Soft Computing (2022) 26:3663–3677
https://doi.org/10.1007/s00500-022-06838-x

FOUNDATIONS

A group incremental approach for feature selection on hybrid data

Feng Wang[1] · Wei Wei[1] · Jiye Liang[1]

Accepted: 21 January 2022 / Published online: 26 February 2022

Abstract

Feature selection for dynamic data sets has been perceived as a very significant hot research problem in data mining. In practice, most real-world data usually are hybrid, which means both include categorical data and numerical data. For dynamic hybrid data, this paper first introduces a new neighborhood relation and information entropy based on neighborhood accordingly. Secondly, the single incremental mechanism and group incremental mechanism are analyzed and proofed to construct feature significance. On this basis, two incremental approaches to feature selection are developed for dealing with hybrid data. To better demonstrate the new algorithm, four common classifiers and twelve UCI data sets are introduced in the experiments. The experimental results further validate the feasibility of the incremental algorithms, and especially the efficiency of the group incremental algorithm.

Keywords Incremental feature selection · Dynamic data sets · information entropy · Neighborhood relation

1 Introduction

With the rapid development of the times, the word of big data has come into all walks of life thoroughly. As it turns out, data scientists usually describe " big data" as having five distinct dimensions, i.e., the 5V's of big data. And one of most important characteristic of big data is velocity which refers to the speed at which vast amounts of data are being generated, acquired and analyzed. Obviously, this speed becomes higher and higher nowadays. In data mining, incremental technique is a kind of common and effective approaches used to discover knowledge from dynamic data. However, the emergence of big data brings unprecedented challenges for traditional incremental approaches, thus exploring more efficient incremental approaches has urgently become a key issue in many areas. Feature selection is a kind of well-known techniques for data preprocessing and has been widely introduced in many areas such as pattern recognition, machine learning and data mining (Liu et al. 2002, 2005, 2007, 2009; Liang et al. 2012, 2014). It aims to improve the performance of models and save training time through removing redundant or irrelevant features from data sets. Therefore, this paper focuses on efficient incremental feature selection strategies.

Most of the traditional feature selection algorithms are suitable for single type of data, i.e. categorical or numerical data (Almuallim and Dietterich 1994; Benabdeslem and Hindawi 2004; Dash and Liu 1997, 2003; Guyon and Elisseeff 2003; Pedrycz and Vukovich 2002). However, in the age of big data, the types of data become more complex. Lots of real data no longer be single type, but take on hybrid forms. In other words, categorical and numerical features coexist in many application fields such as medical data, financial data and biological data. For feature selection on hybrid data, many researchers have proposed lots of considerable algorithms Hall et al. and Liu defined the correlation for categorical features and numerical features and proposed correlation-based feature selection algorithms for hybrid data (Hall 2000; Yu and Liu 2004). In Tang and Mao (2007), hybrid data sets are divided into a series of feature subspaces according to categorical features; and the class separability is measured in each subspace. The performance of this algorithm mainly depend on the classification ability of categorical features (Tang and Mao 2007). On the basis of Parzen window, Kwak and Choi proposed a new feature selection

✉ Jiye Liang
ljy@sxu.edu.cn

Feng Wang
sxuwangfeng@126.com

Wei Wei
weiwei@sxu.edu.cn

[1] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China

algorithm based on mutual information (Kwak and Choi 2002). In addition, by introducing fuzzy rough set theory, Hu et al. proposed a series of feature selection algorithms based on fuzzy equivalence relation, which can be used to find target feature selection from hybrid data (Hu et al. 2007, 2006, 2008). Wei et al. developed a new information entropy and put forward an efficient accelerated feature selection algorithm for hybrid data (Wei et al. 2009). By plugging a dedicated kernel, Paul et al. proposed two heterogeneous feature selection algorithms which can handle both continuous and nominal features (Paul et al. 2015). For hybrid data with missing values, Zhao et al. introduces a neighborhood-tolerance relation and neighborhood-tolerance entropy to deal with selecting useful features in incomplete hybrid data(Zhao and Qin 2014). To further reduce improve efficiency, Wang et al. studied an efficient rough feature selection algorithm from a multi-granulation view which can find a target feature subset in much shorter time(Wang and Liang 2016). However, these algorithm are only suitable for static data. With the coming of dynamic data, feature selection on dynamic hybrid data has become a very significant issue and arouses attention increasingly.

As is well known, both objects and features in dynamic data may vary with time. Even feature values may also change in some specific application data. In order to more effectively deal with dynamic data, there have existed many researches on dynamic data sets. Incremental learning, as a common used technique, has been applied in all aspects of data mining and machine learning. Whether in supervised learning algorithm, unsupervised learning algorithm, or semi-supervised learning algorithm, study on incremental learning models all attracted much attention (Gama 2012; Huang et al. 2020; Shu et al. 2019; Yang et al. 2021). Various of classic data processing techniques or learning models such as decision tree, SVM and manifold learning have been extended to incremental solution mechanism. In addition, online learning and data stream mining are also explorations of dynamic data. As research continues, in view of variation of data distribution, concept drift is a kind of strategy which can be used to handle dynamic data with varying data distribution (Lu et al. 2019; Mera et al. 2019; Neto and Canuto 2021). In above-mentioned numerous studies, incremental feature selection also caused extensive concern. Existing researches on incremental feature selection refer to two main kinds of methods. One devotes to construct incremental solution framework for feature selection. And the other one focuses on analyzing incremental mechanism of measures used in feature selection. For dynamic data with increasing scale, Liang et al. proposed a group incremental feature selection algorithm based on rough set theory (Liang et al. 2014). This algorithm devotes to find an effective feature subset in in a much shorter time when objects are added in groups. On this basis, Jing et al. presents an incremental feature selection algorithm from a view of multi-granulation (Jing et al. 2017). For dynamic data with varying features, dimension incremental mechanisms for feature selection provide strategies of selecting useful features from dynamic data sets with varying features (Wang et al. 2013). Researches on feature stream also attract much attention and enrich exploration of models for feature variation (Liu et al. 2018; Zhou et al. 2019). Finally, for dynamic data with varying feature values, Wang et al. presents a dynamic feature selection algorithm for data sets with varying feature values (Wang et al. 2013). In summary, most of the existing incremental feature selection algorithms are only suitable for single type of data. This paper focus on incremental feature selection algorithm for hybrid data.

The concept of neighborhood has been introduced to deal with imbalance data, small labeled data, numerical data et al. in many data mining methods(Chen et al. 2019; Chang et al. 2019; Wang 2006). Hu et al. introduced neighborhood rough set theory to construct fuzzy equivalence relation and feature selection algorithms for hybrid data (Hu et al. 2007, 2006, 2008). Hence, making reference to the thought of neighborhood, we study incremental feature selection algorithms for hybrid data in this paper. Firstly, this paper defines a new similarity measure for categorical data and on this basis, a new neighborhood relation is introduced to develop the neighborhood rough set model. Then, a new information entropy is proposed to measure the uncertainty of hybrid data sets. To further reduce computational time, the single incremental mechanism and group incremental mechanism are analyzed respectively, which can be used to compute entropy value of dynamic hybrid data more efficiently. Finally, based on the incremental mechanisms, two incremental feature selection algorithms are constructed to select useful features. To better demonstrate the new algorithms, four common classifiers and twelve UCI data sets are introduced in the experiments. The experimental results further validate the feasibility of the incremental algorithms, and especially the efficiency of the group incremental algorithm.

The rest of this paper is organized as follows: some preliminaries are briefly reviewed in Sect. 2. In Sect. 3, a new similarity measure for categorical data is defined to construct the similarity of hybrid data. In Sect. 4, based on the new similarity measure, we develop the neighborhood rough set model and introduce a new information entropy to measure the uncertainty of hybrid data. Section 5 analyzes the incremental mechanisms of the new information entropy and feature significance. Section 6 introduces the incremental feature selection algorithms based on incremental mechanisms of feature significance. In Sect. 7, twelve UCI data sets are employed to illustrate the effectiveness of the proposed algorithm. Section 8 concludes the paper with some discussions.

2 Preliminary knowledge

2.1 Basic concepts

Rough set theory, proposed by Pawlak in 1982, has been conceived as a representative granular computing model (Pawlak 1998; Pawlak and Skowron 2007). It has become a popular mathematical framework and has been successfully applied in many areas such as pattern recognition, image processing and feature selection (Swiniarski and Skowron 2003; Shao and Zhang 2005; Wang et al. 2013; Wu et al. 2003; Xu et al. 2009; Yao 2006; Yao and Zhao 2008).

Classic rough set theory is only suitable for categorical data. In rough set theory, U is called the universe and represents a finite and nonempty set of objects; and A denotes the features set of objects in U. Let quadruple $S = (U, A, V, f)$ be a data table, where $V = \bigcup_{a \in A} V_a$ with V_a being the domain of feature a, and $f : U \times A \rightarrow V$ is an information function with $f(x, a) \in V_a$ for $a \in A$ and $x \in U$. For convenience, S is often simplified as $S = (U, A)$.

A feature subset $B \subseteq A$ induces an equivalence relation R_B on U, where

$$R_B = \{(x, y) \in U \times U | f(x, a) = f(y, a), \forall a \in B\}.$$

$U/R_B = \{[x]_B | x \in U\}$ can generate a family of equivalence classes denoted by

$$[x]_B = \{y \in U | (x, y) \in R_B\},$$

where U/R_B is simplified as U/B.

For $X \subseteq U$ and an equivalence relation R on U, the lower approximation and upper approximation operators of X are defined by

$$\underline{R}X = \bigcup \{x \in U | [x]_R \subseteq X\},$$

and

$$\overline{R}X = \bigcup \{x \in U | [x]_R \cap X \neq \emptyset\},$$

respectively. The positive region of X is defined by $POS_R(X) = \underline{R}X$.

In rough set theory, a labeled data set is often called a decision table. Let $S = (U, C \cup D)$ be a decision table, where D is called decision attribute which means class information, and C is called conditional attribute set which means all features in U. Decision attribute D can also induce an equivalence partition on U, which is denoted by $U/D = \{D_1, D_2, \ldots, D_r\}$. For $P \subseteq C$, the positive region of D with respect to P is defined by $POS_P(D) = \bigcup_{i=1}^{r} \underline{P} D_i$.

2.2 Complementary information entropy

It is well known that the expression of information usually includes uncertainty, vagueness, approximate expression, and so on. In rough set theory, entropy or its variants is introduced to measure uncertainty of data sets. Liang introduced the definition of complementary entropy which not only can measure the uncertainty, but also the fuzziness of a rough set (Liang et al. 2002).

Definition 1 Let $S = (U, A)$ be a unlabeled data table and $B \subseteq A$. Then, one can obtain a partition $U/B = \{X_1, X_2, \ldots, X_m\}$. On this basis, a complementary entropy of B is defined as

$$E(B) = \sum_{i=1}^{m} \frac{|X_i|}{|U|}(1 - \frac{|X_i|}{|U|}). \quad (1)$$

Definition 2 Let $S = (U, C \cup D)$ be a labeled data table and $B \subseteq C$. Then, one can obtain the partitions $U/B = \{X_1, X_2, \ldots, X_m\}$ and $U/D = \{Y_1, Y_2, \ldots, Y_n\}$. On this basis, a complementary conditional entropy of B relative to D is defined as

$$E(D|B) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|X_i \cap Y_j|}{|U|} \frac{|X_i - Y_j|}{|U|}. \quad (2)$$

3 A new distance measure for hybrid data ($D_r d$)

As mentioned above, classic rough set theory is only suitable for categorical data and cannot directly deal with numerical data. To handle hybrid data, researchers develop the neighborhood rough set theory which has been recognized as an effective technique for tackling complex data. In neighborhood rough set theory, one of the key problems is the definition of neighborhood. This section introduces a new distance measure for hybrid data.

3.1 Distance measure for categorical data

It is well known that numerical data have natural geometric property, and one can measure its distance or similarity through computing the difference between their feature values. However, unlike numerical data, categorical data lack such geometric property and need further exploration for its similarity (or distance). For categorical data, the earliest and widely used work on its similarity is Hamming distance. This distance employs 0 and 1 to measure the similarity of different and identical categorical values. Moreover, the frequency of feature value and the co-occurrences between different values have also been applied to learn categorical similarity. Objectively, the difference between categorical feature values not only relies on themselves, but also relies on the environment. In other words, after capturing the difference within same feature, we should also focus on the contribution of other features to similarity of categorical feature values. On this basis, we propose a new distance measure for categorical feature values from the view of data distribution in this section. This distance measure not only considers difference of feature values within a same feature but also the discrimination induced by other features.

As previously described, entropy is always used to measure uncertainty or disorder of a data table. In machine learning, entropy is also considered as a measure for data distribution. For two feature values in a feature, their distribution is obviously related to their distance. The more similar their distribution are, the bigger their similarities. Accordingly, this section introduces the complementary entropy to construct the distance measure for categorical data. Firstly, the inner categorical values distance considered the relationship between data distribution is defined as follows.

Definition 3 Let $S = (U, C \cup D)$ be a data table and $U/D = \{Y_1, Y_2, \ldots, Y_n\}$. $\forall x, y \in U$ and $a_i \in C$, the inner distance between objects x and y on feature a_i is defined as

$$D_{in}(i, x, y) = \frac{1}{n}\sum_{j=1}^{n} |[x]_{a_i} \cap Y_j| \cdot |Y_j - [x]_{a_i}| - |[y]_{a_i} \cap Y_j| \cdot |Y_j - [y]_{a_i}|. \quad (3)$$

Proposition 1 *If* $\forall Y_j \in U/D$, $|[x]_{a_i} \cap Y_j| = |[y]_{a_i} \cap Y_j|$, *then* $S_{in}(i, x, y)$ *reaches its maximum* 1. *If* $\forall 1 \leq j \leq n$, $min(S_1, S_2) = 0$, *then* $S_{in}(i, x, y)$ *reaches its minimum* 0.

The inner distance of categorical values shows the difference between the feature values of feature a_i. In what follows, we study the outer feature similarity, which involves the contribution of other features $a_k \in C(i \neq k)$ to the distance of feature values in a_i. In rough set theory, the upper approximation operator means those equivalence classes which include the target concept. Here introduces the upper approximation of feature values to construct the outer feature similarity in this section. For feature values $[x]_{a_i}$ and $[y]_{a_i}$, the more similar their upper approximations relative to feature $a_k (k \neq i)$ are, the higher their outer similarity are. Similarly, based on the concept of entropy, we also use the relationship between data distribution of upper approximations as outer distance of feature values, which is defined as follows.

Definition 4 Let $S = (U, C \cup D)$ be a data table and $U/D = \{Y_1, Y_2, \ldots, Y_n\}$. $\forall x, y \in U$ and $a_i \in C$, the outer distance between objects x and y on another feature $a_k \in C(i \neq k)$is defined as

$$D_{out}(i, x, y) = \frac{1}{n}\sum_{j=1}^{n} |\overline{a_k}([x]_{a_i}) \cap Y_j| \cdot |\overline{a_k}([x]_{a_i}) - Y_j| - |\overline{a_k}([y]_{a_i}) \cap Y_j| \cdot |\overline{a_k}([y]_{a_i}) - Y_j|. \quad (4)$$

Based on above definitions of inner and outer distances for feature values, we introduce the distance between objects on feature a_i by combing its inner and outer distance. Moreover, calculating the distance between two objects on each feature $a_i \in B$ can obtain their distance on feature subset B.

Definition 5 Let $S = (U, C \cup D)$ be a data table and $U/D = \{Y_1, Y_2, \ldots, Y_n\}$. The distance between $x, y \in U$ on categorical feature a_i is defined as

$$D(i, x, y) = \alpha D_{in}(i, x, y) + \beta \frac{\sum_{k=1, k\neq i}^{|C|} D_{out}(i, x, y)}{|C| - 1}, \quad (5)$$

where $\alpha + \beta = 1$.

Definition 6 Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$ be categorical features. The distance between $x, y \in U$ on feature subset B is defined as

$$D'_B(x, y) = \frac{1}{|B|}\sum_{a_i \in B} D(i, x, y). \quad (6)$$

3.2 Similarity measure for hybrid data

For numerical data, here introduces a common distance measure, i.e., Euclidean distance, to measure the distance between two objects on numerical features.

Definition 7 Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$ be numerical features. The distance between $x, y \in U$ on feature subset B is defined as

$$D''_B(x, y) = \left(\sum_{i=1}^{|B|} (f'(x, a_i) - f'(y, a_i))^2\right)^{\frac{1}{2}}, \quad (7)$$

where $f'(x, a_i)$ and $f'(y, a_i)$ are the normalized feature values of $f(x, a_i)$ and $f(y, a_i)$, respectively.

According to Definitions 6 and 7, following definition shows the distance between objects on hybrid features.

Definition 8 Let $S = (U, C \cup D)$ be a data table, $B_1 \subseteq C$ be categorical feature, and $B_2 \subseteq C$ be numerical features. The distance between $x, y \in U$ on feature subset $B = B_1 \cup B_2$ is defined as

$$D_r d_B(x, y) = \frac{|B_1| D'_{B_1}(x, y) + |B_2| D''_{B_2}(x, y)}{|B|}. \quad (8)$$

4 Neighborhood rough set model based on the new distance measure $D_r d$

4.1 Basic concepts

As mentioned above, classic rough set theory is only suitable for categorical data and its approximation operators are induced by equivalence relation. To deal with complex data, the idea of neighborhood is introduced to develop approximation operators based on neighborhood relation and neighborhood rough set theory. Accordingly, based on the hybrid feature similarity defined in Sect. 3, we introduce some basic concepts in neighborhood rough set theory in this section.

Definition 9 Let $S = (U, A)$ be a data table, $B \subseteq A$ and $x \in U$. The neighborhood $\delta_B(x)$ of x on feature subset B is defined as

$$\delta_B(x) = \{y | D_r d_B(x, y) \leq \delta, y \in U\}, \quad (9)$$

where D_B means the distance induced on B, and δ is the neighborhood radius.

Definition 10 Let $S = (U, A)$ be a data table and $B \subseteq A$. Feature subset B can induce an neighborhood relation R_B on U, which is defined as

$$R_{B,\delta} = \{(x, y) \in U \times U | y \in \delta_B(x)\}. \quad (10)$$

Obviously, $U/R_{B,\delta} = \{\delta_B(x) | x \in U\}$ is a family of neighborhood granules on the universe. According to classic rough set theory, the lower and upper approximation operators based on δ neighborhood are defined as follows.

Definition 11 Let $S = (U, A)$ be a data table, $B \subseteq A$ and $X \subseteq U$. The lower and upper approximation operators of X are defined as

$$\underline{R_{B,\delta}} X = \{x | \delta_B(x) \subseteq X, x \in U\} \quad (11)$$

and

$$\overline{R_{B,\delta}} X = \{x | \delta_B(x) \cap X \neq \emptyset, x \in U\}, \quad (12)$$

respectively. The positive region of X is defined by $POS_R(X) = \underline{R} X$.

4.2 Rough feature selection based on information entropy

On the basis of the new neighborhood relation, this subsection defines a new information entropy for uncertainty in labeled data.

Definition 12 Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$. On the basis of neighborhood rough set model, the information entropy on B is defined as

$$E(D|B) = \sum_{i=1}^{|U|} \frac{|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|}{|U|^2}. \quad (13)$$

According to the new information entropy, following definition introduces two types of feature significance to select useful feature while keeping the same uncertainty. The inner feature significance is usually used to detect redundant features, and the out feature significance can be used to add new features to existing feature subset.

Definition 13 Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$. On the basis of complementary entropy, the inner significance of feature $a \in B$ is defined as

$$Sin_{in}(a, B) = E(D|B - \{a\}) - E(D|B). \quad (14)$$

The outer significance of feature $a \in C - B$ is defined as

$$Sin_{out}(a, B) = E(D|B) - E(D|B \cup \{a\}). \quad (15)$$

Following definition shows the stop criteria and select condition in the process of rough feature selection. The first condition guarantees that the target feature subset has the same distinguish ability as the whole feature set, and the second condition guarantees that there is no redundant features in the target feature subset.

Definition 14 Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$. Then, the attribute set B is a target feature subset if B satisfies:

1) $E(D|B) = E(D|C)$;
2) $\forall a \in B$, $E(D|B) < E(D|B - \{a\})$.

5 Dynamic solution mechanism for feature significance

According to Definition 13, in the process of computing feature significance, a key step is the solution of entropy value. Hence, in order to find incremental mechanisms for feature significance, this section first introduces the incremental mechanisms for complementary entropy. Subsection 5.1 presents detailed analysis and proof of single incremental mechanism and group incremental mechanism for complementary entropy separately.

5.1 Incremental mechanism for complementary entropy

In the process of feature selection, according to Definition 13, the calculation of entropy is obviously a key step. However, with the increase of objects, recomputing entropy is very time-consuming for a dynamic data table. To address this issue, this section introduces incremental mechanisms for computing entropies. When new objects are added to a given table, instead of re-computation on new table, the incremental mechanisms aim to calculate new entropy values by combing the changes of granules into the existing entropies of the original table. In this section, Theorem 1 shows the single incremental mechanism for computing entropy of a hybrid data table, which means the new object being added one by one.

Theorem 1 *Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$. When a new object x is added to table S, the new complementary conditional entropy on $U \cup \{x\}$ becomes*

$$E_{U\cup\{x\}}(D|B) = \frac{1}{(|U|+1)^2} \times \left(|U|^2 E_U(D|B) + 2|\delta'_B(x) - [x]'_D|\right), \tag{16}$$

where, $\delta'_B(x)$ means the neighborhood of x on $U \cup \{x\}$, and $[x]'_D$ is the equivalence class of x on on $U \cup \{x\}$.

Proof When x is added to U, supposed $U \cup \{x\}/R_{B,\delta} = \{\delta'_B(x_i)|x_i \in U\cup\{x\}\}$ and $U\cup\{x\}/D = \{[x_i]'_D|x_i \in U\cup\{x\}\}$, there are four situations about neighborhood relation between x and objects belonged to U:

1) $\forall y \in U - \delta'_B(x) \cup [x]'_D$, the relation between x and y is not satisfied neighborhood relation;
2) $\forall y \in \delta'_B(x) - [x]'_D$, x and y satisfy neighborhood relation induced by feature subset B;
3) $\forall y \in [x]'_D - \delta'_B(x)$, x and y satisfy equivalence relation on D;
4) $\forall y \in [x]'_D \cap \delta'_B(x)$, x and y satisfy equivalence relation on D and neighborhood relation on B.

On this basis, let $X = |\delta'_B(x)|$, $Y = |[x]'_D|$ and $t = |U - X \cup Y|$, the complementary entropy induced by B on $U \cup \{x\}$ is computed as

$$\begin{aligned}
&E_{U\cup\{x\}}(D|B) \\
&= \sum_{i=1}^{|U|+1} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{(|U|+1)^2} \\
&= \frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{t}(|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|) \\
&\quad + \sum_{i=t}^{|U-Y|}(|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|) \\
&\quad + \sum_{i=t+|X|}^{|U|}(|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|) \\
&\quad + \sum_{i=|U-Y|}^{t+|X|}(|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|) \\
&\quad + |\delta'_B(x)| - |\delta'_B(x) \cap [x]'_D|\Big) \\
&= \frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{t}(|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|) \\
&\quad + \sum_{i=t}^{|U-Y|}(|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]_D|) \\
&\quad + \sum_{i=t+|X|}^{|U|}(|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]'_D|) \\
&\quad + \sum_{i=|U-Y|}^{t+|X|}(|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|) \\
&\quad + |\delta'_B(x)| - |\delta'_B(x) \cap [x]'_D|\Big) \\
&= \frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{t}(|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|) \\
&\quad + \sum_{i=t}^{|U-Y|}(|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D| + 1) \\
&\quad + \sum_{i=t+|X|}^{|U|}(|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|) \\
&\quad + \sum_{i=|U-Y|}^{t+|X|}(|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|) + |X - Y|\Big) \\
&= \frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{|U|}(|\delta_B(x_i)| \\
&\quad - |\delta_B(x_i) \cap [x_i]_D|) + 2|X - Y|\Big) \\
&= \frac{1}{(|U|+1)^2}(|U|^2 E_U(D|B) + 2|X - Y|) \\
&= \frac{1}{(|U|+1)^2}(|U|^2 E_U(D|B) + 2|\delta'_B(x) - [x]'_D|).
\end{aligned}$$

This completes the proof. □

For a dynamic data table, when multiple objects are added, the incremental mechanism introduced in Theorem 1 obviously need to repeat the operation many times. To overcome this limitation, Theorem 2 further introduces a group incremental mechanism of computing entropy of a dynamic hybrid data table.

Theorem 2 *Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$. For a group of new objects U_n, $\forall y \in U_n$, its neighborhood induced by B indicates as $\delta_B(y)$, and its equivalence class on D indicates as $[y]_D$. When U_n is added to U, let $X_1 \subseteq U$ and $X_2 \subseteq U_n$ be objects whose neighborhood changes on $U \cup U_n$ and $Y_1 \subseteq U$ and $Y_2 \subseteq U_n$ be objects whose equivalence class induced by D changes on $U \cup U_n$, the new complementary conditional entropy on $U \cup U_n$ becomes*

$$E_{U\cup U_n}(D|B) = \frac{1}{(|U|+|U_n|)^2} \times \Big(|U|^2 E_U(D|B) + |U_n|^2 E_{U_n}(D|B) + \Delta_1 + \Delta_2\Big), \quad (17)$$

where

$$\Delta_1 = \sum_{i=t_1}^{t_1+|X_1|} |\delta_B^{U_n}(x_i)| + \sum_{i=t_2}^{t_2+|X_2|} |\delta_B^{U}(y_i)|,$$

$$\Delta_2 = \sum_{i=|U-Y_1|}^{t_1+|X_1|} |\delta_B^{U_n}(x_i) \cap [x_i]_D^{U_n}| + \sum_{i=|U_n-Y_2|}^{t_2+|X_2|} |\delta_B^{U}(y_i) \cap [y_i]_D^{U}|,$$

$t_1 = |U - X_1 \cup Y_1|$ *and* $t_2 = |U - X_2 \cup Y_2|$.

Proof When new objects set U_n is added to U, the neighborhood and equivalence class of a part of objects may change. As mentioned above, let $X_1 \subseteq U$ and $X_2 \subseteq U_n$ be objects whose neighborhood changes on $U \cup U_n$ and $Y_1 \subseteq U$ and $Y_2 \subseteq U_n$ be objects whose equivalence class induced by D changes on $U \cup U_n$, the new complementary conditional entropy on $U \cup U_n$ is computed as

$$E_{U\cup U_n}(D|B) = \sum_{i=1}^{|U|} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2} + \sum_{i=1}^{|U_n|} \frac{|\delta'_B(y_i)| - |\delta'_B(y_i) \cap [y_i]'_D|}{|U \cup U_n|^2}.$$

For the first item in above formula, one has

$$\sum_{i=1}^{|U|} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2} = \sum_{i=1}^{t_1} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2} + \sum_{i=t_1}^{|U-Y_1|} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2} + \sum_{i=t_1+|X_1|}^{|U|} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2} + \sum_{|U-Y_1|}^{t_1+|X_1|} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2}$$

$$= \sum_{i=1}^{t_1} \frac{|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|}{|U \cup U_n|^2} + \sum_{i=t_1}^{|U-Y_1|} \frac{|\delta'_B(x_i)| - |\delta'_{B,\sigma}(x_i) \cap [x_i]_D|}{|U \cup U_n|^2} + \sum_{i=t_1+|X_1|}^{|U|} \frac{|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2} + \sum_{|U-Y_1|}^{t_1+|X_1|} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2}.$$

Let $\delta'_B(x_i) = \delta_B(x_i) \cup \delta_B^{U_n}(x_i)$ and $[x_i]'_D = [x_i]_D \cup [x_i]_D^{U_n}$, where $\delta_B^{U_n}(x_i)$ means neighborhood of $x_i \in U$ on U_n and $[x_i]_D^{U_n}$ indicates equivalence class induced by D on U_n. Then, one has

$$\sum_{i=1}^{|U|} \frac{|\delta'_B(x_i)| - |\delta'_B(x_i) \cap [x_i]'_D|}{|U \cup U_n|^2} = \sum_{i=1}^{t_1} \frac{|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|}{|U \cup U_n|^2} + \sum_{i=t_1}^{|U-Y_1|} \frac{|\delta_B(x_i) \cup \delta_B^{U_n}(x_i)| - |(\delta_B(x_i) \cup \delta_B^{U_n}(x_i)) \cap [x_i]_D|}{|U \cup U_n|^2} + \sum_{i=t_1+|X_1|}^{|U|} \frac{|\delta_B(x_i)| - |\delta_B(x_i) \cap ([x_i]_D \cup [x_i]_D^{U_n}|)}{|U \cup U_n|^2} + \sum_{|U-Y_1|}^{t_1+|X_1|} \frac{|\delta_B(x_i) \cup \delta_B^{U_n}(x_i)| - |(\delta_B(x_i) \cup \delta_B^{U_n}(x_i)) \cap ([x_i]_D \cup [x_i]_D^{U_n})}{|U \cup U_n|^2}$$

$$= \sum_{i=1}^{|U|} \frac{|\delta_B(x_i)| - |\delta_B(x_i) \cap [x_i]_D|}{|U \cup U_n|^2} + \sum_{i=t_1}^{|U-Y_1|} \frac{|\delta_B^{U_n}(x_i)|}{|U \cup U_n|^2} + \sum_{|U-Y_1|}^{t_1+|X_1|} \frac{|\delta_B^{U_n}(x_i)| - |\delta_B^{U_n}(x_i) \cap [x_i]_D^{U_n}|}{|U \cup U_n|^2} = \frac{|U|^2}{|U \cup U_n|^2} E_U(D|B) + \sum_{i=t_1}^{t_1+|X_1|} \frac{|\delta_B^{U_n}(x_i)|}{|U \cup U_n|^2} - \sum_{|U-Y_1|}^{t_1+|X_1|} \frac{|\delta_B^{U_n}(x_i) \cap [x_i]_D^{U_n}|}{|U \cup U_n|^2}.$$

Similarly, one can also get

$$\sum_{i=1}^{|U_n|} \frac{|\delta'_B(y_i)| - |\delta'_B(y_i) \cap [y_i]'_D|}{|U \cup U_n|^2} = \frac{|U_n|^2}{|U \cup U_n|^2} E_{U_n}(D|B) + \sum_{i=t_2}^{t_2+|X_2|} \frac{|\delta_B^{U}(y_i)|}{|U \cup U_n|^2} - \sum_{|U_n-Y_2|}^{t_2+|X_2|} \frac{|\delta_B^{U}(y_i) \cap [y_i]_D^{U}|}{|U \cup U_n|^2}.$$

Let $\Delta_1 = \sum_{i=t_1}^{t_1+|X_1|} |\delta_B^{U_n}(x_i)| + \sum_{i=t_2}^{t_2+|X_2|} |\delta_B^{U}(y_i)|$ and $\Delta_2 = \sum_{i=|U-Y_1|}^{t_1+|X_1|} |\delta_B^{U_n}(x_i) \cap [x_i]_D^{U_n}| + \sum_{i=|U_n-Y_2|}^{t_2+|X_2|} |\delta_B^{U}(y_i) \cap [y_i]_D^{U}|$, one has $E_{U\cup U_n}(D|B) = \frac{1}{(|U|+|U_n|)^2}(|U|^2 E_U(D|B) + |U_n|^2 E_{U_n}(D|B) + \Delta_1 + \Delta_2)$.

This completes the proof.

5.2 Incremental mechanism for feature significance

On the basis of incremental mechanisms shown in Theorems 1-2, feature significance for dynamic data sets is redefined in this section.

Definition 15 Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$. When a new object x is added to table S, the inner significance of feature $a \in B$ is defined as

$$Sin_{U\cup\{x\}}^{in}(a, B) = E_{U\cup\{x\}}(D|B - \{a\}) - E_{U\cup\{x\}}(D|B). \tag{18}$$

The outer significance of feature $a \in C - B$ is defined as

$$Sin_{U\cup\{x\}}^{out}(a, B) = E_{U\cup\{x\}}(D|B) - E_{U\cup\{x\}}(D|B \cup \{a\}). \tag{19}$$

Definition 16 Let $S = (U, C \cup D)$ be a data table and $B \subseteq C$. When U_n is added to U, the inner significance of feature $a \in B$ is defined as

$$Sin_{U\cup U_n}^{in}(a, B) = E_{U\cup U_n}(D|B - \{a\}) - E_{U\cup U_n}(D|B). \tag{20}$$

The outer significance of feature $a \in C - B$ is defined as

$$Sin_{U\cup U_n}^{out}(a, B) = E_{U\cup U_n}(D|B) - E_{U\cup U_n}(D|B \cup \{a\}). \tag{21}$$

For dynamic data sets, with the increase of new objects, incremental mechanisms for feature significance are effective for reducing time-consuming and improving computational efficiency of selecting useful features. Definition 15 presents the single incremental mechanism for computing feature significance. And when new data generate with a fast speed, Definition 16 gives a kind of group incremental mechanism which is able to solve feature significance more efficiently.

6 Incremental feature selection algorithm for hybrid data

Based on incremental mechanism of computing entropy and feature significance, this section introduces two incremental feature selection algorithms for dynamic hybrid data in the framework of rough set theory. For the convenience of reading, Fig. 1 shows the general framework of the main idea introduced in above sections. Algorithm 1 is a single incremental feature selection algorithm which is suitable for new object being added one by one.

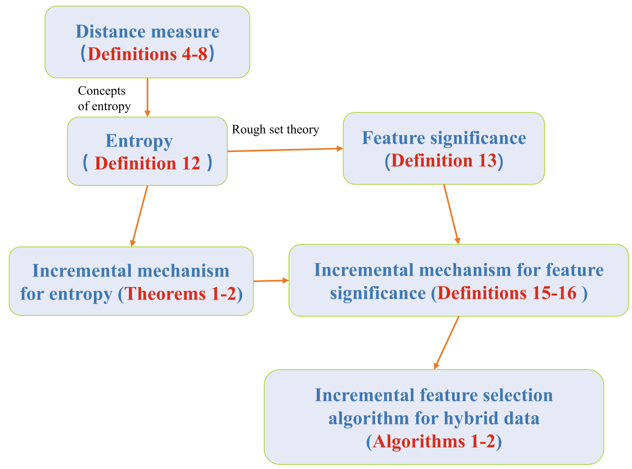

Fig. 1 Framework

Algorithm 1 An incremental feature selection algorithm for hybrid data (IFSA)

Require: A dynamic data table $S = (U, C \cup D)$, target feature subset r_1 on U, and a new object x;
Ensure: Target feature subset r on $U \cup \{x\}$;
1: $B \leftarrow r_1$. Compute $\delta'_{B,\sigma}(x)$ and $[x]'_D$ on $U \cup \{x\}$;
2: **while** $E_{U\cup\{x\}}(D|B) \neq E_{U\cup\{x\}}(D|C)$ **do**
3: $\forall a \in C - B$, compute $Sig_{U\cup\{x\}}^{out}(a, B, D)$ according to definition 15 and theorem 1;
4: Select $a_0 = max\{Sig_{U\cup\{x\}}^{out}(a, B, D)\}, a \in C - B$;
5: $B \leftarrow B \cup \{a_0\}$.
6: **end while**
7: **for all** $a \in B$ **do**
8: Compute $Sig_{U\cup\{x\}}^{in}(a, B, D)$ according to definition 15 and theorem 1;
9: **if** $Sig_{U\cup\{x\}}^{in}(a, B, D) = 0$ **then** $B \leftarrow B - \{a\}$.
10: **end if**
11: **end for**
12: $r \leftarrow B$;
13: **return** r.

The following is the time complexities of Algorithm 1. According to the analysis in Theorem 1, when x is added to the table, the time complexity of computing entropy is $O(|U||C|)$. In the algorithm $IFSA$, the time complexity of Steps 2 -6 is $O(|U||C|^2)$. In Step 7-11, the time complexity of deleting redundant features is $O(|U||C|^2)$. Hence, the total time complexity of algorithm $IFSA$ is $O(|U||C|^2)$.

Above algorithm works those data sets that new objects are increased one by one or the scale of new objects is not very large. However, many real data in applications may be generated in groups instead of one by one. If the scale of new objects is very large, above feature selection algorithm may be also time-consuming. In other words, above algorithm needs to be re-performed repeatedly to deal with multiple objects. To overcome this deficiency, following algorithm introduces a group incremental feature selection algorithm, which aims to deal with multiple objects at a time efficiently.

Algorithm 2 A group incremental feature selection algorithm for hybrid data (GIFSA)

Require: A dynamic data table $S = (U, C \cup D)$, target feature subset r_1 on U, and a group of new objects U_n;
Ensure: Target feature subset r on $U \cup U_n$;
1: $B \leftarrow r_1$. $\forall x \in U$, $y \in U_n$, compute $\delta_{B,\sigma}(x)$, $\delta_{B,\sigma}(y)$, $\delta^{U_n}_{B,\sigma}(x)$ and $\delta^{U}_{B,\sigma}(y)$;
2: **while** $E_{U \cup U_n}(D|B) \neq E_{U \cup U_n}(D|C)$ **do**
3: $\forall a \in C - B$, compute $Sig^{out}_{U \cup U_n}(a, B, D)$ according to definition 16 and theorem 2;
4: Select $a_0 = max\{Sig^{out}_{U \cup U_n}(a, B, D)\}, a \in C - B$;
5: $B \leftarrow B \cup \{a_0\}$.
6: **end while**
7: **for all** $a \in B$ **do**
8: Compute $Sig^{in}_{U \cup U_n}(a, B, D)$ according to definition 16 and theorem 2;
9: **if** $Sig^{in}_{U \cup U_n}(a, B, D) = 0$ **then** $B \leftarrow B - \{a\}$.
10: **end if**
11: **end for**
12: $r \leftarrow B$;
13: **return** r.

The time complexity of Algorithm 2 is introduced as follows. Based on the analysis in Theorem 2, when U_n is added to the table, the time complexity of computing entropy is $O(|U||C||X|)$ ($|X| = max(|X_1|, |X_2|, |Y_1|, |Y_2|)$, the explanations of X_1, X_2, Y_1, Y_2 are shown in Theorem 2). In the algorithm $GIFSA$, the time complexity of Steps 2 -6 is $O(|U||C|^2|X|)$. In Step 7-11, the time complexity of deleting redundant features is $O(|U||C|^2|X|)$. Hence, the total time complexity of algorithm $GIFSA$ is $O(|U||C|^2|X|)$.

In conclusion, when massive new objects in the databases are generated at once, the time complexity of single incremental algorithm (Algorithm 1) is $O(|U||C|^2|U_n|)$; but the time complexity of group incremental algorithm (Algorithm 2) is $O(|U||C|^2|X|)$. Because of that $|X|$ is usually much smaller than $|U_n|$, the computational time of Algorithm 2 is usually much smaller than that of Algorithm 1. In other words, the group incremental algorithm can deal with dynamic data sets more timely instead of one by one.

7 Experimental analysis

This section aims to illustrate effectiveness and efficiency of the two new algorithms $IFSA$ and $GIFSA$. The data sets employed in this section are shown in Table 1, which are all downloaded from UCI repository of machine learning databases. All the experimental programs have been carried out on a personal computer with Windows 10, Inter(R) Core (TM) i7-2600 CPU (3.4 GHz) and 8.00 GB memory. The software being used is Microsoft Visual Studio 2010 and programming language is C#. On the basis of data sets shown in Table 1, subsection 7.1 illustrates the effectiveness by comparing the two new algorithms($IFSA$ and $GIFSA$) and the non-incremental algorithm CFS. Algorithm CFS denotes the non-incremental feature selection algorithm based on complementary entropy. In subsection 7.2, efficiency of algorithms $GIFSA$ are further evaluated by comparing the computational time of algorithms $GIFSA$ and $IFSA$.

7.1 Effectiveness analysis

In this subsection, four classical classifiers and tenfold cross-validation are employed to evaluate the target feature subsets obtained by $IFSA$, $GIFSA$ and CFS. The four classifiers are decision tree(C4.5), Naive Bayes classifier(NBC), random forest (RF) and decision rule(Jrip). For each data set shown in Table 1, 60% objects are selected as the basic data set, and the rest of the objects (40%) are assumed as the incremental data. When a part of new objects are added to the basic data set, algorithms $IFSA$, $GIFSA$ and CFS are used to find the final feature subset separately. The experimental results include comparison of feature subsets, classification accuracy and computational time. Table 2 shows the number of selected features and computational time. Table 3 shows the classification accuracy on the basis of the above four classifiers. The highest accuracy values of each data set are highlighted in boldface in this table. For simplicity, N means the Number of Selected Features in Table 2.

According to the results shown in Table 2, computational time of incremental algorithms $IFSA$ and $GIFSA$ are much smaller than that of non-incremental algorithm CFS, especially, the group incremental algorithm $GIFSA$ is the most efficient one. From the results of data sets Letter and shuttle, the advantages of group incremental algorithm are clear. The computational efficiency of group incremental algorithm (GIFSA) is 90 percentage points higher than that of the classical non-incremental algorithm (CAR), and 60 percentage points higher than that of the single incremental algorithm (IFSA). Hence, for larger size data sets, the high efficiency of group incremental algorithm is more obvious. In other words, the group in incremental algorithm is helpful for deal with larger data sets. In Table 3, for each data set, its four classification accuracies induced by a same classifier are very close to each other, without obvious superiority and inferiority. Experimental results shown in Table 3 validate the feasibility of the two proposed incremental algorithm further. Compared with CAR, the two incremental algorithm also can find an effective feature subset. In summary, for a

Table 1 Description of data sets

	Data sets	*Samples*	Features Total	Numeric	Categorical	Classes
1	Credit	690	15	6	9	2
2	Anneal	898	38	6	32	6
3	Vowel	990	13	10	3	11
4	German	1000	20	7	13	2
5	kr-vs-kp	3196	36	0	36	2
6	Sick	3772	29	7	22	2
7	Hypothyroid	3772	29	6	23	4
8	Waveform	5000	40	40	0	3
9	Ticdata2000	5822	85	0	85	2
10	Thyroid	9172	29	7	22	2
11	Letter	20000	16	0	16	26
12	Shuttle	58000	9	7	2	7

Table 2 Comparison of computational time

	Data sets	CAR		$IFSA$		$GIFSA$	
		N	Time/ s	N	Time/ s	N	Time/ s
1	Credit	11	17.25	12	9.38	12	4.25
2	Anneal	12	95.75	11	49.13	11	22.53
3	Vowel	8	29.83	9	12.45	9	7.63
4	German	8	11.25	9	4.75	9	2.02
5	kr-vs-kp	29	81.36	29	14.63	29	2.39
6	Sick	7	1031.43	6	602.35	6	289.13
7	Hypothyroid	6	899.28	7	401.13	7	227.63
8	Waveform	7	2632.83	8	1402.37	8	899.25
9	Ticdata2000	24	486.93	25	128.81	25	68.97
10	Thyroid	11	9872.84	11	2461.35	11	878.95
11	Letter	11	3553.14	11	332.76	11	98.38
12	Shuttle	5	5036.76	5	1066.37	5	283.32

hybrid dynamic data set, the two proposed incremental algorithms can find an effective feature subset in very short time, especially for the group incremental algorithm $GIFSA$.

7.2 Efficiency analysis

To further illustrate the efficiency of $GIARC$, this subsection compares $GIARC$ with $IARC$ based on different sizes of incremental data. For example, let U be the size of a data set, 60% ($0.6|U|$)objects are selected as the basic data set; then 20%, 40%, $\cdots$, 100% of remaining objects ($0.2*0.4|U|$,$0.4*0.4|U|$, $\cdots$,$0.4|U|$) are used as the incremental data sets and added to the basic data in turn. When different incremental data set are added, the two incremental algorithm are employed to update the feature subset, respectively. The experimental results are shown in Figs. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. In these figures, the y-coordinate is the computational time for updating feature subsets, and the x-coordinate is the size of incremental data set, that is, coordinate value 1, 2, 3, 4 and 5 correspond to adding $0.2*0.4|U|$,$0.4*0.4|U|$, $\cdots$,$0.4|U|$, respectively.

According to the results shown in Figs. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13, one can easily get that, $GIARC$ is more efficient than $IARC$ when multiple objects are added. Furthermore, with the size of added objects increasing, the efficiency of $GIARC$ is more and more obvious. Compared with the single incremental algorithm, the group incremental algorithm saves nearly half of the computational time spent on all data sets. The data sets "letter and "kr-vs-kp" save nearly two thirds of computational time. In other words, the group incremental algorithm is very effective, especially for those larger data sets. Hence, on the basis of feasibility validated in above subsection (Subsection 7.1), experimental results shown in Figs. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Table 3 Comparison of classification accuracy

	Data sets	Algorithm	C4.5	NBC	Jrip	RF
		CAR	0.8449±0.2082	0.8231±0.1936	0.8493±0.2327	**0.8667±0.2185**
1	Credit	$IFSA$	**0.8507±0.2059**	**0.8304±0.1899**	**0.8580±0.2146**	0.8551±0.2261
		$GIFSA$	**0.8507±0.2059**	**0.8304±0.1899**	**0.8580±0.2146**	0.8551±0.2261
		CAR	0.9554±0.0171	**0.7127±0.0959**	0.9565±0.0206	**0.9688±0.0183**
2	Anneal	$IFSA$	**0.9555±0.0181**	0.7105±0.0964	**0.9577±0.0229**	0.9666±0.0109
		$GIFSA$	**0.9555±0.0181**	0.7105±0.0964	**0.9577±0.0229**	0.9666±0.0109
		CAR	**0.8081±0.0380**	0.6667±0.0842	0.7111±0.0618	0.9465±0.0479
3	Vowel	$IFSA$	0.8030±0.0384	**0.6687±0.0843**	**0.7202±0.0608**	**0.9525±0.0481**
		$GIFSA$	0.8030±0.0384	**0.6687±0.0843**	**0.7202±0.0608**	**0.9525±0.0481**
		CAR	**0.7010±0.3886**	0.7040±0.3787	0.6870±0.4082	0.6610±0.3834
4	German	$IFSA$	0.6930±0.3841	**0.7100±0.3756**	**0.6890±0.4109**	**0.6700±0.3850**
		$GIFSA$	0.6930±0.3841	**0.7100±0.3756**	**0.6890±0.4109**	**0.6700±0.3850**
		CAR	0.9011±0.1301	0.7854±0.2943	0.8758±0.1847	0.8667±0.1351
5	kr-vs-kp	$IFSA$	0.9011±0.1301	0.7854±0.2943	0.8758±0.1847	0.8667±0.1351
		$GIFSA$	0.9011±0.1301	0.7854±0.2943	0.8758±0.1847	0.8667±0.1351
		CAR	**0.9388±0.1130**	0.9356±0.1045	**0.9379±0.1047**	**0.9287±0.0901**
6	Sick	$IFSA$	0.9387±0.1150	**0.9385±0.1133**	0.9377±0.1153	0.9281±0.1041
		$GIFSA$	0.9387±0.1150	**0.9385±0.1133**	0.9377±0.1153	0.9281±0.1041
		CAR	0.9735±0.0184	0.9457±0.0339	0.9729±0.0193	**0.9716±0.0192**
7	Hypothyroid	$IFSA$	**0.9737±0.0208**	**0.9478±0.0336**	**0.9753±0.0174**	0.9552±0.0247
		$GIFSA$	**0.9737±0.0208**	**0.9478±0.0336**	**0.9753±0.0174**	0.9552±0.0247
		CAR	0.7550±0.1974	**0.7846±0.1703**	**0.7690±0.2190**	**0.7948±0.1876**
8	Waveform	$IFSA$	**0.7674±0.1974**	0.7828±0.1758	0.7580±0.2286	0.7888±0.1954
		$GIFSA$	**0.7674±0.1974**	0.7828±0.1758	0.7580±0.2286	0.7888±0.1954
		CAR	0.9395±0.1127	**0.9089±0.1283**	**0.9398±0.1118**	0.9246±0.1065
9	Ticdata2000	$IFSA$	**0.9402±0.1124**	0.8782±0.1445	0.9390±0.1128	**0.9287±0.1075**
		$GIFSA$	**0.9402±0.1124**	0.8782±0.1445	0.9390±0.1128	**0.9287±0.1075**
		CAR	0.6018±0.1586	**0.7106±0.2349**	0.6341±0.1725	0.6325±0.1525
10	Thyroid	$IFSA$	**0.6220±0.1541**	0.7034±0.2337	**0.6418±0.1696**	**0.6495±0.1529**
		$GIFSA$	**0.6220±0.1541**	0.7034±0.2337	**0.6418±0.1696**	**0.6495±0.1529**
		CAR	0.7784±0.0194	0.7298±0.0251	**0.9023±0.0262**	**0.9086±0.0231**
11	Letter	$IFSA$	**0.7848±0.0186**	**0.7320±0.0252**	0.9020±0.0281	0.9072±0.0252
		$GIFSA$	**0.7848±0.0186**	**0.7320±0.0252**	0.9020±0.0281	0.9072±0.0252
		CAR	0.9967±0.0013	**0.8804±0.0449**	0.9932±0.0028	0.9976±0.0015
12	Shuttle	$IFSA$	**0.9968±0.0013**	0.8794±0.0443	**0.9946±0.0023**	**0.9977±0.0009**
		$GIFSA$	**0.9968±0.0013**	0.8794±0.0443	**0.9946±0.0023**	**0.9977±0.0009**

and 13 validates that the group incremental reduction algorithm proposed in this paper is very efficient and suitable for dealing with larger data sets. Nowadays, with the speed of generating new data being more and more quickly, efficient and feasible incremental feature selection algorithms for big dynamic data sets are obviously necessary and significant. The group incremental mechanism gives new ways and thinking for exploring effective mining techniques.

8 Conclusions

Incremental feature selection is still a challenging issue in data mining. For hybrid data, this paper proposes two feasible and efficient incremental feature selection algorithms in the framework of neighborhood rough set theory. This paper firstly defines a new feature similarity for hybrid data and develops neighborhood relation based on the new measure. Then, a new information entropy and its incremental mechanisms are introduced to construct feature significance.

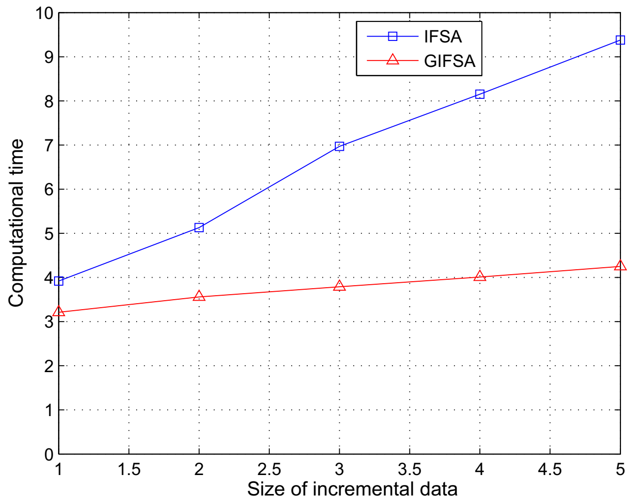

Fig. 2 Credit

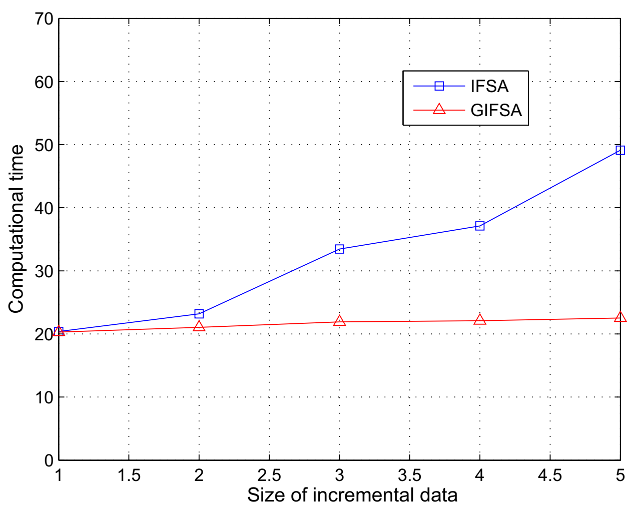

Fig. 3 Anneal

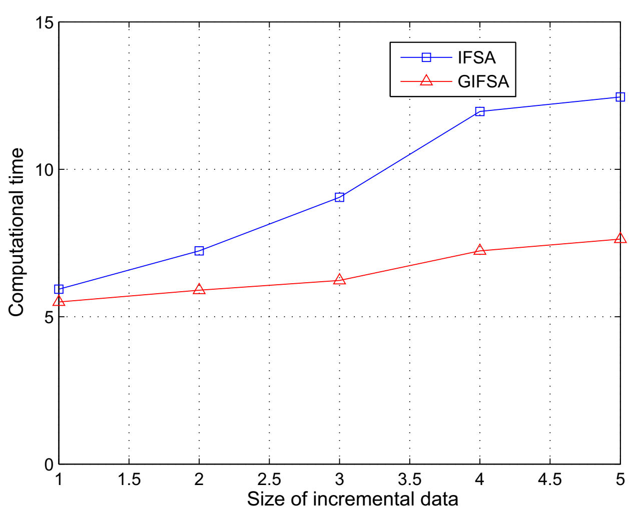

Fig. 4 Vowel

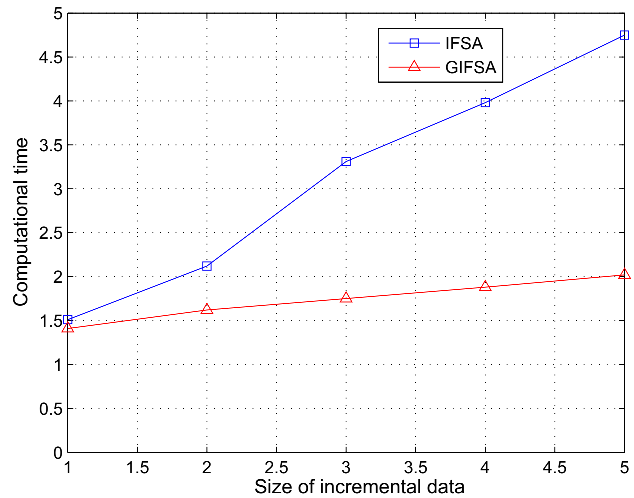

Fig. 5 German

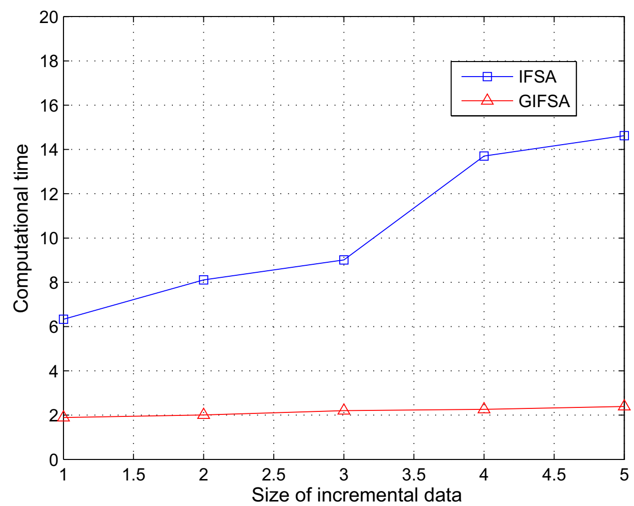

Fig. 6 kr-vs-kp

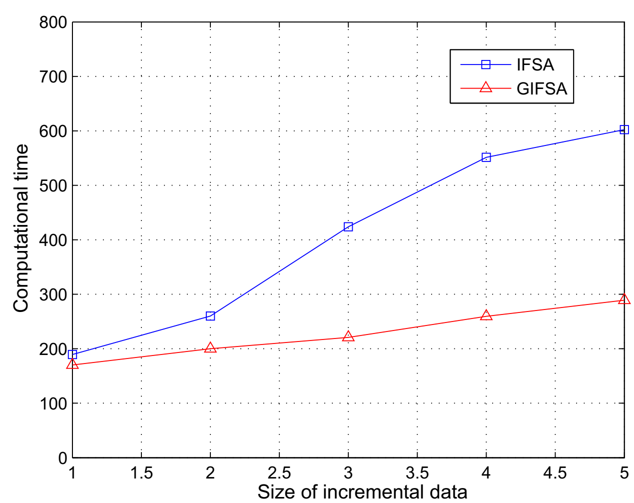

Fig. 7 Sick

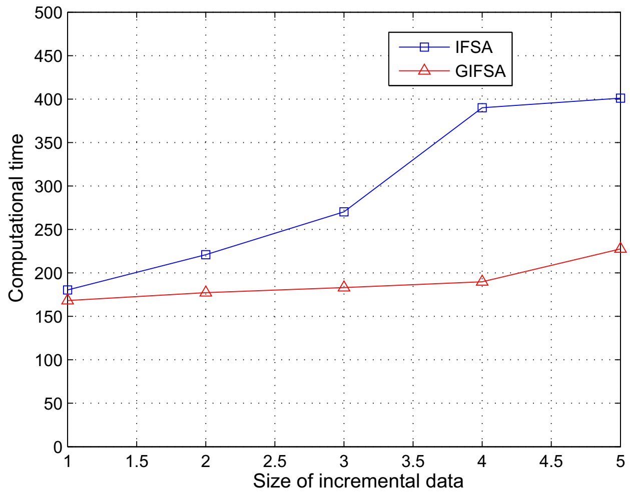

Fig. 8 Hypothyroid

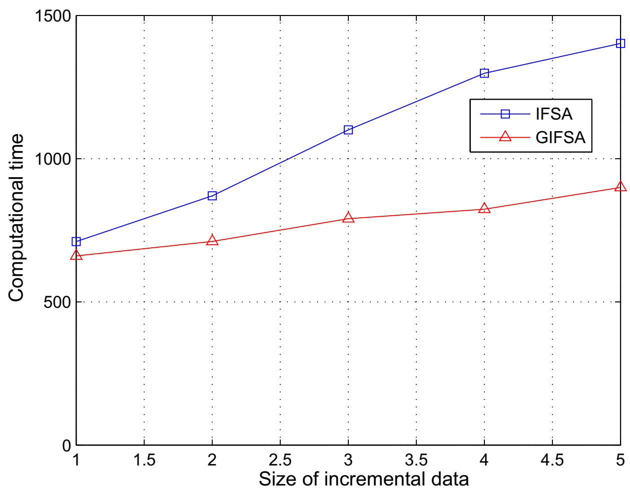

Fig. 9 Waveform

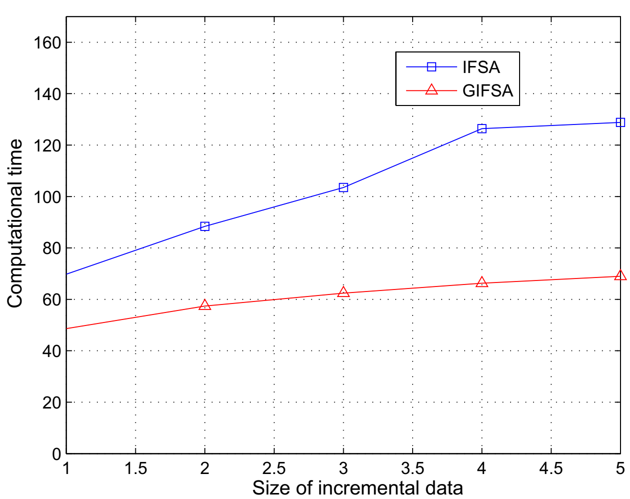

Fig. 10 Ticdata2000

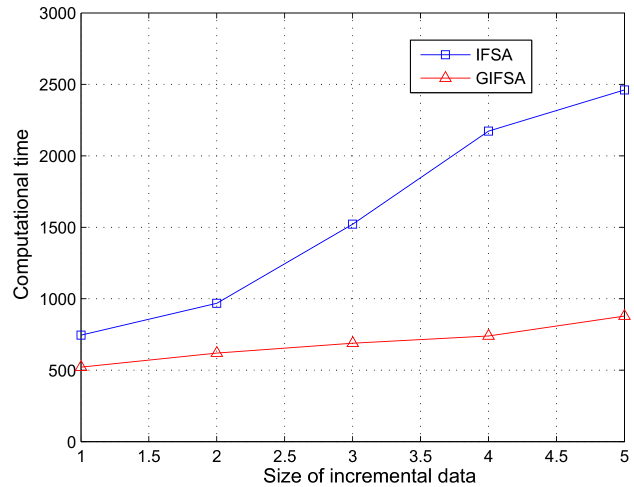

Fig. 11 Thyroid

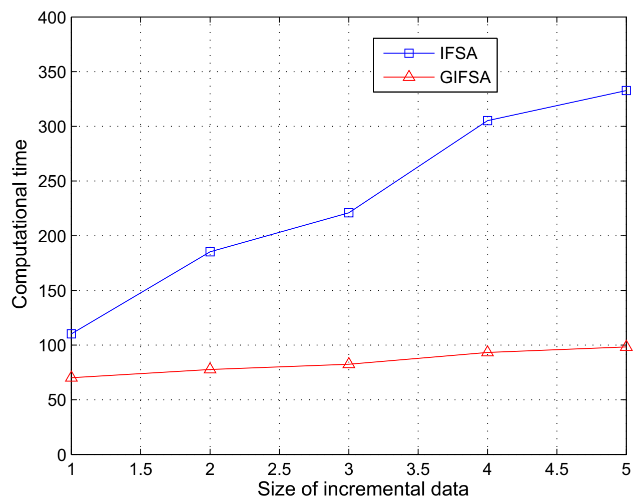

Fig. 12 Letter

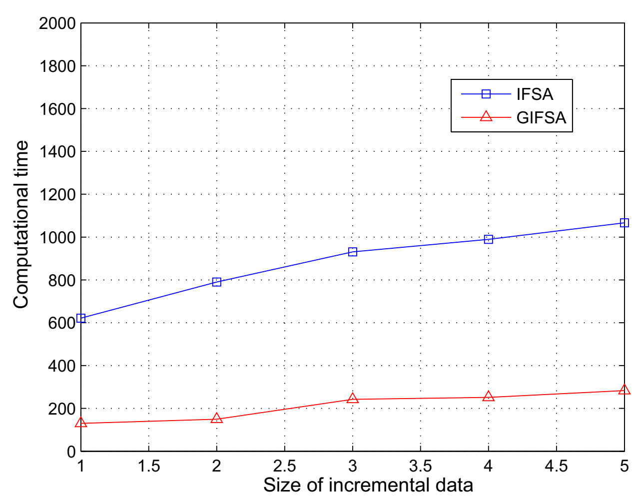

Fig. 13 Shuttle

Finally, two incremental rough feature selection algorithms are proposed to deal with dynamic hybrid data sets. Theoretical analysis and experimental results show that the new incremental algorithms are very effective, especially the efficiency of the group incremental algorithm. It is our wish that this study provides views on measuring categorical data and complex data.

Funding This study was funded by the Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi Province, China (No. 2016111), Applied Basic Research Programs of Shanxi Province (No. 201801D221170), Research Project Supported by Shanxi Scholarship Council of China(No. 2021-007) and National Natural Science Fund of China (No. 61876103).

Data Availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors certify that there is no conflict of interest with any individual/organization for the present work.

Human and animal participants This article does not contain any studies with human participants performed by any of the authors.

Informed consent In addition, informed consent was obtained from all individual participants included in the study.

References

Almuallim H, Dietterich TG (1994) Learning boolean concepts in the presence of many irrelevant features. Artif Intell 69(1–2):279–305

Benabdeslem K, Hindawi M (2004) Efficient semi-supervised feature selection: constraint, relevance and redundancy. IEEE Trans Knowl Data Eng 26:1131–1143

Chen HM, Li TR, Fan X, Luo C (2019) Feature selection for imbalanced data based on neighborhood rough sets. Inf Sci 483:1–20

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156

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

Gama J (2012) A survey on learning from data streams: current and future trends. Prog Artif Intell 1(1):45–55

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

Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the 17th International Conference on Machine Learning 359-366

Hu QH, Xie ZX, Yu DR (2007) Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recogn 40:3509–3521

Hu QH, Yu DR, Xie ZX (2006) Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recogn Lett 27(5):414–423

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

Huang QQ, Li TR, Huang YY, Yang X (2020) Incremental three-way neighborhood approach for dynamic incomplete hybrid data. Inf Sci 541:98–122

Jing YG, Li TR, Fujita H et al (2017) An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view. Inf Sci 411:23–38

Kwak N, Choi CH (2002) Input feature selection by mutual information based on Parzen window. IEEE Trans Pattern Anal Mach Intell 24(12):1667–1671

Liu H, Hussain F, Dash M (2002) Discretization: an enabling technique. Data Min Knowl Discov 6(4):393C423

Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502

Li TR, Ruan D, Geert W, Song J, Xu Y (2007) A rough sets based characteristic relation approach for dynamic attribute generalization in data mining. Knowl-Based Syst 20(5):485–494

Liu D, Li TR, Ruan D, Zou WL (2009) An incremental approach for inducing knowledge from dynamic information systems. Fundamenta Inform 94:245C260

Liang JY, Wang F, Dang CY, Qian YH (2014) A group incremental approach to feature selection applying rough set technique. IEEE Trans Knowl Data Eng 26(2):294–308

Lu J, Liu AJ, Dong F, Gu F et al (2019) Learning under concept drift: a review. IEEE Trans Knowl Data Eng 31(12):2346–2363

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

Liang JY, Chin KS, Dang CY, Yam Richid CMA (2002) new method for measuring uncertainty and fuzziness in rough set theory. Int J Gen Syst 31(4):331–342

Liu JH, Lin YJ, Li YW et al (2018) Online multi-label streaming feature selection based on neighborhood rough set. Pattern Recogn 84:273–287

Mera C, Alzate MO, Branch J (2019) Incremental learning of concept drift in multiple instance learning for industrial visual inspection. Comput Ind 109:153–164

Neto AF, Canuto AMP (2021) EOCD: an ensemble optimization approach for concept drift applications. Inf Sci 561:81–100

Pedrycz W, Vukovich G (2002) Feature analysis through information granulation and fuzzy sets. Pattern Recogn 35:825–834

Paul J, Ambrosio RD, Dupont P (2015) Kernel methods for heterogeneous feature selection. Neurocomputing 169:187–195

Pawlak Z (1998) Rough set theory and its applications in data analysis. Cybern Syst 29:661–688

Pawlak Z, Skowron A (2007) Rough sets and boolean reasoning. Inf Sci 177(1):41–73

Swiniarski RW, Skowron A (2003) Rough set methods in feature selection and recognition. Pattern Recogn Lett 24:833–849

Shu WH, Qian WB, Xie YH et al (2019) Incremental approaches for feature selection from dynamic data with the variation of multiple objects. Knowl-Based Syst 163:320–331

Shao MW, Zhang WX (2005) Dominance relation and rules in an incomplete ordered information system. Int J Intell Syst 20:13–27

Tang WY, Mao KZ (2007) Feature selection algorithm for mixed data with both nominal and continuous features. Pattern Recogn Lett 28(5):563–571

Chang CZ, Shi YP, Fan XD, Shao MW (2019) Attribute reduction based on k-nearest neighborhood rough sets. Int J Approx Reason 106:18–31

Wang F, Liang JY, Qian YH (2013) Attribute reduction: a dimension incremental strategy. Knowl-Based Syst 39:95–108

Wang F, Liang JY, Dang CY (2013) Attribute reduction for dynamic data sets. Appl Soft Comput 13:676–689

Wang F, Liang JY (2016) An efficient feature selection algorithm for hybrid data. Neurocomputing 193:33–41

Wang H (2006) Nearest neighbors by neighborhood counting. IEEE Trans Pattern Anal Mach Intell 28(6):942–953

Wei W, Liang JY, Qian YH, Wang F (2009) An attribute reduction approach and its accelerated version for hybrid data. In: The 8th IEEE International Conference on Cognitive Informatics 167-173
Wu WZ, Mi JS, Zhang WX (2003) Generalized fuzzy rough sets. Inf Sci 151:263–282
Xu WH, Zhang XY, Zhang WX (2009) Knowledge granulation, knowledge entropy and knowledge uncertainty measure in ordered information systems. Appl Soft Comput 9(4):1244–1251
Yang X, Liu D, Yang XB, Liu KY, Li TR (2021) Incremental fuzzy probability decision-theoretic approaches to dynamic three-way approximations. Inf Sci 550:71–90
Yao YY (2006) Neighborhood systems and approximate retrieval. Inf Sci 176(23):3431–3452
Yao YY, Zhao Y (2008) Attribute reduction in decision-theoretic rough set models. Inf Sci 178(17):3356–3373
Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
Zhao H, Qin KY (2014) Mixed feature selection in incomplete decision table. Knowl Based Syst 57:181–190
Zhou P, Hu XG, Li PP, Wu XD (2019) Online streaming feature selection using adapted neighborhood rough set. Inf Sci 481:258–279