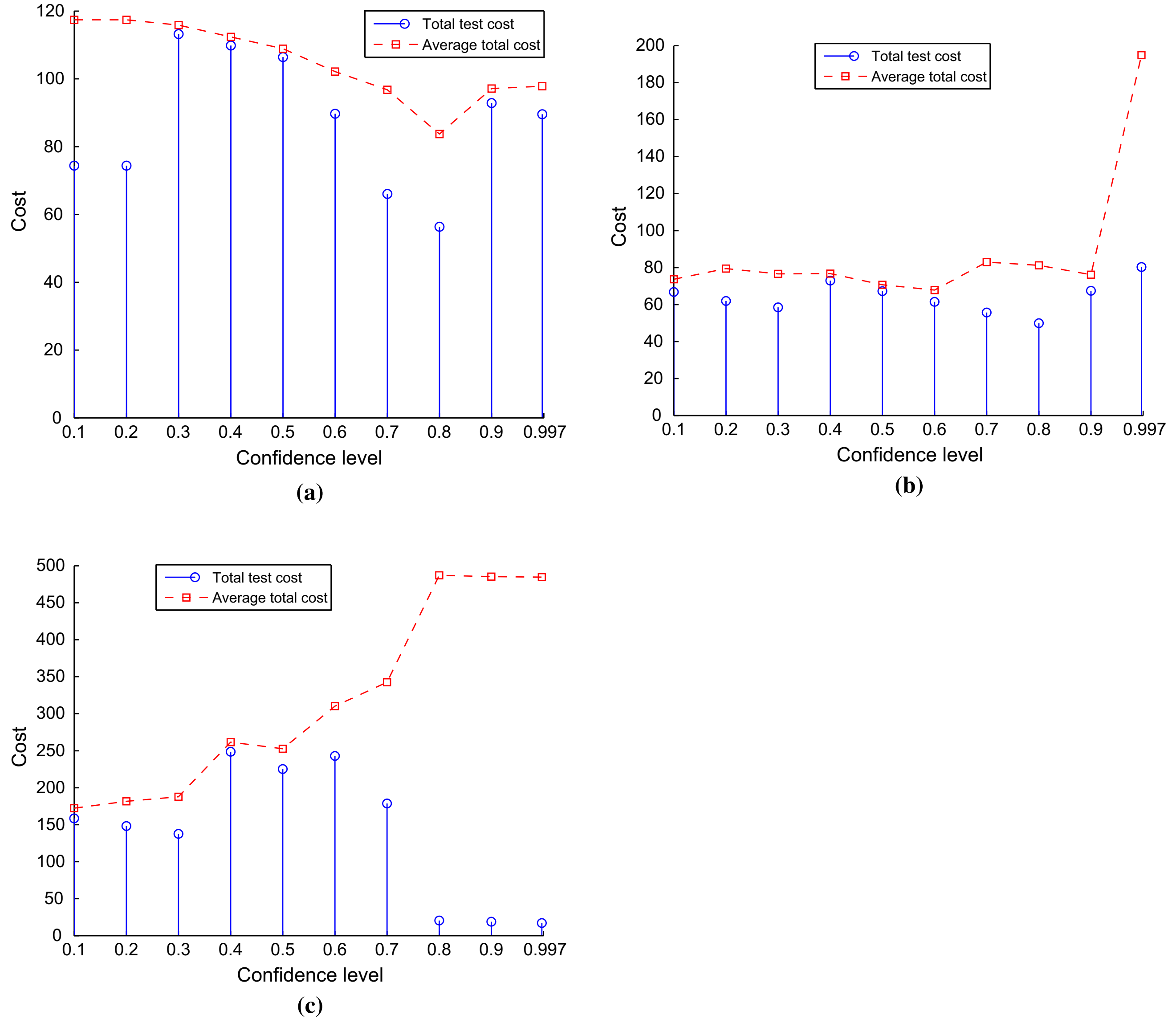


Fig. 4 Total test costs and average total costs: **a** Cylinder, **b** Diabetes, **c** Diabetic

range or the error interval, is chosen to evaluate the data granularity.

2. Some representative results of optimal feature–granularity selection

For each dataset, the optimal feature–granularity selection algorithm is run with four different types of cost setting. Concretely, let l and p denote the linear function and the piecewise constant function, respectively, then l–l-type means that both test costs and misclassification costs are linear functions. The other three types l–p, p–l and p–p are done in the same manner. For each dataset and each type of cost setting, the algorithm is run with multiple different cost parameter values whose generating methods are introduced in Sect. 6.1. It is found from the experiments that the experimental results have no distinct difference between different types of cost setting. Three representative results of Heart are shown in Tables 7, 8 and 9, where AEB denotes the average error bound for all numeric features, TTC denotes the total test cost for each object, AMC and ATC, respectively, denote the average misclassification cost and the average total cost for all objects, the boldface numbers in the fifth columns of the tables are the minimal average total costs, and the integers in the sixth columns are the indexes of selected features, with "o" and "u" in the seventh columns, denoting that the corresponding feature is nominal and numeric, respectively. In the experiments, for numeric feature a, the piecewise constant test cost is $\text{TC}_i^u(a)$ ($\text{TC}_m^u(a) > \text{TC}_n^u(a)$ if $m < n$) when $p \in [0.2i - 0.2, 0.2i)$, $i = 1, 2, \ldots, 5$, namely, the test cost is set to be a 5-segment function. The piecewise constant misclassification cost functions are set in a similar way.

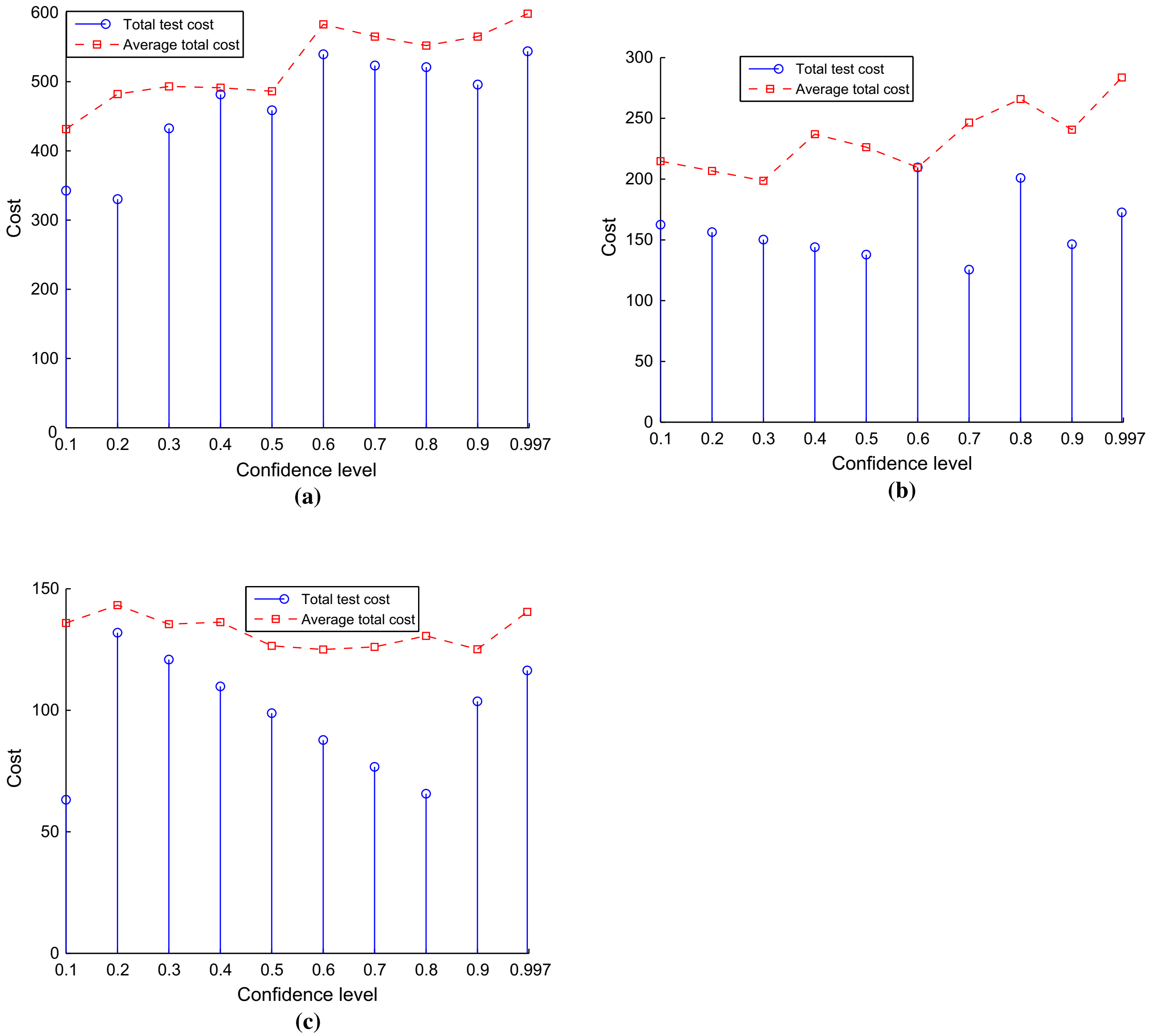
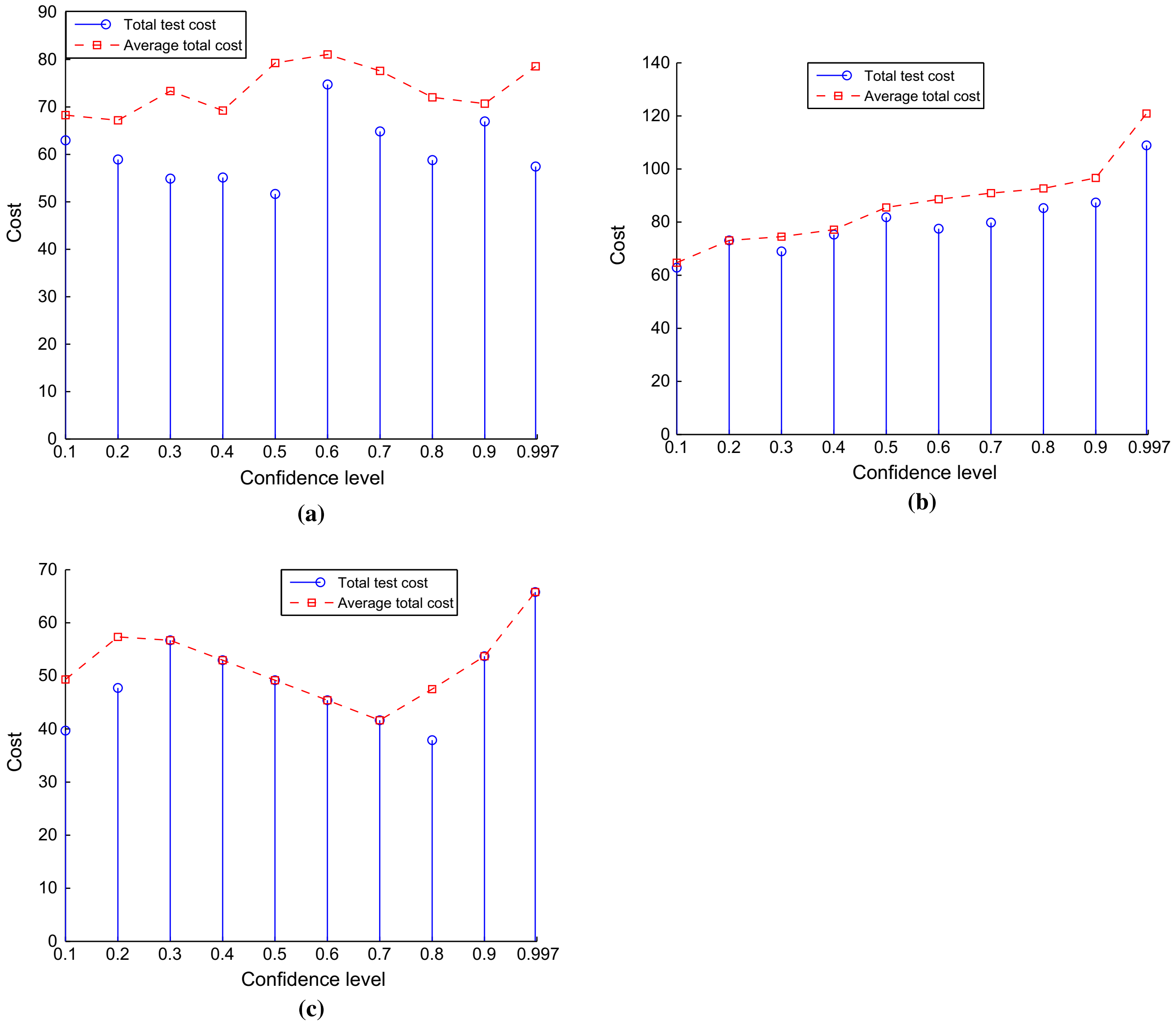


Fig. 5 Total test costs and average total costs: **a** German, **b** Hepatitis, **c** Image

The following observations could be made from the three tables:

(1) With the increase in error confidence level, the error bound increases, but there are not certain change rules for TTC, AMC and ATC. In most cases, the three types of cost change between two adjacent confidence levels. Even if the individual test costs and/or the individual misclassification costs are in the form of piecewise constant function (e.g., Tables 7, 8, 9), only a part of TTC, AMC and/or ATC may be equal. So people cannot know which confidence level is optimal in advance. By using the feature–granularity selection algorithm, the minimal average total cost can be obtained. Consequently, the optimal feature subset and optimal confidence level can be known. This validates the effectiveness of the proposed algorithm. A good trade-off among feature dimension reduction, data granularity selection and total cost minimization can be achieved by the algorithm.

(2) Although there are some differences between the three tables, in general the number of selected features (i.e., the dimension of selected feature subset) increases gradually as the error confidence level becomes large. In particular, if the number of selected features gets small suddenly in the process, the average misclassification cost will become significantly high, especially for $p = 0.997$ whose quantile is much larger than those of $p \leq 0.9$ (see Table 1). The reason is that, usually more features are needed to avoid the increase in misclassification rate which arises from large error range. If there are not enough features to discriminate the incon-



(a)

(b)

(c)

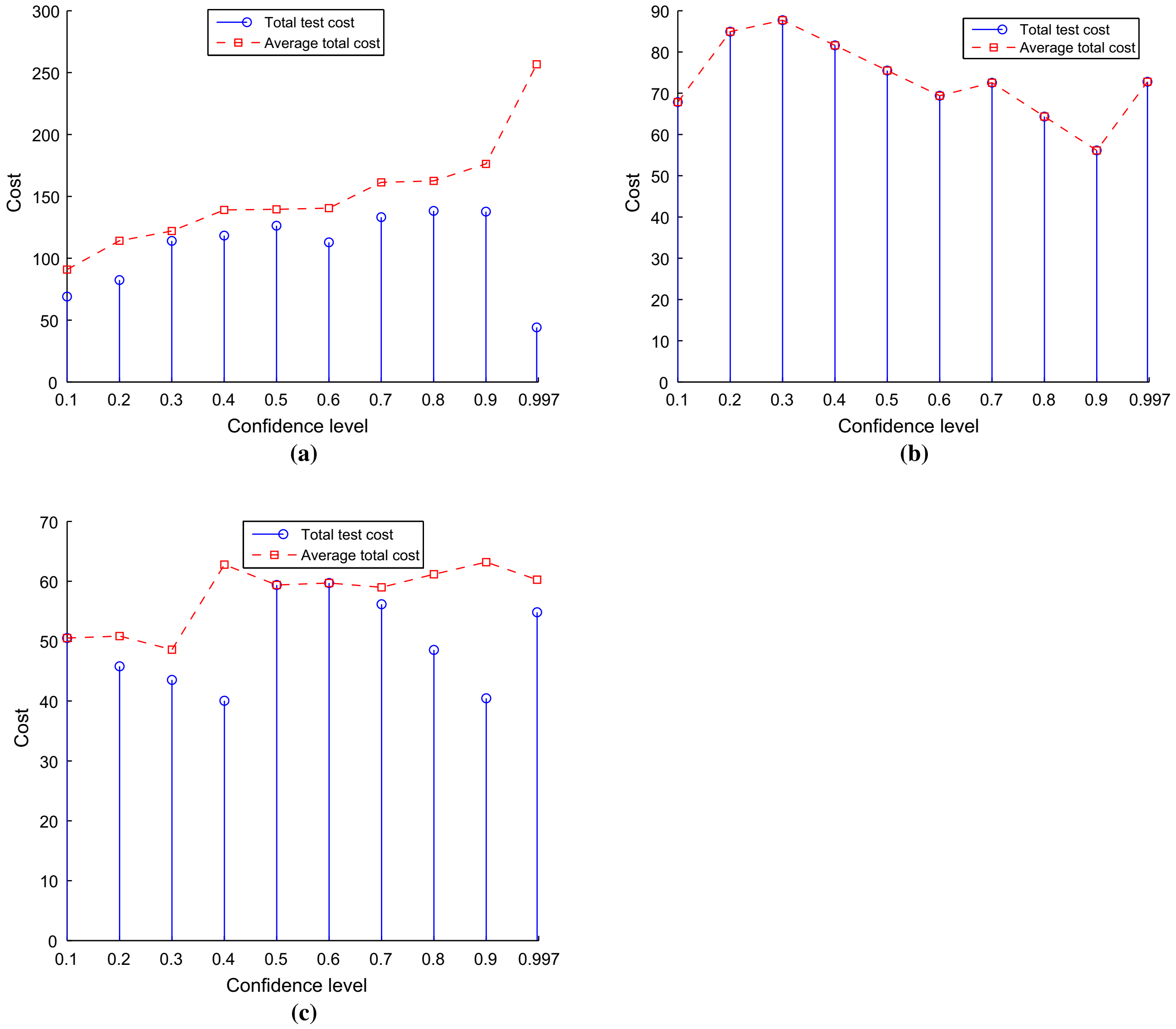
Fig. 6 Total test costs and average total costs: **a** Ionosphere, **b** Mice, **c** Sonar

sistent objects in the neighborhood granules effectively, the total misclassification cost will be large.

To visualize the cost changes, the total test costs and the average total costs, which are obtained under some *l*–*l*-type cost settings for other fifteen datasets, are plotted in Figs. 3, 4, 5, 6, 7. Combining the five figures with Tables 7, 8 and 9, it is found that there is not a universally optimal value or a rational value range for the error confidence level, namely, the data granularity. In fact, with the increase in confidence level, the average total cost may grow, drop or even stay the same. Its change depends on the decision system and the cost setting, and there is not a certain rule for the change. By using the proposed algorithm, the minimal average total cost and the optimal feature–granularity pair can be obtained, which verifies the effectiveness of the algorithm.

3. Comparisons with multiple existing feature selection algorithms

As mentioned in Sect. 4.3, the proposed feature–granularity selection approach is essentially a cost-sensitive variable granularity–feature selection approach. To further investigate its performance, the corresponding optimal feature-granularity selection (OFGS) algorithm is compared with eight existing feature selection algorithms, including a mutual information-based (MI) algorithm (Doquire and Verleysen 2011), a manifold learning-based (ML) algorithm (Huang and Zhu 2017), a random forest-based (RF) algorithm (Zhou et al. 2016), a particle swarm optimization-based (PSO) algorithm (Zhang 2017), a histogram comparison-based (HC) algorithm (Weiss et al. 2013), a rough set-based (RS) algorithm (Zhao et al. 2013), a rough set and Lapla-



(a)

(b)

(c)

Fig. 7 Total test costs and average total costs: **a** Wdbc, **b** Wine, **c** Wpbc

cian score-based (RSLS) algorithm (Yu and Zhao 2018) and a $l_{2,1}$-norm-based (LN) algorithm (Zhao and Yu 2019). It is notable that originally the cost settings are different between these algorithms; even the first algorithm does not take cost information into consideration. Moreover, all the existing algorithms consider only features but not the data granularity. To facilitate the comparisons, in the eight previous algorithms the cost settings are supposed to be the same as those discussed in Sect. 4.1, except that the individual test costs do not change with the error confidence level because the date granularity is not considered in these algorithms. For each dataset, the nine compared algorithms are run with multiple groups of cost parameter values whose generating methods are introduced in Sect. 6.1. Meanwhile, the calculation method presented in Sect. 4.2 is used to compute the average total costs. A group of representative experimental results is shown in Fig. 8.

From the figure, it can be found that the OFGS algorithm significantly outperforms the eight previous feature selection algorithms on minimizing average total costs. The reason is that, as mentioned above, the existing algorithms take into account only features but not the data granularity. Consequently, the individual test costs do not reduce with the increase in the error confidence level, which results in relative large values of total test cost, average misclassification cost and average total cost, while the OFGS algorithm selects not only the optimal feature subset but also the optimal data granularity to minimize the average total cost. As discussed in Sect. 5, the OFGS algorithm invokes a back-

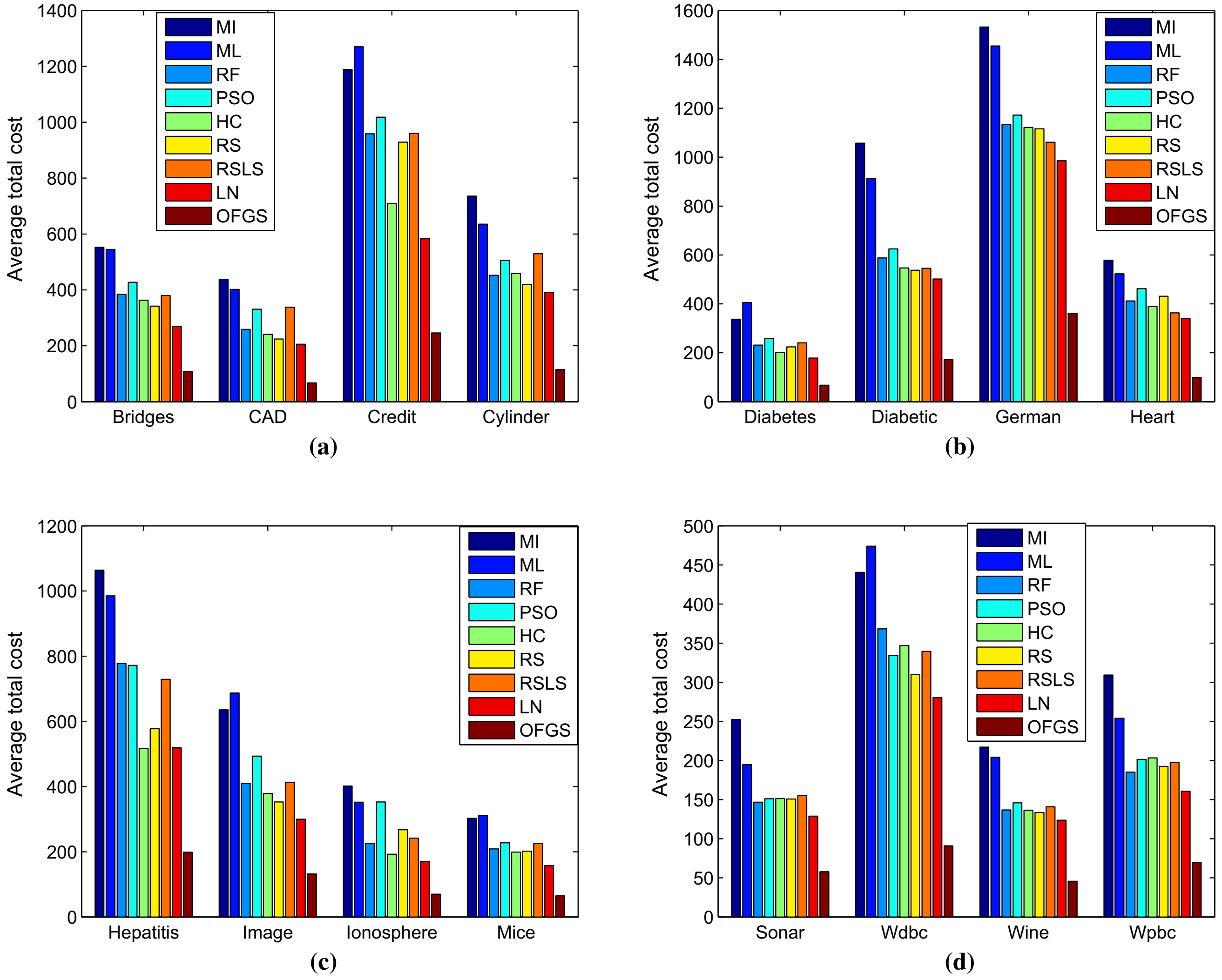
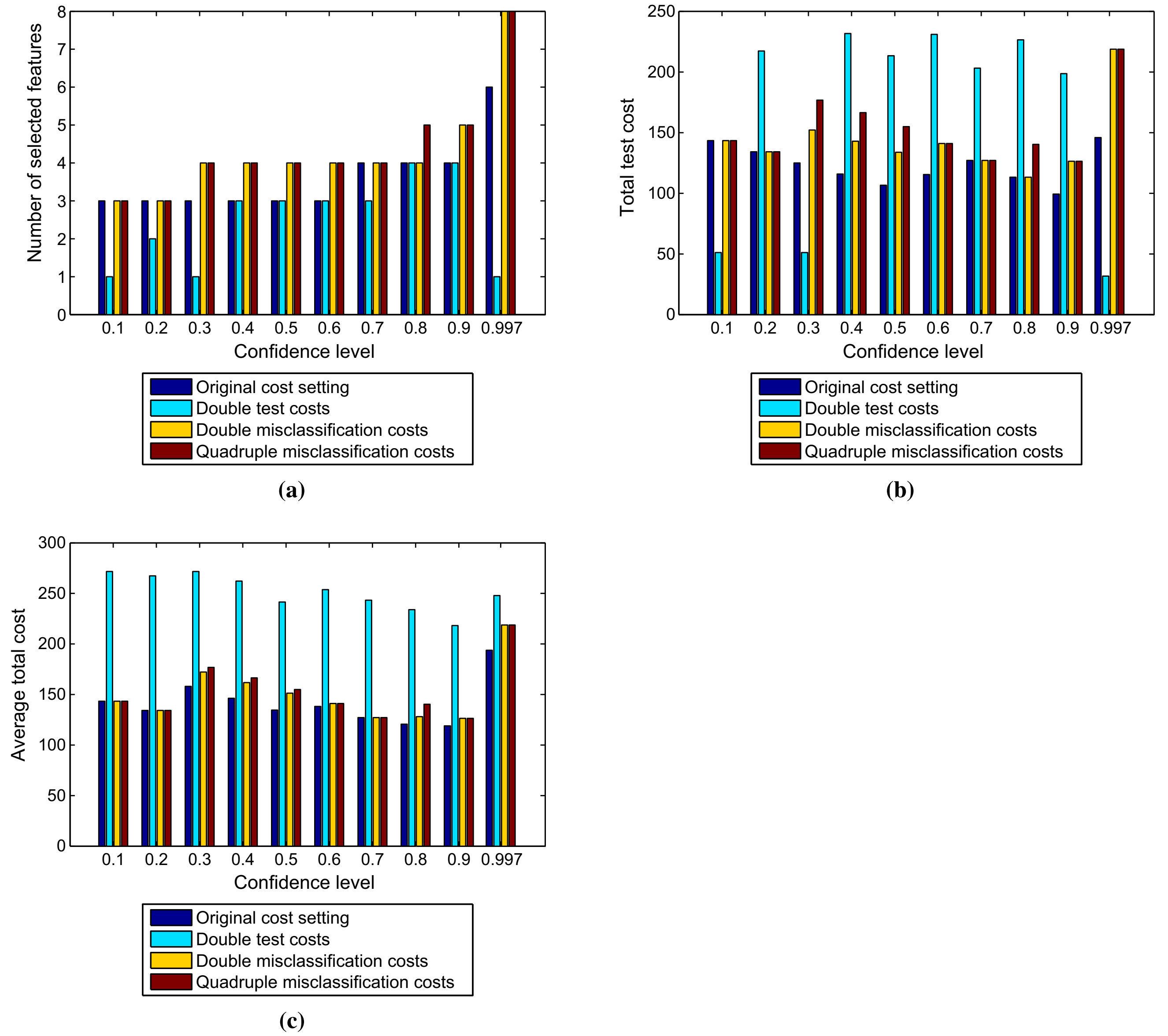


Fig. 8 Representative results of comparisons between the proposed OGFS algorithm and eight existing feature selection algorithms

tracking algorithm (namely Algorithm 2) to find the locally minimal average total cost and the corresponding feature subset at each tried confidence level. Then, the locally minimal average total costs are compared among different confidence levels to choose the globally minimal average total cost and the optimal feature–granularity pair. Hence, the average total cost obtained by the OFGS algorithm is often much lower than those obtained by the eight existing algorithms. For the same reason, it can be deduced that even if the cost settings are set to be identical between the previous algorithms and the OFGS algorithm, namely, the individual test costs are supposed to be changing with the data granularity in the existing algorithms, the algorithms still cannot perform better than the OFGS algorithm on minimizing average total costs because the latter can always achieve the optimal results through invoking the backtracking algorithm. However, it is worth mentioning that the OFGS algorithm has no advantage in the computational efficiency.

4. Influences of different cost settings to the feature-granularity selection

The influences of cost setting changes to the feature-granularity selection are discussed in Sect. 4.3. Here, the influences are further investigated by experimentation. The feature–granularity selection algorithm is run with four different cost settings, which include an original cost setting as a reference, double individual test costs and fixed individual misclassification costs, fixed test costs and double misclassification costs, fixed test costs and quadruple misclassification costs. Five evaluation metrics are used, which are the number of selected features, the total test cost, the misclassification rate, the average misclassification cost and the average total cost. The representative results of Heart dataset are depicted in Fig. 9 and Tables 10, 11.



(a)

(b)

(c)

Fig. 9 Comparisons between different cost settings: **a** number of selected features, **b** total test cost, **c** average total cost

Table 10 Misclassification rates of Heart with respect to different cost settings and different confidence levels, where TCs and MCs are the abbreviations of test costs and misclassification costs, respectively

Confidence levels	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.997
Original costs	0	0	0.0132	0.0132	0.0132	0.0099	0	0.0033	0.0099	0.0165
Double TCs	0.4323	0.0231	0.4323	0.0132	0.0132	0.0099	0.0198	0.0033	0.0099	0.4323
Double MCs	0	0	0.0033	0.0033	0.0033	0	0	0.0033	0	0
Quadruple MCs	0	0	0	0	0	0	0	0	0	0

The following observations could be made from the results:

(1) With the increase in individual test costs, except that the number of selected features usually remains unchanged or gets smaller, other four kinds of quantity often become large; especially, all the average total costs increase. In particular, when the number of selected features drops, the total test cost may decrease, but both the misclassification rate and the average misclassification cost increase significantly. These observations are in line with real-world applications. Take

Table 11 Average misclassification costs of Heart with respect to different cost settings and different confidence levels

Confidence levels	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.997
Original costs	0	0	33.0033	30.363	27.9868	22.7723	0	7.4587	19.604	47.8548
Double TCs	220.495	50.132	220.495	30.495	28.1188	22.7723	40.198	7.4587	19.604	216.1716
Double MCs	0	0	20.066	18.7459	17.5578	0	0	14.9175	0	0
Quadruple MCs	0	0	0	0	0	0	0	0	0	0

the medical diagnosis as an example. If the cost of each kind of test increases while a person only has a certain amount of money, he/she may have to give up some necessary tests, which results in an increasing possibility of misdiagnosis.
(2) With the increase in individual misclassification costs, the number of selected features, the total test cost and the average total cost increase or stay the same, while the misclassification rate and the average misclassification cost usually drop even to zero. Particularly, when the individual misclassification costs are the quadruples of the original ones, namely, they are high enough, all of misclassification rates and average misclassification costs are equal to zero. The reason is that now each misclassification will induce a large cost. To reduce the total misclassification cost, more necessary features are needed. Hence, the misclassification rate is low and even reaches zero. Interestingly, in some practical applications, minimizing the misclassification rate or equivalently minimizing the total misclassification cost is more attractive than minimizing the total cost. It can be found from the above experimental analysis that, if allowed, setting high individual misclassification costs is a feasible solution for this kind of optimization objective. This finding is useful for decision making.

In summary, the effectiveness of the proposed optimal feature–granularity selection algorithm is demonstrated through experiments on multiple data sets from multiple perspectives. A good trade-off among feature dimension reduction, data granularity selection and total cost minimization can be achieved by the algorithm. In addition, the in-depth experimental analysis provides some feasible schemes for decision making. It is notable that although most of the individual test costs for numeric features and the individual misclassification costs are, respectively, given in a single form (either linear functions or piecewise constant functions) in the above experiments for simplicity, one could also set the cost functions in a mixed form or design other types of cost functions according to reality. The observations are similar; namely, the proposed algorithm is an effective solution to obtain the minimal total cost and the optimal feature–granularity pair in any case.

7 Conclusions

In this paper, a feature–granularity selection approach was proposed to find the optimal feature subset and the optimal data granularity simultaneously for hybrid data in terms of measurement errors and variable costs. First, an adaptive neighborhood model was constructed, in which the neighborhood granules are adaptive to the types of features. Then, multiple types of variable cost setting were developed according to reality, and the influences of cost setting changes to the feature–granularity selection were analyzed. Finally, an optimal feature–granularity selection algorithm was designed. Experimental results have validated the effectiveness of the proposed algorithm. By the algorithm, an optimal pair of feature subset and data granularity can be selected to minimize the total cost for processing hybrid data. Particularly, the influences of different cost settings were further investigated in the experiments, which could offer some desirable schemes for decision making. In the future, we will develop a parallel method to solve the feature–granularity selection problem for massive hybrid data.

Acknowledgements This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 61603173 and 11871259, the Natural Science Foundation of Fujian Province, China, under Grant No. 2017J01771, the Institute of Meteorological Big Data-Digital Fujian and Fujian Key Laboratory of Data Science and Statistics.

Compliance with ethical standards

Funding This study was funded by the National Natural Science Foundation of China under Grant Nos. 61603173 and 11871259, the Natural Science Foundation of Fujian Province, China, under Grant No. 2017J01771, the Institute of Meteorological Big Data-Digital Fujian and Fujian Key Laboratory of Data Science and Statistics.

Conflict of interest Author Shujiao Liao declares that she has no conflict of interest. Author Qingxin Zhu declares that he has no conflict of interest. Author Yuhua Qian declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ansorge S, Schmidt J (2015) Visualized episode mining with feature granularity selection. In: Industrial conference on data mining. Springer, Cham, pp 201–215

Bian J, Peng XG, Wang Y, Zhang H (2016) An efficient cost-sensitive feature selection using chaos genetic algorithm for class imbalance problem. Math Probl Eng 2016:1–9

Blake CL, Merz CJ (1998) UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/mlrepository.html

Boussouf M, Quafafou M (2000) Scalable feature selection using rough set theory. In: Proceedings of rough sets and current trends in computing, vol. 2005. LNCS, pp 131–138

Cao P, Zhao DZ, Zaiane O (2013) An optimized cost-sensitive SVM for imbalanced data learning. In: Advances in knowledge discovery and data mining, vol 7819. LNCS, pp 280–292

Chai XY, Deng L, Yang Q, Ling CX (2004) Test-cost sensitive Naïve Bayes classification. In: Proceedings of the 5th international conference on data mining, pp 51–58

Chen DG, Yang YY (2014) Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models. IEEE Trans Fuzzy Syst 22(5):1325–1334

Dai JH, Wang WT, Xu Q, Tian HW (2012) Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. Knowl Based Syst 27:443–450

Dash M, Liu H (2003) Consistency-based search in feature selection. Artif Intell 151:155–176

Domingos P (1999) Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of the 5th international conference on knowledge discovery and data mining, pp 155–164

Doquire G, Verleysen M (2011) An hybrid approach to feature selection for mixed categorical and continuous data. In: Proceedings of the international conference on knowledge discovery and information retrieval, pp 394–401

Du J, Cai ZH, Ling CX (2007) Cost-sensitive decision trees with prepruning. In: Proceedings of Canadian AI, No. 4509. LNAI, pp 171–179

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc Lond Ser A Contain Pap Math Phys Charact 222:309–368

Greiner R, Grove AJ, Roth D (2002) Learning cost-sensitive active classifiers. Artif Intell 139(2):137–174

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hu QH, Yu DR, Liu JF, Wu CX (2008) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178(18):3577–3594

Hu QH, Pedrycz W, Yu DR, Lang J (2010) Selecting discrete and continuous features based on neighborhood decision error minimization. IEEE Trans Syst Man Cybern Part B Cybern 40(1):137–150

Huang TY, Zhu W (2017) Cost-sensitive feature selection via manifold learning. J Shandong Univ 52(3):91–96

Iswandy K, Koenig A (2006) Feature selection with acquisition cost for optimizing sensor system design. Adv Radio Sci 4:135–141

Jia XY, Liao WH, Tang ZM, Shang L (2013) Minimum cost attribute reduction in decision-theoretic rough set models. Inf Sci 219:151–167

Kannan SS, Ramaraj N (2010) A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm. Knowl Based Syst 23:580–585

Liang JY, Wang F, Dang CY, Qian YH (2012) An efficient rough feature selection algorithm with a multi-granulation view. Int J Approx Reason 53:912–926

Liao SJ, Zhu QX, Min F (2014) Cost-sensitive attribute reduction in decision-theoretic rough set models. Math Probl Eng 2014:1–9

Liao SJ, Zhu QX, Liang R (2017) An efficient approach of test-cost-sensitive attribute reduction for numerical data. Int J Innov Comput Inf Control 13(6):2099–2111

Liao SJ, Zhu QX, Qian YH, Lin GP (2018) Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs. Knowl Based Syst 158:25–42

Liu GL, Sai Y (2009) A comparison of two types of rough sets induced by coverings. Int J Approx Reason 50(3):521–528

Luo C, Li TR, Chen HM, Lu LX (2015) Fast algorithms for computing rough approximations in set-valued decision systems while updating criteria values. Inf Sci 299:221–242

Min F, He HP, Qian YH, Zhu W (2011) Test-cost-sensitive attribute reduction. Inf Sci 181:4928–4942

Min F, Hu QH, Zhu W (2014) Feature selection with test cost constraint. Int J Approx Reason 55:167–179

Pendharkar PC (2013) A maximum-margin genetic algorithm for misclassification cost minimizing feature selection problem. Expert Syst Appl 40(10):3918–3925

Shu WH, Shen H (2016) Multi-criteria feature selection on cost-sensitive data with missing values. Pattern Recogn 51:268–280

Turney PD (2000) Types of cost in inductive concept learning. In: Proceedings of the workshop on cost-sensitive learning at the 17th ICML, pp 1–7

Wang T, Qin ZX, Jin Z, Zhang S (2010) Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning. J Syst Softw 83(7):1137–1147

Weiss Y, Elovici Y, Rokach L (2013) The cash algorithm-cost-sensitive attribute selection using histograms. Inf Sci 222:247–268

Yao YY (2004) A partition model of granular computing. Lect Notes Comput Sci 3100:232–253

Yao YY, Zhao Y (2008) Attribute reduction in decision-theoretic rough set models. Inf Sci 178(17):3356–3373

Yu SL, Zhao H (2018) Rough sets and Laplacian score based cost sensitive feature selection. PLoS ONE 13(6):1–23

Zhang SC, Liu L, Zhu XF, Zhang C (2008) A strategy for attributes selection in cost-sensitive decision trees induction. In: IEEE 8th international conference on computer and information technology workshops, Sydney, QLD, pp 8–13

Zhang Y, Gong DW, Cheng J (2017) Multi-objective particle swarm optimization approach for cost-based feature selection in classification. IEEE/ACM Trans Comput Biol Bioinform 14(1):64–75

Zhao H, Yu SL (2019) Cost-sensitive feature selection via the $l_{2,1}$-norm. Int J Approx Reason 104:25–37

Zhao H, Zhu W (2014) Optimal cost-sensitive granularization based on rough sets for variable costs. Knowl Based Syst 65:72–82

Zhao H, Min F, Zhu W (2013) Cost-sensitive feature selection of numeric data with measurement errors. J Appl Math 2013:1–13

Zhou YH, Zhou ZH (2016) Large margin distribution learning with cost interval and unlabeled data. IEEE Trans Knowl Data Eng 28(7):1749–1763

Zhou QF, Zhou H, Li T (2016) Cost-sensitive feature selection using random forest: selecting low-cost subsets of informative features. Knowl Based Syst 95:1–11

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.