Applied Soft Computing 113 (2021) 107993

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Online early terminated streaming feature selection based on Rough Set theory

Peng Zhou [a,b,c], Peipei Li [d,e,*], Shu Zhao [a,b,c], Yanping Zhang [a,b,c]

[a] Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui Province 230601, PR China
[b] School of Computer Science and Technology, Anhui University, Hefei, Anhui Province 230601, PR China
[c] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Province 230601, PR China
[d] Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Anhui Province 230601, PR China
[e] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui Province 230601, PR China

ARTICLE INFO

Article history:
Received 14 July 2021
Received in revised form 28 September 2021
Accepted 8 October 2021
Available online 23 October 2021

Keywords:
Online feature selection
Streaming features
Early terminated
Rough Set
Dependency degree

ABSTRACT

Feature selection is a vital dimensionality reduction technology for machine learning and data mining that aims to select a minimal subset from the original feature space. Traditional feature selection methods assume that all features can be required before learning, while features may exist in a stream mode for some real-world applications. Therefore, online streaming feature selection was proposed to handle streaming features on the fly. When the feature dimension is extraordinarily high or even infinite, it is time-consuming or impractical to wait for all the streaming features to arrive. Motivated by this, we study and solve the exciting issue of whether we can terminate the online streaming feature selection early for efficiency while maintaining satisfactory performance for the first time. Specifically, we first formally define the problem of online early terminated streaming feature selection and summary two properties that the early terminated mapping function should satisfy. Then we choose the dependency degree function in Rough Set theory as our early terminated mapping function and demonstrate that it satisfies the two properties. Based on this, we propose a novel Early Terminated Online Streaming Feature Selection framework, named OSFS-ET, which could terminate the streaming feature selection early before the end of streaming features and guarantee a competing performance with the currently selected features. Extensive experiments on twelve real-world datasets demonstrate that OSFS-ET can be far faster than state-of-the-art streaming feature selection methods while maintaining excellent performance on predictive accuracy.

Code metadata

Permanent link to reproducible Capsule: https://codeocean.com/capsule/8154265/tree/v1.

1. Introduction

Feature selection aims to select a minimal subset from the original feature space that can retain the optimum salient characteristics and is treated as a data preprocessing component for high-dimensional datasets before machine learning and data mining [1]. From the perspective of whether using classifiers or not during feature selection, we can divide feature selection methods into three categories: filter, wrapper, and embedded [2]. Filter methods select the features in terms of specific feature measurements, while wrapper methods use predefined classifiers as a black box to evaluate the selected features [3]. Embedded methods perform feature selection in the process of model construction [4]. From a data perspective, we can divide feature selection into traditional feature selection methods for static data, and online feature selection methods for stream data [5].

Traditional feature selection has been studied for decades, and there are many papers in this research area [5–8]. In general, traditional feature selection methods select features from the feature space according to specific strategies and assume that all features can be required before learning. However, the feature space may be unknown in advance, and features can exist in a stream mode for some real-world applications, such as impact crater detection [9] and multiple descriptors image analysis [10]. Besides, as the increasing of data volume and dimensionality [11], it cannot load all the data into memory at once before feature selection for some big datasets. Therefore, for big datasets, they are more

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Correspondence to: 193 Tunxi Road, Hefei, Anhui, PR China.
E-mail addresses: doodzhou@ahu.edu.cn (P. Zhou), peipeili@hfut.edu.cn (P. Li), zhaoshuzs@ahu.edu.cn (S. Zhao), 87025@ahu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.asoc.2021.107993
1568-4946/

appropriate to be processed by rows (samples) or by columns (features) in feature selection [12,13]. Stream data can be further divided into the instance stream and feature stream [5]. In this paper, we focus on feature selection with feature streams. Feature streams are defined as the features flow in one by one over time, whereas the number of instances remains fixed [14]. Online streaming feature selection deals with feature streams on the fly and can handle extraordinarily high-dimensional or infinite feature space datasets [15]. The flow charts of traditional feature selection and online streaming feature selection are shown as Fig. 1(a) and (b), respectively. There are two main differences between traditional feature selection and online streaming feature selection. First, the former can require all features in the dataset before learning, while the latter just gets one streaming feature once without the information of the entire feature space. Second, the former can compare each feature in the feature space multiple times while the latter must decide to retain or discard the new arriving feature immediately. Once a feature is discarded, it cannot be selected or used again. Thus, it is more challenging for the latter to retain optimal features on the fly.

Specifically, Perkins et al. [16] first proposed the problem of online feature selection with stream features. Based on streamwise regression, Zhou et al. [17] proposed two penalized likelihood ratio methods for the online streaming feature selection problem. Wu et al. [14] proposed the OSFS (Online Streaming Feature Selection) framework that consists of two main components: online relevance analysis and online redundancy analysis. Considering the feature interaction between features during online streaming feature selection, Zhou et al. [18] proposed a new method that could select features to interact with each other. Besides, many researchers have recently begun applying the Rough Set theory for online streaming feature selection, such as the classical Rough Set based method OS-NRRSARA-SA [19], and the neighborhood Rough Set based methods OFS-A3M [20] and OFS-Density [21]. In terms of different technologies and theories, all these methods mentioned above have been proved efficiently in the experiments. However, if the feature space is extraordinarily immense or even infinite, is it necessary to wait for all the streaming features to arrive? For example, in bioinformatics [22], for the high cost of conducting wet-lab experiments, acquiring the complete set of features for every training instance is prohibitive, and it is impossible to wait for a complete set of features. If we can terminate the selection early with currently generated streaming features, it could save the time cost. Thus, it is an exciting issue to study whether we can terminate the streaming feature selection early for efficiency while maintaining satisfactory performance. There are three main reasons for early terminated online streaming feature selection. (1) The streaming features are infinite, and it is impractical to wait for all features to arrive. (2) The currently selected features are "good enough", so it is unnecessary to wait for more features. (3) The time cost of waiting is too high to have to end the selection as soon as possible. In other words, if the currently selected features are "good enough" and the expectation of the increase using the future arriving features is much lower than the cost of the time consumption, it is worthy of terminating the selection immediately.

The flow chart of early terminated online streaming feature selection is shown as Fig. 1(c). The main difference between online streaming feature selection and early terminated online streaming feature selection lies in whether the method contains a stopping criterion that can terminate the selection early before the end of streaming features. However, there are three main challenges for online early terminated streaming feature selection. First, we cannot provide a fixed stopping criterion in advance without the complete information of feature space before learning. In other words, we cannot terminate the streaming feature selection with some specific thresholds. Second, during online streaming feature selection, which features will arrive in the upcoming period can be independent of the history. The values of the target mapping function on the currently selected features may fluctuate. Thus, it is not easy to choose an adequately terminated timestamp. Third, once we terminate the streaming selection, all the following arriving features will be discarded directly. Therefore, it is a challenge to ensure the currently selected features are good enough. Besides, for some real-world applications, such as bioinformatics, feature generation is expensive or time-consuming for the target instances. Thus, if the streaming features are extraordinarily immense or even infinite, it is not cost-effective or impractical to wait for all streaming features to arrive. Thus, it is essential to address these challenges for online early terminated streaming feature selection.

Without the information of the entire feature space, streaming feature selection methods usually apply filter mode, which selects the features in terms of a specific feature mapping function [15]. The main target of online streaming feature selection methods is to maximize the value of the mapping function with the arrived features at each time. Therefore, if the expectation of the increase in mapping function is much lower than the cost of time consumption, then we can terminate the selection immediately for efficiency. In other words, at timestamp t, if the value of the mapping function on the currently selected features is "big enough" and the expectation of the increase for the future arriving features is much lower than the cost of time consumption, we should terminate the selection immediately. Besides, as we know, feature selection is an NP-hard problem. Therefore, most of filter mode based feature selection methods adopt greedy strategies and choose the best feature in each round [6]. With a specific mapping function h, we aim to maximize $h(S^t)$ where S^t is the selected feature subset at timestamp t. Thus, to terminate the selection safely, we point out two properties that an early terminated mapping function must satisfy: (1) having an upper bound and (2) being non-decreasing. Property (1) is used to judge whether the currently selected feature is "good enough", while property (2) makes it possible to predict the future growth of the early terminated mapping function.

Recently, a batch of Rough Set-based online streaming feature selection methods [19–21,23] have been proposed for the two most important advantages: (1) Rough Set based data mining methods do not require any domain knowledge other than the given dataset; (2) By using the dependency degree of the candidate feature subset, these methods can measure the selected feature subset as an integral. The main target of Rough Set-based online streaming feature selection methods is to maximize the dependency degree of the selected feature subset at each timestamp. In this paper, we choose the dependency degree as our early terminated mapping function and demonstrate that it satisfies the two properties in Theorem 3. Theoretically, the maximal value of the dependency degree can reach one. However, the increase in the dependency degree of selected feature subset is not linear and smooth, and it may no longer grow after a particular timestamp. For example, we apply the baseline Rough Set based online streaming feature selection method RS-OSFS (as shown in Algorithm 1) on dataset Nova (from WCCI 2006, the number of features is 16,969). Suppose we only get one feature at each timestamp. Fig. 2 reports the growth trend of the dependency degree with the selected feature subset varying with different timestamps (the number of arrived streaming features).

The dependency degree of the selected feature subset only increases 0.0092 from 8000 to 10,000. Besides, the increase in the dependency degree is less than 0.013 for the next 6969 features (from 10,001 to 16,969). Thus, the motivation for terminating the streaming feature selection early before the end is that the increase of dependency degree is much lower than the cost of time

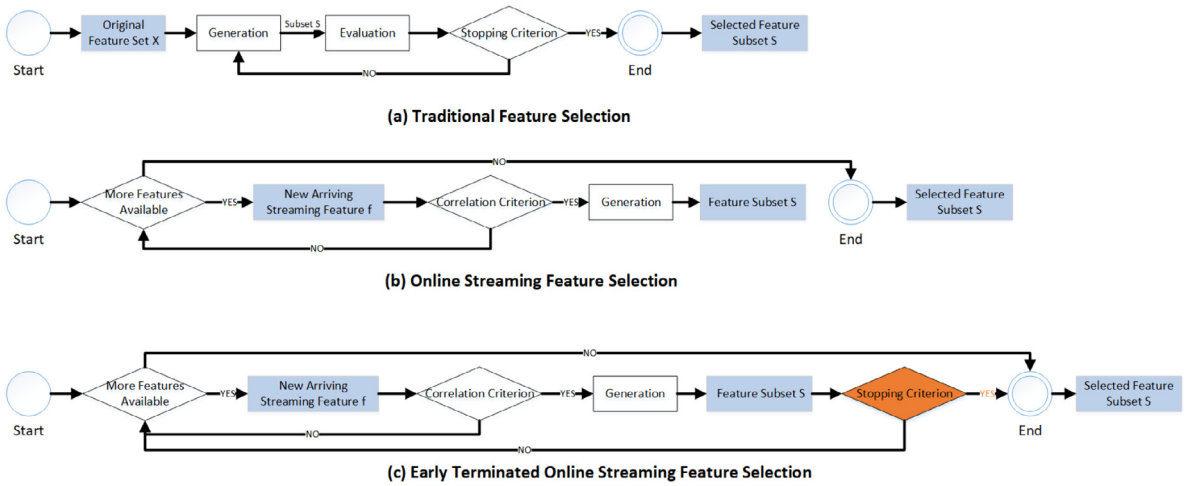

Fig. 1. Flow charts of traditional feature selection (a), online streaming feature selection (b) and early terminated online streaming feature selection (c).

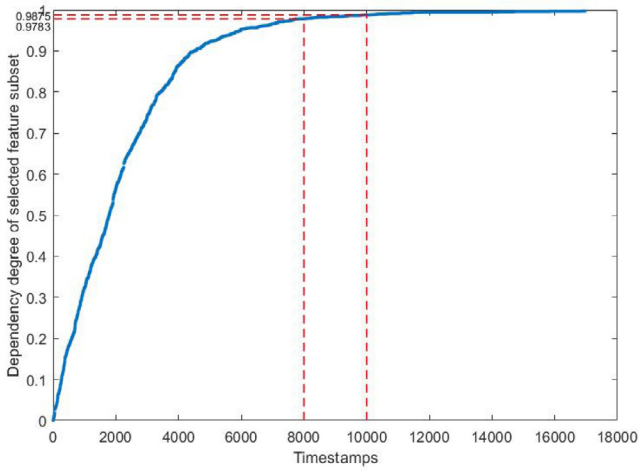

Fig. 2. The growth trend of dependency degree on Nova varying with different timestamps. With the arriving of half of all features (around timestamp 8000), the dependency degree of the selected feature subset nearly reaches the maximal value 1.

consumption. In other words, if we can terminate the streaming feature selection early while guaranteeing the competing performance of the currently selected feature subset, it is worthy of doing it. Nevertheless, this interesting problem has not been studied as far as we know.

In this paper, we study how to terminate the streaming feature selection early and propose a new Early Terminated Online Streaming Feature Selection framework, named OSFS-ET. The main innovations and contributions of this paper are as follows:

- We first propose this novel issue of early terminated online streaming feature selection. With the in-depth analysis of why we should terminate the streaming feature selection early, we present a formal definition of this issue for the first time. As far as we know, none of the existing online streaming feature selection methods can automatically terminate the selection early before the end of streaming features.
- To terminate online streaming feature selection early and safely, we summarize two properties for the early terminated mapping function which should satisfy. Meanwhile, we prove the finite limit for the satisfied mapping function even though the feature streams are infinite.
- We choose the dependency degree function in Rough Set theory as our early terminated mapping function and demonstrate that it satisfies the two properties. Based on this, we propose a new framework OSFS-ET that can terminate the online streaming feature selection early while maintaining a competing performance using the currently selected features.
- We conduct extensive experimental comparisons between OSFS-ET and seven state-of-the-art competing online streaming feature selection algorithms on twelve real-world datasets. Experiment results demonstrate that our new early terminated method can be far faster than the competing streaming feature selection algorithms while maintaining an outstanding performance on predictive accuracy.

The rest of this paper is organized as follows. Section 2 briefly introduces some traditional feature selection methods and focuses on the related works of online streaming feature selection. In terms of different theories and technologies, we divide all these online streaming feature selection methods into Rough Set-based and non-Rough Set-based. In Section 3, we first present the formal definition of online early terminated online streaming feature selection. Then, we analyze the main properties of the early terminated mapping function that should satisfy and choose the dependency degree function in Rough Set Theory in this paper. Based on these, we propose the new early terminated online streaming feature selection framework OSFS-ET. Experimental analyses are presented in Section 4, including the experiment settings, the parameter analysis of OSFS-ET, and the experimental comparison between OSFS-ET and seven state-of-the-art competing algorithms. In Section 5, we make a brief conclusion of this paper, point out the limitations of the proposed framework, and give some suggestions in our future work.

2. Related work

Feature selection aims to select a minimal subset from the original feature space and is essential to speed up learning and improve concept quality [1]. According to different data types, we can divide feature selection into two categories: traditional feature selection for static data and online feature selection for stream data [5].

2.1. Traditional feature selection for static data

For static data, traditional feature selection has been studied for decades, and there are many papers in this research

area [7]. Considering whether using a classifier or not in feature selection, we can further divide feature selection into three categories: filter, wrapper and embedded [2]. For example, Wei et al. [3] designed a new measure named Dynamic Feature Importance (DFI) and its corresponding feature selection algorithm for high-dimensional and small-sample-size data. In order to obtain higher classification accuracy with a smaller number of features, an effective hybrid feature selection framework was proposed. Manikandan et al. [24] propose an efficient feature selection framework based on mutual information and Monte Carlo tree, which aims to select optimal features from the high dimensional datasets. Besides, based on the Rough Set theory, many feature selection (attribute reduction) methods were proposed in the literature [25]. Recently, Yuan et al. [8] gave a comprehensive review of attribute reduction methods based on fuzzy Rough Set theory.

2.2. Online feature selection for stream data

However, for some real-world applications, data may exist in a stream mode. There are two types of stream data: instance stream and feature stream [5]. The instance stream assumes that the number of features on training data is fixed while the number of instances changes over time [26,27]. This paper focuses on the online feature selection problem with feature streams, which assumes the features exist in a stream mode and flows in one by one over time, whereas the number of instances remains fixed [15]. Based on feature selection whether using Rough Set theory, we further divide online streaming feature selection into Rough Set-based methods and non-Rough Set-based methods.

2.2.1. Non-rough set-based online streaming feature selection

Grafting [16] was the first method which deals with the problem of online feature selection. Based on the stagewise gradient descent, Grafting was an embedded method that treats feature selection as an integral part of learning a predictor within a regularized framework. Information-investing and alpha-investing [17] were two penalized likelihood ratio methods that focus on the online streaming feature selection problem. Based on streamwise regression, these two methods did not need to determine any prior parameters in advance and run very fast. OSFS (Online Streaming Feature Selection) and fast-OSFS [14] were two online streaming feature selection methods based on conditional independence/dependence tests. In terms of online relevance analysis (discarding irrelevant features) and online redundancy analysis (eliminating redundant features), these two methods selected a very compact feature subset. SAOLA [28] (a Scalable and Accurate Online feature selection Approach) was a novel online pairwise comparison technique based method that aimed to addresses two challenges in many big data applications: extremely high dimensionality and its highly scalable requirement of feature selection. GFSSF [29] was Mutual Information based method that could work on both the group and individual feature levels, by exploiting entropy and mutual information in information theories. OSFSMI and OSFSMI-k [30] were two Mutual Information based streaming feature selection algorithms that attempt to handle the challenges of high computational cost, the stability of the generated results, and the size of the final features subset. SFS-FI [18] considered the interaction between features during online streaming feature selection and proposed a new method that can select features to interact with each other. All these methods mentioned above aim to handle the online streaming feature selection problem on the fly and have been demonstrated to be efficient in the experiments. Inspired by these methods, our new online streaming feature selection framework also consists of two main components: relevant feature selection and redundancy feature removal.

Recently, some new works have studied the online streaming feature selection from other perspectives. For example, GF-CSF [31] conducted online feature selection from capricious streaming features, where features flow in one by one with some random missing entries while the number of data instances remains fixed. GF-CSF adopted latent factor analysis to preprocess capricious streaming features for completing their missing entries before conducting feature selection. I-SFS and G-SFS [32] were two streaming feature selection methods for multi-label data where the multiple labels are reduced to a lower-dimensional space. These two methods grouped the similar labels before performing the selection method to improve the selection quality and make the model efficient. LOSSA [33] was a latent-factor-analysis-based online sparse-streaming-feature selection algorithm, which aims to implement online feature selection from sparse streaming features. The main idea of LOSSA is to apply latent factor analysis to pre-estimate missing data in sparse streaming features before conducting feature selection, thereby addressing the missing data issue effectively and efficiently. OCFSSFs [34] was an online causal feature selection method for streaming features through mining Markov blanket containing parents and children (PC) and spouse. Furthermore, OCFSSFs distinguished PC and spouse in real-time and could identify children with parents online when identifying spouses.

2.2.2. Rough set-based online streaming feature selection

Without requiring any additional information, the Rough Set theory can reduce dimensionality using intrinsic information within the data. There are many feature selection methods based on the classical Rough Set model and its extended models (such as the Neighborhood Rough Set model, Probabilistic Rough Set model, and Fuzzy Rough Set model) [8,35]. Besides, many researchers have recently begun applying the Rough Set theory for online streaming feature selection. Specifically, [36] was one of the early works based on fuzzy-rough set theory, in an attempt to deal with feature selection scenarios where features and instances may be dynamically added or removed throughout the training process. OS-NRRSARA-SA [19] was proposed as a classical Rough Set based method for online streaming feature selection which considers both the boundary and positive regions. Meanwhile, OS-NRRSARA-SA uses a noise-resistant dependency measure to search for reduces. CIE-OSFS [37] was an OSFS based uncertainty measure framework to address the online streaming feature selection problem from the Rough Set perspective. By specifying the uncertainty measure with conditional information entropy (CIE), CIE-OSFS did not need prior knowledge to deliver credible results and was robust to changing streaming orders. K-OFSD [23] was a k-nearest neighborhood relation based online streaming feature selection method from the Neighborhood Rough Set perspective for high-dimensional and class-imbalanced data. With the information of K nearest neighbors, K-OFSD selected relevant features that can get a higher separability between the majority class and the minority class. OFS-A3M [20] was a new non-parametric online streaming feature selection method that need not specify any optimal parameter values before learning. OFS-A3M can select features with high correlation, high dependency, and low redundancy in terms of the maximal-dependency, maximal-relevance, and maximal-significance evaluation criteria. OFS-Density [21] was proposed for online streaming feature selection where the sample distribution of instances is usually not uniform. Meanwhile, Rough Set-based methods usually use the increase of dependency degree to select features while the exactly equal constraint is too strict for real-world applications during streaming feature selection. With the density information of the surrounding instances and the fuzzy equal constraint, OFS-Density can select features with

low redundancy. These Rough Set-based methods have demonstrated the effectiveness of applying Rough Set theory for the online streaming feature selection problem. Therefore, this paper applied the dependency degree function in the Rough Set theory to construct our new early terminated framework.

Table 1 summarizes the analysis of some existing streaming feature selection methods, including four main properties: feature selection type, core technology/theory, the main advantages, and disadvantages.

For real-world streaming feature selection applications, the features are generated one by one over time with specific time intervals. Therefore, when the feature space is extraordinarily immense or even infinite, it will take a very long time or impractical to wait for all the features to arrive. However, as far as we know, none of the existing online streaming feature selection methods can automatically terminate the selection early before the end of streaming features. Thus, this prompted us to study whether we could terminate the streaming feature selection early before the end.

3. The proposed framework

This section first defines online streaming feature selection and early terminated online streaming feature selection. With the in-depth analysis of the reasons for early termination, we point out two properties that the mapping function should satisfy to terminate the selection before the end while maintaining competing performance. Then, we introduce the dependency degree in Rough Set theory and demonstrate that it satisfies these two early terminated properties. After that, we propose our new early terminated online streaming feature selection framework, named OSFS-ET. Finally, we discuss the time complexity of our new framework.

3.1. Problem definition

Definition 1 (*Online Streaming Feature Selection*)**.** Let $\{f_t|t = 1,\ldots,T\}$ be a sequence of streaming features, where $f_t \in R^{n\times 1}$ is a pattern of n samples received at the tth time. $D \in R^{n\times 1}$ is the observed class labels of the n samples. At timestamp t, we get a new feature f_t and should decide whether retain or discard f_t on the fly. Meanwhile, the discarded features cannot be selected again. Suppose the selected feature subset is S^t at timestamp t, the number of features in S^t is s ($s \le t$), and $h : R^{n\times s} \to R$ is the mapping function which measure the quality of selected features. Online streaming feature selection aims to maximize the quality of selected features at each timestamp t, defined as follows:

$$Max \quad h(S^t). \tag{1}$$

For streaming features, if the feature space is extraordinarily immense or even infinite, is it necessary to wait for all the features to arrive? There are three main reasons for early terminated online streaming feature selection. (1) The streaming features are infinite, and it is impractical to wait for all features to arrive. (2) The currently selected features have satisfied the requirement, so it is unnecessary to wait for more features and waste much more time. (3) The time cost of waiting is too high to have to end the selection as soon as possible. In other words, if the currently selected features are good enough and the expectation of the increase for the future arriving features is much lower than the cost of time consumption, it is worthy of terminating the selection immediately. Thus, it is an exciting issue to study whether we can terminate the streaming feature selection early for efficiency while maintaining satisfactory performance.

Definition 2 (*Early Terminated Online Streaming Feature Selection*)**.** For a streaming feature selection method, suppose the mapping function is $h(S^t)$ for the selected feature subset S^t at timestamp t. Let $\phi(t, t+w)$ denote the expected increase of mapping function from timestamp t to $t+w$, and $\varphi(t, t+w)$ denote the cost of time consumption from timestamp t to timestamp $t+w$. Then, we should **terminate** the selection at timestamp t and return the currently selected feature subset S^t if:

- $h(S^t)$ satisfies the requirement;
- OR $\phi(t, t+w) < \varphi(t, t+w)$.

For filter mode feature selection methods, them usually adopt greedy strategies. In each round, they always choose the best features so far. With a specific mapping function h, we aim to increase $h(S)$ with the new selected features. During online streaming feature selection, which features will arrive in the upcoming period can be independent of the history. Thus, to terminate the selection safely, the mapping function $h()$ must have an upper bound and be non-decreasing.

Definition 3 (*Early Terminated Mapping Function*)**.** For a streaming feature selection method, suppose the mapping function is $h(S^t)$ for the selected feature subset S^t at timestamp t. In order to terminate the selection safely, $h()$ must satisfy the following two properties:

- 1°. For $\forall S^t$, $h(S^t) \le \lambda$, where λ is a definite constant;
- 2°. For $\forall S^t$, $h(S^t) \le h(S^{t+1})$.

We call $h()$ is an **early terminated mapping function**.

Specifically, the primary motivation that we should terminate the selection early is (1) the currently selected features are "good enough"; or (2) the expected increase in mapping function is lower than the cost of time consumption. Thus, property 1° is used to judge whether the currently selected feature is "good enough", while property 2° makes it possible to predict the future growth of the early terminated mapping function.

Theorem 1. *For $\forall t \in \{1, 2, \ldots, +\infty\}$, suppose $h(S^{t+1}) \ge h(S^t)$ and $h(S^t) \le \lambda$, then when $t \to +\infty$, $h(S^t)$ has a finite limit.*

Proof. For $\forall t$, $h(S^t) \le \lambda$, $\therefore$ $h(S^t)$ has a supremum and suppose it is A. Given an arbitrary number $\varepsilon > 0$, $\because$ $Sup\{h(S^t)\} = A$, $\therefore$ there must have a t' which makes $h(S^{t'}) > A - \varepsilon$. $\because$ For $\forall t$, $h(S^{t+1}) \ge h(S^t)$, $\therefore$ when $t > t'$, $h(S^t) \ge h(S^{t'}) > A - \varepsilon$. Meanwhile, $\because$ For $\forall t$, $h(S^t) \le A < A + \varepsilon$, $\therefore$ when $t > t'$, $|h(S^t) - A| < \varepsilon$. $\therefore$ $h(S^t)$ has a finite limit.

According to Theorem 1, even the feature streams are infinite, and they always have a finite limit for the satisfied mapping function. Thus, with an appropriate mapping function, we can predict the expectation and terminate the selection correctly.

3.2. A satisfied early terminated mapping function

First, let us take a brief introduction to the Rough Set theory and the definition of the dependency degree. For the classical Rough Set model [38], in terms of feature subset B, the objects with the same feature values are drawn together and form an equivalence class, denoted by $[x]_B$. The family of elemental granules $\{[x_i]_B \mid x_i \in U\}$ builds a concept system to describe an arbitrary subset of the sample space. Let $U = \{x_1, x_2, \ldots, x_n\}$ denote a nonempty set which contains all the samples, where x_i denotes the ith sample. For feature subset B and sample subset $X (X \subseteq U)$, the elemental granules of lower approximation and upper approximation are defined as follows:

$$\underline{B}X = \{[x_i]_B \mid [x_i]_B \subseteq X, x_i \in U\} \tag{2}$$

Table 1
Analysis of some existing streaming feature selection methods.

Methods	Type	Technique	Main advantages	Main disadvantages
Grafting	Embedded	Stagewise gradient descent	Can have a single global optimal solution	High time cost and difficult to choose a good value for the regularization parameter
Alpha-investing	Filter	Penalized likelihood ratio	Easy to implement and run very fast	Does not reevaluate the selected features
OSFS	Filter	Conditional independence an dependence tests	Can achieve better compactness	High time complexity and need a large number of training instances
SAOLA	Filter	Mutual Information	Can handle extremely high dimensionality data and highly scalable	Only consider pairwise correlations between features
GFSSF	Filter	Mutual Information	Work on both the group and individual feature levels	Need to specify parameter in advance and can only handle discrete data
OSFSMI	Filter	Mutual information	Low computational cost and can generate more robust results	Cannot consider the selected features as an integral
OS-NRRSARA-SA	Filter	Classical rough set	Does not require any domain knowledge	Cannot handle continuous data directly
K-OFSD	Filter	Neighborhood rough set	Design for high-dimensional and class-imbalanced data	Need to specify parameter in advance
OFS-A3M	Filter	Neighborhood rough set	Non-parametric	Unfit the dataset with unbalanced sample distribution
OFS-Density	Filter	Neighborhood rough set	Fuzzy equal constraint	Cannot handle discrete data well

$$\overline{B}X = \{[x_i]_B \mid [x_i]_B \cap X \neq \emptyset, x_i \in U\} \tag{3}$$

The upper approximation is the minimal union of elemental granules containing X, and the lower approximation is the maximal union of elemental granules consistently contained in X. The lower approximation is also called the positive region, denoted as POS_B.

Definition 4. For an arbitrary feature subset B and the decision feature (class attribute) D, the dependency degree of B to D is defined as the ratio of consistent objects:

$$\gamma_B(D) = \frac{CARD(POS_B(D))}{|U|} \tag{4}$$

where $CARD(POS_B(D))$ denotes the number of positive region objects.

Obviously, $0 \leq \gamma_B(D) \leq 1$.

Theorem 2 ([39]). *Suppose B is a subset of conditional features, f is an arbitrary conditional attribute that belongs to the dataset, and D is the decision attribute. Then $\gamma_{B\cup f}(D) \geq \gamma_B(D)$.*

Proof. The proof of this theorem is available in [39] on page 90.

For Rough Set based streaming feature selection, the primary goal is to maximize $\gamma_{S^t}(D)$ at each timestamp t.

Theorem 3. *Dependency degree function $\gamma_{S^t}(D)$ is an early terminated mapping function.*

Proof. According to Definition 4, $\gamma_{S^t}(D) \leq 1$, thus $\gamma_{S^t}(D)$ satisfies property 1° in Definition 3. In terms of Theorem 2, for feature f_t at timestamp t, $\gamma_{S^{t-1}\cup f_t}(D) \geq \gamma_{S^{t-1}}(D)$. Regardless of whether f_t is selected, $\gamma_{S^t}(D) \geq \gamma_{S^{t-1}}(D)$. So, $\gamma_{S^t}(D)$ satisfies property 2° in Definition 3. Thus, $\gamma_{S^t}(D)$ is an early terminated mapping function.

Definition 5. Suppose the selected feature subset is S and f' is a feature in S. The significance of f' to S can be defined as:

$$\sigma_D(f', S) = \gamma_S - \gamma_{S-\{f'\}}. \tag{5}$$

For the selected feature subset S^t at timestamp t, if there exist a feature $f'(f' \in S)$ making $\sigma_D(f', S) = 0$, then f' can be discarded as a redundant feature.

To illustrate the characteristics of Rough Set-based streaming feature selection methods, we present a baseline Rough Set-based streaming feature selection algorithm RS-OSFS, shown as Algorithm 1. We use the classical Rough Set and k-nearest Neighborhood Rough Set models for nominal data and continuous data.

Definition 6. For feature subset B and instance subset X, we define the lower and upper approximations in terms of the k-nearest neighborhood relation as

$$\underline{B_K}X = \{x_i \mid K(x_i) \subseteq X, x_i \in U\} \tag{6}$$

$$\overline{B_K}X = \{x_i \mid K(x_i) \cap X \neq \emptyset, x_i \in U\} \tag{7}$$

where $K(x_i)$ denotes the k-nearest neighbors around x_i.

We apply RS-OSFS on dataset Nova. The growth trend of dependency degree and the number of selected features varying with difference timestamps are shown as Figs. 2 and 3 respectively.

Algorithm 1 Rough Set Based Online Streaming Feature Selection Baseline Method (RS-OSFS)

Input: k (the number of neighbors for k-nearest Neighborhood Rough Set model) / none (for classical Rough Set model)
Output: the selected feature set S;

```
S: the selected feature set, initialized to {};
Repeat
Get a new feature f_t at timestamp t;
IF γ_{S∪f_t}(D) > γ_S(D)
    S = S ∪ f_t;
END IF
Until no more features are available;
return S;
```

From Fig. 3, we can observe that RS-OSFS only selects 16 new features among 2000 streaming features from timestamp 8000 to 10,000. Meanwhile, for the following 6969 features, RS-OSFS selects less than 20 new features among them. Meanwhile, in Fig. 2, the dependency degree of the selected feature subset increased slowly or even no longer increased after a certain number of features have arrived (or after a particular timestamp). For dataset Nova, after the arrival of 12,000 features, the dependency degree of the selected feature subset nearly reaches the maximum. Thus, we can terminate the streaming feature selection early before the end.

3.3. Early terminated streaming feature selection framework

The key issue for early terminated online streaming feature selection is how to choose the proper terminated timestamp t.

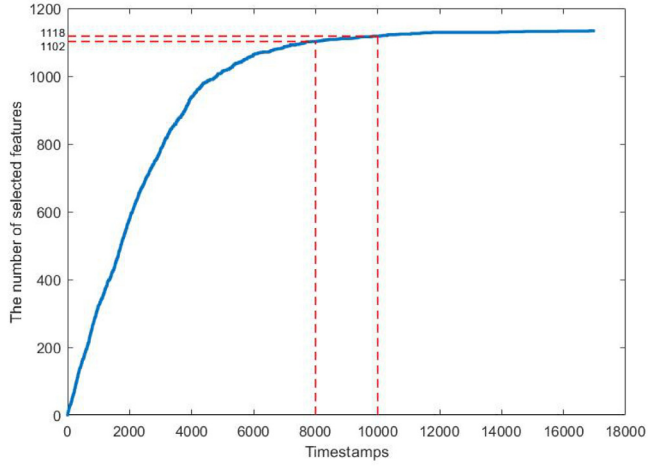

Fig. 3. The growth trend of the number of selected features on Nova varying with different timestamps. The growth rate is very slow after timestamp 8000.

The native idea for this is to specify a threshold $\alpha(\alpha < \lambda)$ and we terminate the selection when $\gamma(S^t) \geq \alpha$. However, this is difficult or even unrealistic for real-world datasets. There are two main reasons: (1) Without the information of the entire feature space before learning, it is difficult to specify the value of α. (2) Although the maximal value of $\gamma_S(D)$ is one theoretically, it cannot always be reached. For real-world datasets, an inappropriate parameter value of α may never be reached. Thus, we cannot just use a predefined threshold for termination.

For early terminated online streaming feature selection, at timestamp t, according to Definition 2, we need to solve two problems:

- (1) how can we define $\varphi(t, t+w)$.
- (2) how can we estimate the value of $\phi(t, t+w)$;

For problem (1), suppose we can require one feature per second. From timestamp t to $t+w$, we need to wait w seconds. The main reason we terminate the selection early is the expected increase of the mapping function for the next w features lower than the cost of time consumption. For example, we can afford the wait of 100 more seconds if the increase of the mapping function is bigger than 0.01. Then, $w = 100$, and $\varphi(t, t+100) = 0.01$. We want to terminate the selection early before the end while without much loss in the increase of mapping function. Therefore, it is necessary to consider the growing trend in a large interval for future arrival features to avoid local stagnation. Thus, $\varphi(t, t+w)$ consists of two factors: (1) the number of features w we can afford to wait; (2) the least expected increases of mapping function from these w features

For high-dimensional datasets, there are a large number of irrelevant and redundant features. For a new arriving feature f_t at timestamp t, if f_t can increase the dependency degree of S, we do not need to consider terminating the selection. However, if f_t is discarded and the last arrived $w-1$ features bring a small increase or even no increase to the dependency degree of the selected feature subset, then we can check whether the selection should be terminated. Therefore, for the problem (2), at timestamp t, we use the increase of the last arrived w features to estimate the dependency degree at timestamp $t+w$. Specifically, according to Fig. 2, we observe that the growth trend of dependency degree is gradually slowing down. Thus, if the increase of the last arrived w features $(\gamma(S^t) - \gamma(S^{t-w}))$ is smaller than $\varphi(t, t+w)$, $\phi(t, t+w)$ has a high probability of being smaller than $\varphi(t, t+w)$. Meanwhile, to avoid local stagnation, we should specify a larger value of w in practice.

Based on this, we propose the Early Terminated Online Streaming Feature Selection framework, named OSFS-ET, as shown in Algorithm 2.

Algorithm 2 Early Terminated Online Streaming Feature Selection Framework (OSFS-ET)

Input: w: the number of features expected to wait;
β: the dependency degree expected to increase;
Output: the selected feature set S;

1: S: initialized to {};
2: $depArray = []$, $depSet = 0$;
3: **Repeat**
4: Get a new feature f_t at timestamp t;
5: $depNew = h(S \cup \{f_t\})$;
6: $depArray_{[t]} = depNew$
7: IF $depNew > depSet$
8: $S = S \cup \{f_t\}$;
9: $depSet = depNew$;
10: END IF
11: IF $depNew - depArray_{[t-w]} < \beta$
12: remove redundant features in S;
13: **terminate** the selection and **goto** step 16;
14: END IF
15: **Until** no more features are available;
16: **return** S;

At timestamp t, if the new arriving feature f_t can increase the dependency degree of S, it will be selected into S in Step 8. With the predefined parameter values of w and β, if the increase of the last arrived w features is smaller than β, OSFS-ET removes the redundant features in Step 12 and terminates the selection in Step 13.

3.4. The time complexity of SFS-ET

For different feature mapping functions, the time complexity is different. Suppose the time complexity of the mapping function $h()$ is $O(h)$. From step 5 to step 10, the time complexity is $O(h)$. In step 12, we remove redundant features in terms of Eq. (5) and the time complexity is $O(|S| * h)$. This paper uses the dependency degree of Rough Set theory as the target mapping function, and the time complexity of $O(h)$ is $O(n^2)$. Therefore, the time complexity of OSFS-ET is $O(|S| * n^2)$. OSFS-ET gets the worst time complexity when the online streaming feature selection algorithm selects all the features, $|S| = m$. Thus, the worst time complexity is $O(m*n^2)$.

However, for real-world applications, it is unrealistic to select all the features. Meanwhile, according to the experiment results in Section 4, OSFS-ET always terminates the selection very early before the end (about one-third of the entire feature streams). Thus, OSFS-ET can run very fast with early termination.

4. Experiments

4.1. Experimental settings

4.1.1. Datasets

In this section, we apply the proposed OSFS-ET and its competing algorithms on twelve real-world high-dimensional datasets [40,41],[1] as shown in Table 2.

4.1.2. Evaluation metrics

We use two basic classifiers, KNN(k = 9) and SVM (with the linear kernel) in Matlab R2017a, to evaluate a selected feature subset in our experiments. We perform 5-fold cross-validation on each dataset. Feature selection is training on 4/5 samples and

[1] Public available at http://www.cs.binghamton.edu/~lyu/KDD08/data/, http://csse.szu.edu.cn/staff/zhuzx/Datasets.html, http://archive.ics.uci.edu/ml/index.php.

Table 2
Real-world datasets.

Index	Datasets	Instances	Features	Classes	Type
1	Srbct	63	2308	4	Real
2	Lymphoma	62	4026	3	Real
3	Prostate	102	6033	2	Real
4	Dlbcl	77	7129	2	Integer
5	Leukemia	72	7129	2	Real
6	Arcene	200	10,000	2	Integer
7	LungCancer	181	12,533	2	Real
8	Ovarian	253	15,154	2	Real
9	Breast	97	24,481	2	Real
10	Madelon	2000	500	2	Integer
11	Gina	3153	970	2	Integer
12	Gisette	6000	5000	2	Integer

testing on the rest 1/5. All competing algorithms use the same training and testing data for each fold. The order of streaming features is random, and we run ten times for each dataset.

To validate whether there is a significant difference in prediction accuracy and running time, we conduct the Friedman test at a 95% significance level under the null-hypothesis and the Nemenyi test as a post-hoc test [42].

4.1.3. Comparing algorithms

We compare OSFS-ET with seven state-of-art streaming feature selection methods, including: Alpha-Investing [17], Fast-OSFS [14], GFSSF [29], SAOLA [28], OSFSMI [30], OFS-A3M [20], and OFS-Density [21]. All algorithms mentioned above are implemented in MATLAB[2] [43], and GFSSF is implemented by ourselves. For Alpha-investing, the parameters are set to the values used in [17]. The significance level is set to 0.01 for Fast-OSFS, SAOLA, and OSFSMI. We apply the k-nearest neighborhood rough set model in OSFS-ET for nominal datasets to calculate the dependency degree, and the value of k is set to 9 as the KNN classifier.

4.1.4. Streaming simulation

To simulate a natural feature stream environment, we specify each dataset in Table 2 to generate features one by one over time. The generation speed is set as one feature per 0.1 s in our experiments. In other words, we send one random feature to the algorithm every 0.1 s until there are no features in the target dataset. In real-world applications, we may need much more than 0.1 s to require a new streaming feature.

4.1.5. Computational device

All experimental results are conducted on a PC with AMD(R) 3700X, 3.6 GHz CPU, and 32 GB memory.

4.2. Analysis of parameters

In OSFS-ET, the parameters w and β are related. For example, we can expect the increase of the next 100 features reaching 0.01 or the next 200 features reaching 0.02. For the convenience of parameters analysis, we fix the value of β as 0.01 and test different values of $w = 100, 200, 400, 600, 800, 1000$. The predictive accuracy on KNN and SVM varying with different values of w on datasets Srbct, Prostate, Dlbcl, Arcene, Breast, and Gisette can be seen as Fig. 4. Besides, Fig. 5 shows the dependency degree of the final selected feature subset and the terminated ratio of the whole streaming feature space varying with different values of w on these six datasets.

[2] Public available at https://github.com/kuiy/LOFS and https://github.com/doodzhou/OSFS.

Table 3
The average ranks varying with different values of w.

	$w = 100$	$w = 200$	$w = 400$	$w = 600$	$w = 800$	$w = 1000$
KNN	5.6667	4.7500	3.3750	3.0417	2.2083	**1.9583**
SVM	5.6250	4.6250	3.4167	2.9167	2.2083	**2.2083**
Dependency degree	6.0000	5.0000	3.9167	2.8333	1.8750	**1.3750**
Terminated ratio	**1.0000**	2.0000	3.1250	4.1667	4.9167	5.7917

The p-values of the Friedman test on KNN, SVM, dependency degree, and terminated ratio are 5.5038e−10, 2.6727e−08, and 3.2153e−31, 6.1388e−32 respectively. Thus, there is a significant difference among these six cases on the predictive accuracy, dependency degree, and terminated ratio of final selected features. According to the Nemenyi test, the value of CD (critical difference) is 2.1764. We list the average ranks in Table 3.

From Figs. 4, 5, and Table 3, we have the following observations:

- On predictive accuracy, according to the statistical test, $w = 1000$ gets the best performance in both cases of KNN and SVM. There is a significant difference between $w = 100, 200$ and $w = 1000$ on predictive accuracy. Meanwhile, there is no significant difference between $w = 400, 600, 800$ and $w = 1000$. A too-small $w(w = 100, 200)$ will terminate the selection too early and cannot guarantee satisfactory performance. In general, with the increase of w, OSFS-ET checks and selects more streaming features, and the final dependency degree can be bigger at the same time. However, a bigger value of dependency degree does not necessarily mean better predictive accuracy. If there are many irrelevant and redundant features in the datasets, such as Dlbcl, Breast, and Arcene, more selected features may decrease the performance. Thus, the parameter w should big enough and can get the best performance with different values for different datasets.
- On the dependency degree of the final selected features, $w = 1000$ gets the maximal dependency degree among these six different parameter values. In total, with the increase of w, the dependency degree of final selected features will increase too. However, there is no significant difference between $w = 600, 800$ and $w = 1000$. When the value of w is big enough, the increase in the value of the dependency degree is very limited. This demonstrates that our proposed method can terminate the selection early with only a small loss in the final dependency degree when w is big enough.
- On the terminated ratio of the whole streaming feature space, the values grow almost linearly from $w = 100$ to 1000. The terminated ratios grow fast when the dimension of target datasets is very low, such as dataset Srbct. For high-dimensional datasets, the maximal terminated ratio is less than 0.3 even $w = 1000$. This indicates that OSFS-ET can save much waiting time for high-dimensional datasets while maintaining a competing performance on dependency degree and predictive accuracy.

In sum, with the increase of w, the dependency degree of the final selected feature subset and the terminated ratio of the whole streaming feature space grows linearly. However, the final dependency degree and predictive accuracy are not affected too much when the value of w is big enough. Thus, with a proper value of w, we can terminate the streaming feature selection early and get a competing performance. In the following experiments, we set $w = 800$ as an experience value. For dataset Madelon and Gina, the total numbers of features are 500 and 970, respectively. Thus, we set $w = 200$ for these two datasets.

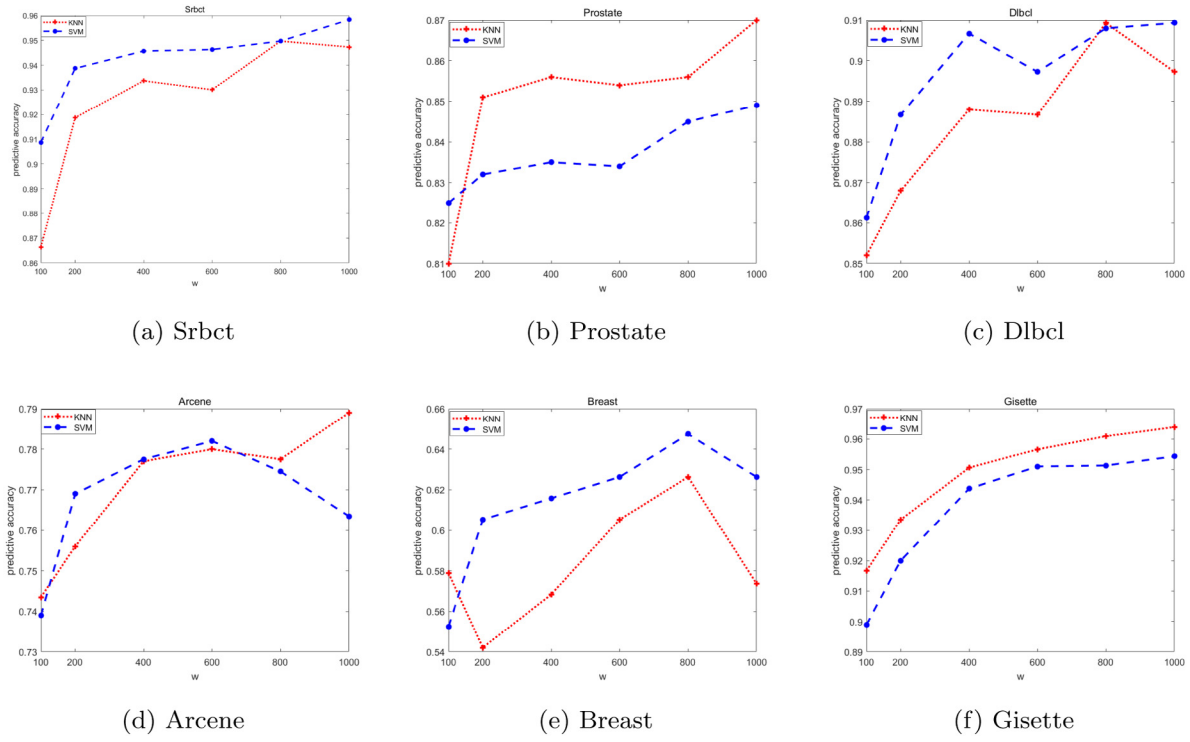

Fig. 4. Predictive accuracy on KNN and SVM varying with different values of w.

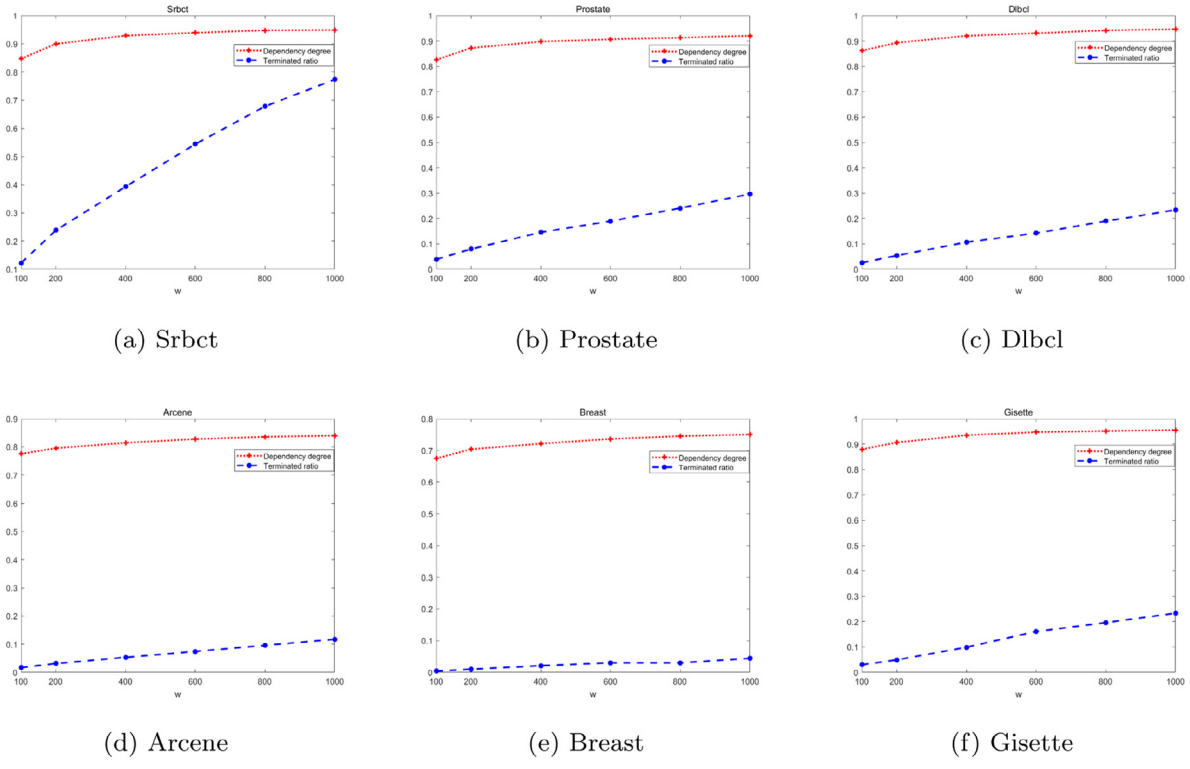

Fig. 5. The dependency degree of final selected feature subset and terminated ratio varying with different values of w.

4.3. OSFS-ET vs. State-of-art streaming feature selection methods

In this section, we compare OSFS-ET with some state-of-art streaming feature selection methods, including: Alpha-Investing [17], Fast-OSFS [14], GFSSF [29], SAOLA [28], OSFSMI [30], OFS-A3M [20], and OFS-Density [21].

Tables 4 and 5 summarize the predictive accuracy of these competing algorithms using the KNN and SVM classifiers. Tables 6 and 7 show the running time and the mean number of selected features for these competing algorithms. Table 8 presents the terminated ratio of OSFS-ET on these datasets. The p-values of the Friedman test on KNN, SVM, running time, and the mean number

Table 4
OSFS-ET vs. competing algorithms with KNN.

Dataset	OSFS-ET	Alpha-Investing	Fast-OSFS	GFSSF	SAOLA	OSFSMI	OFS-A3M	OFS-Density
Srbct	0.9301	0.3843	0.6971	0.4378	0.708	0.7449	**0.9442**	0.8492
Lymphoma	**0.9633**	0.8567	0.925	0.745	0.9483	0.915	0.9233	0.9567
Prostate	0.87	0.702	0.882	0.617	0.883	0.913	0.879	**0.924**
Dlbcl	0.8707	0.8027	0.8507	0.748	**0.904**	0.84	0.872	0.872
Leukemia	0.9029	0.76	0.9171	0.6329	0.9386	0.9043	0.9057	**0.94**
Arcene	**0.7845**	0.6925	0.7	0.6425	0.646	0.683	0.772	0.7705
LungCancer	0.9739	0.8561	0.975	0.835	**0.9833**	0.9667	0.97	0.9789
Ovarian	0.9881	0.9585	**0.9984**	0.7466	0.9913	0.9766	0.9917	0.992
Breast	0.5947	0.5737	0.6632	0.5421	0.6421	**0.6684**	0.6263	0.5421
Madelon	**0.8794**	0.6896	0.5883	0.5194	0.5633	0.6267	0.5019	0.5019
Gina	**0.943**	0.9236	0.8669	0.6622	0.8268	0.7695	0.8443	0.829
Gisette	**0.9609**	0.9339	0.8626	0.7476	0.9133	0.9066	0.9507	0.9257
AVG.	**0.8818**	0.7454	0.8239	0.6480	0.8213	0.8189	0.8391	0.8323
AVG. RANKS	**3.0833**	5.6667	3.8333	7.7083	3.7500	4.8333	3.7500	3.3750

Table 5
OSFS-ET vs. competing algorithms with SVM.

Dataset	OSFS-ET	Alpha-Investing	Fast-OSFS	GFSSF	SAOLA	OSFSMI	OFS-A3M	OFS-Density
Srbct	**0.9524**	0.3655	0.7526	0.4085	0.7941	0.7848	0.9414	0.8655
Lymphoma	**0.955**	0.8533	0.915	0.7533	0.9317	0.8867	0.9417	0.9467
Prostate	0.855	0.722	0.896	0.642	0.869	0.912	0.868	**0.922**
Dlbcl	0.8787	0.8027	0.8573	0.7533	**0.9067**	0.8547	0.876	0.8827
Leukemia	0.9171	0.7771	0.9257	0.6629	**0.9557**	0.9271	0.9086	0.9371
Arcene	**0.7965**	0.6865	0.707	0.627	0.656	0.6915	0.7735	0.7405
LungCancer	0.9761	0.8611	0.9806	0.8311	**0.9883**	0.98	0.9778	0.9794
Ovarian	0.9937	0.9988	**1**	0.7717	0.9917	0.977	0.9924	0.992
Breast	0.6053	0.5632	0.6368	0.4632	0.6368	**0.6842**	0.5842	0.5789
Madelon	**0.6148**	0.6121	0.6113	0.5246	0.6033	0.6146	0.4912	0.4921
Gina	**0.8697**	0.8685	0.8434	0.673	0.8211	0.7733	0.8253	0.8123
Gisette	0.9513	**0.971**	0.8586	0.7559	0.9101	0.8691	0.9233	0.8806
AVG.	**0.8638**	0.7568	0.8320	0.6555	0.8387	0.8295	0.8419	0.8358
AVG. ranks	**2.8333**	5.3333	3.8750	7.7500	3.7083	4.4167	4.2500	3.8333

Table 6
SFS-ET vs. competing algorithms on running time (seconds).

Dataset	OSFS-ET	Alpha-Investing	Fast-OSFS	GFSSF	SAOLA	OSFSMI	OFS-A3M	OFS-Density
Srbct	**164.37552**	230.8381	230.9804	231.0358	231.0361	231.1129	231.4123	231.3791
Lymphoma	**91.91844**	402.7656	403.0678	403.3025	403.4002	404.3673	403.5953	403.7921
Prostate	**165.9691**	603.6197	603.7551	605.1014	603.8984	604.5919	605.8381	605.7123
Dlbcl	**119.87406**	713.5553	713.4179	714.6096	713.5833	714.0763	715.1124	715.0268
Leukemia	**147.50412**	713.3263	713.4408	714.5696	713.7111	714.0995	714.9921	714.8043
Arcene	**102.4125**	1001.4564	1000.9277	1005.8441	1001.4744	1003.2078	1011.2401	1012.2934
LungCancer	**122.3184**	1254.5654	1255.0218	1258.5482	1256.21	1258.0364	1266.4384	1265.3466
Ovarian	**145.01134**	1521.5953	1517.7559	1524.1189	1517.6255	1523.009	1542.3416	1541.0659
Breast	**86.1621**	2451.8381	2449.5551	2460.8634	2449.8293	2450.1306	2458.5503	2459.6492
Madelon	181.4094	50.0854	50.0636	51.521	**50.0402**	50.0429	219.1359	218.3765
Gina	417.4275	102.301	119.8029	106.5134	**97.2163**	97.392	473.4378	563.4573
Gisette	2301.0541	1122.1065	549.6397	947.8269	**502.1072**	502.8906	10 346.1451	11 167.5612
AVG.	**337.1**	869.1	800.3	836.2	795.0	796.0	1665.6	1741.4
AVG. ranks	**2.2500**	3.0833	2.9167	5.4167	3.1667	4.5000	7.4167	7.2500

of selected features are 2.4793e−06, 1.0320e−05, 1.4469e−13, and 2.7008e−16, respectively. Thus, these algorithms have a significant difference in predictive accuracy, running time, and the mean number of selected features. According to the Nemenyi test, the value of CD (critical difference) is 3.0335.

From Tables 4 to 8, we can observe that:

- On predictive accuracy, OSFS-ET gets the highest average accuracies and lowest average ranks in both cases of KNN and SVM. Thus, according to the statistical test, OSFS-ET performs best among these eight competing algorithms. Meanwhile, OSFS-ET gets five of the ten best performances with KNN and SVM, respectively. From Table 8, we can see that the final dependency degree of OSFS-ET can achieve a very high value even it terminates the selection early. For example, on datasets Lymphoma, LungCancer, Ovarian, and Gisette, the dependency degree of selected features nearly achieve the maximal value 1 while the terminated ratio just around 0.2 or even less than 0.1. Thus, although our new method terminates the streaming feature selection very early before the end (0.3 on average), it still can get outstanding performance on predictive accuracy compared with other streaming feature selection methods.
- On the running time, OSFS-ET is the fastest while OFS-A3M and OFS-Density are the slowest. For Rough Set models, the time complexity of dependency degree calculation is $O(n^2)$, where n is the number of instances in datasets. For high-dimensional small sample datasets, OSFS-ET terminates the selection very early and can save much waiting time compared with other competing algorithms. However, for low-dimensional large sample datasets, such as Madelon and Gina, OSFS-ET spends much more time than those not Rough Set-based algorithms. Thus, with the early termination during streaming feature selection, the higher the

Table 7
OSFS-ET vs. competing algorithms on the mean number of selected features.

Dataset	OSFS-ET	Alpha-Investing	Fast-OSFS	GFSSF	SAOLA	OSFSMI	OFS-A3M	OFS-Density
Srbct	35.46	1.6	4.48	3.04	18.52	7.34	7.58	5.84
Lymphoma	18.18	3.6	5.5	2.86	39.26	7.78	3.66	10.52
Prostate	35.38	2.18	3.34	3.7	12.94	7.54	20.02	5
Dlbcl	30.48	5.92	4.38	2.94	14.14	7.16	20.36	7.12
Leukemia	35.74	2.54	4.6	3	22.44	8.34	11.7	4.48
Arcene	54.12	5.72	5.24	5.32	17.66	8.64	36.24	21.74
LungCancer	36.14	2.42	6.96	3.64	48.98	9.7	5.9	9.1
Ovarian	23.34	53.76	5.48	4.22	8.36	10.46	3.3	8.6
Breast	24.9	4.7	4.1	3.5	18.7	7.1	27.3	12.6
Madelon	16.8	5.2	5.4	6.4	7.4	4	2	2
Gina	118.8	88	20.9	12.4	12.3	7.4	24.1	10.4
Gisette	103.4	750.2	2.2	29.4	23	16.8	63.4	14.8
AVG.	44.3	77.1	6.0	6.7	20.3	8.5	18.7	9.3
AVG. ranks	7.5833	3.5000	3.0000	2.5833	5.9167	4.3333	5.1250	3.9583

Table 8
The terminated ratio and final dependency degree of OSFS-ET.

Dataset	Final dependency degree	Terminated ratio
Srbct	0.8583	0.7094
Lymphoma	0.9783	0.2274
Prostate	0.886	0.273
Dlbcl	0.9347	0.1674
Leukemia	0.8957	0.2058
Arcene	0.79	0.1008
LungCancer	0.9933	0.096
Ovarian	1	0.0926
Breast	0.6474	0.035
Madelon	0.5506	0.95
Gina	0.8209	0.665
Gisette	0.9519	0.184
AVG.	0.8589	0.3088

dimensionality of the datasets, the more pronounced the time savings for our new method. Meanwhile, compared with other Rough Set-based online streaming feature selection methods, OSFS-ET can significantly reduce the running time and make up for the shortcomings caused by the high time complexity.

- On the mean number of selected features, according to the average ranks, OSFS-ET selects the most number of features, while GFSSF selects the least. However, GFSSF gets the worst performance at the same time. To terminate the streaming feature selection early, OSFS-ET selects features as long as it can increase the dependency degree of the selected feature subset and remove redundant features once after the termination. Thus, OSFS-ET selects more features than other competing algorithms. However, the number of selected features for OSFS-ET is still tiny to the dimensions of the entire feature space. Meanwhile, OSFS-ET gets the best performance on predictive accuracy with these selected features.

In general, our new early terminated method can be far faster than the competing state-of-art streaming feature selection algorithms for high-dimensional datasets while maintains an outstanding performance on predictive accuracy.

5. Conclusion

In this paper, we study the exciting issue of how to terminate the online streaming feature selection early while maintaining a satisfactory performance for the first time. An assumption is proposed that the online streaming feature selection can be terminated early if the expected increase of mapping function is much lower than the time consumption cost for the following arriving features. Based on this, we first present a formal definition on this issue and summarize two properties that the early terminated mapping function should satisfy. We choose the dependency degree function in Rough Set theory as our early terminated mapping function and propose a new framework that can terminate the online streaming feature selection early while maintaining a competing performance using the currently selected features. Extensive experiments on twelve real-world datasets demonstrate that our new framework can significantly reduce waiting time while maintaining an outstanding performance on predictive accuracy.

Our new framework needs to specify two parameters (the number of features expected to wait and the dependency degree expected to increase) before learning. However, the framework cannot terminate the selection early if the parameter values we set are too big. On the contrary, if parameter values are too small, the framework may terminate the selection too early to be good enough. Besides, we choose the dependency degree function in Rough Set theory as the early terminated mapping function in this paper. Nevertheless, we should notice that our new framework can be applied with other mapping functions that satisfy the two properties mentioned in the problem definition. As we know, Rough Set models have high time complexity. Thus, in terms of these two properties, more new efficient mapping functions will be considered in our future work.

CRediT authorship contribution statement

Peng Zhou: Conceptualization, Methodology, Software, Writing – original draft, Funding acquisition. **Peipei Li:** Validation, Investigation, Writing – review & editing, Funding acquisition. **Shu Zhao:** Formal analysis, Project administration, Funding acquisition. **Yanping Zhang:** Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under grants (61906056, 61976077, 61876001).

References

[1] H. Liu, H. Motoda, Computational Methods of Feature Selection, Chapman and Hall/CRC Press, 2007.
[2] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[3] G. Wei, J. Zhao, Y. Feng, A. He, J. Yu, A novel hybrid feature selection method based on dynamic feature importance, Appl. Soft Comput. 93 (2020) 106337.
[4] J. Wu, M. Song, W. Min, J. Lai, W. Zheng, Joint adaptive manifold and embedding learning for unsupervised feature selection, Pattern Recognit. 112 (2021) 107742, http://dx.doi.org/10.1016/j.patcog.2020.107742.
[5] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, Acm Comput. Surv. 50 (6) (2017) 94:1–94:45.
[6] Y. Li, T. Li, H. Liu, Recent advances in feature selection and its applications, Knowl. Inf. Syst. 53 (3) (2017) 551–577.
[7] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.
[8] Z. Yuan, H. Chen, P. Xie, P. Zhang, J. Liu, T. Li, Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions, Appl. Soft Comput. 107 (2021) 107353, http://dx.doi.org/10.1016/j.asoc.2021.107353.

[9] W. Ding, T.F. Stepinski, Y. Mu, L. Bandeira, R. Ricardo, Y. Wu, Z. Lu, T. Cao, X. Wu, Subkilometer crater discovery with boosting and transfer learning, ACM Trans. Intell. Syst. Technol. (TIST) 2 (4) (2011) 1–22.
[10] M. Wang, H. Li, D. Tao, K. Lu, X. Wu, Multimodal graph-based reranking for web image search, IEEE Trans. Image Process. 21 (11) (2012) 4649–4661.
[11] X. Wu, X. Zhu, G. Wu, W. Ding, Data mining with big data, IEEE Trans. Knowl. Data Eng. 26 (1) (2014) 97–107.
[12] G. Ditzler, J. LaBarck, J. Ritchie, G. Rosen, R. Polikar, Extensions to online feature selection using bagging and boosting, IEEE Trans. Neural Netw. Learn. Syst. 29 (9) (2017) 4504–4509.
[13] B. Sang, H. Chen, T. Li, W. Xu, H. Yu, Incremental approaches for heterogeneous feature selection in dynamic ordered data, Inform. Sci. 541 (2020) 475–501.
[14] X. Wu, K. Yu, W. Ding, H. Wang, X. Zhu, Online feature selection with streaming features, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1178–1192.
[15] X. Hu, P. Zhou, P. Li, J. Wang, X. Wu, A survey on online feature selection with streaming features, Front. Comput. Sci. 12 (3) (2018) 479–493.
[16] S. Perkins, J. Theiler, Online feature selection using grafting, in: Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 592–599.
[17] J. Zhou, D.P. Foster, R.A. Stine, L.H. Ungar, Streamwise feature selection, J. Mach. Learn. Res. 3 (2) (2006) 1532–4435.
[18] P. Zhou, P. Li, S. Zhao, X. Wu, Feature interaction for streaming feature selection, IEEE Trans. Neural Netw. Learn. Syst. 32 (10) (2021) 4691–4702.
[19] S. Eskandari, M. Javidi, Online streaming feature selection using rough sets, Internat. J. Approx. Reason. 69 (C) (2016) 35–57.
[20] P. Zhou, X. Hu, P. Li, X. Wu, Online streaming feature selection using adapted neighborhood rough set, Inform. Sci. 481 (2019) 258–279.
[21] P. Zhou, X. Hu, P. Li, X. Wu, Ofs-density: A novel online streaming feature selection method, Pattern Recognit. 86 (2019) 48–61.
[22] J. Wang, P. Zhao, S.C. Hoi, R. Jing, Online feature selection and its applications, IEEE Trans. Knowl. Data Eng. 26 (3) (2013) 698–710.
[23] P. Zhou, X. Hu, P. Li, X. Wu, Online feature selection for high-dimensional class-imbalanced data, Knowl.-Based Syst. 136 (2017) 187–199.
[24] G. Manikandan, S. Abirami, An efficient feature selection framework based on information theory for high dimensional data, Appl. Soft Comput. 111 (2021) 107729, http://dx.doi.org/10.1016/j.asoc.2021.107729.
[25] Q. Zhang, Q. Xie, G. Wang, A survey on rough set theory and its applications, CAAI Trans. Intell. Technol. 1 (4) (2016) 323–333.
[26] S. Fong, R. Wong, A.V. Vasilakos, Accelerated PSO swarm search feature selection for data stream mining big data, IEEE Trans. Serv. Comput. 9 (1) (2015) 33–45.
[27] R. Ma, Y. Wang, L. Cheng, Feature selection on data stream via multi-cluster structure preservation, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1065–1074.
[28] K. Yu, X. Wu, W. Ding, J. Pei, Scalable and accurate online feature selection for big data, ACM Trans. Knowl. Discov. Data 11 (2) (2016) 1–39.
[29] H. Li, X. Wu, Z. Li, W. Ding, Group feature selection with streaming features, in: IEEE 13th International Conference on Data Mining, 2013, pp. 1109–1114.
[30] M. Rahmaninia, P. Moradi, Osfsmi: Online stream feature selection method based on mutual information, Appl. Soft Comput. 68 (2018) 733–746.
[31] D. Wu, Y. He, X. Luo, M. Shang, X. Wu, Online feature selection with capricious streaming features: A general framework, in: 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 683–688.
[32] D. Paul, R. Kumar, S. Saha, J. Mathew, Multi-objective cuckoo search-based streaming feature selection for multi-label dataset, ACM Trans. Knowl. Discov. Data (TKDD) 15 (6) (2021) 1–24.
[33] D. Wu, Y. He, X. Luo, M. Zhou, A latent factor analysis-based approach to online sparse streaming feature selection, IEEE Trans. Syst. Man Cybern.: Syst. (2021) 1–15, http://dx.doi.org/10.1109/TSMC.2021.3096065.
[34] D. You, R. Li, S. Liang, M. Sun, X. Ou, F. Yuan, L. Shen, X. Wu, Online causal feature selection for streaming features, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1–15, http://dx.doi.org/10.1109/TNNLS.2021.3105585.
[35] S. An, Q. Hu, C. Wang, Probability granular distance-based fuzzy rough set model, Appl. Soft Comput. 102 (2021) 107064, http://dx.doi.org/10.1016/j.asoc.2020.107064.
[36] R. Diao, N.M. Parthaláin, Q. Shen, Dynamic feature selection with fuzzy-rough sets, in: 2013 IEEE International Conference on Fuzzy Systems, 2013, pp. 1–7.
[37] H. Wang, G. Wang, X. Zeng, S. Peng, Online streaming feature selection based on conditional information entropy, in: Proceedings of the 8th IEEE International Conference on Big Knowledge, 2017, pp. 230–235.
[38] Z. Pawlak, Rough Sets - Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht , Boston, 1991.
[39] R. Jensen, Q. Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, IEEE Press Series on Computational Intelligence, JOHN WILEY & SONS, 2008.
[40] L. Yu, C. Ding, S. Loscalzo, Stable feature selection via dense feature groups, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 803–811.
[41] K. Yang, Z. Cai, J. Li, G. Lin, A stable gene selection in microarray data analysis, BMC Bioinformatics 7 (2006) 228.
[42] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (1) (2006) 1–30.
[43] K. Yu, W. Ding, X. Wu, Lofs: Library of online streaming feature selection, Knowl.-Based Syst. 113 (2016) 1–3.