OSFSMI: Online stream feature selection method based on mutual information

Maryam Rahmaninia [a], Parham Moradi [b],*

[a] Faculty of Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran
[b] Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

ABSTRACT

Feature selection is used to choose a subset of the most informative features in pattern identification based on machine learning methods. However, in many real-world applications such as online social networks, it is either impossible to acquire the entire feature set or to wait for the complete set of features before starting the feature selection process. To handle this issue, online streaming feature selection approaches have been recently proposed to provide a complementary algorithmic methodology by choosing the most informative features. Most of these methods suffer from challenges such as high computational cost, stability of the generated results and the size of the final features subset. In this paper, two novel feature selection methods called OSFSMI and OSFSMI-k are proposed to select the most informative features from online streaming features. The proposed methods employ mutual information concept in a streaming manner to evaluate correlation between features and also to assess the relevancy and redundancy of features in complex classification tasks. The proposed methods do not use any learning model in their search process, and thus can be classified as filter-based methods Several experiments are performed to compare the performance of the proposed algorithms with the state-of-the-art online streaming feature selection methods The reported results show that the proposed methods performs better than the others in most of the cases.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Rapid improvement of storage and information processing technologies has led to appearance of large-scale datasets with large number of patterns and features [1]. The presence of high dimensional data – known as the curse of the dimensionality problem – reduces the performance of many machine learning methods [2]. A popular approach to tackle this problem is to reduce dimensionality of the feature space [3]. Feature selection is a well-known and effective dimensionality reduction approach that aims at selecting a parsimonious feature subset by identifying and eliminating those of redundant and irrelevant features.

Up to now, many feature selection methods have been proposed to improve the interpretability, efficiency and accuracy of the learning models. Most of these methods require to access the entire feature set to perform their search process [4–11]. However, in many real-world applications it is either impossible to acquire the entire data or it is impractical to wait for the complete data before feature selection starts [12–15]. In other words,

in these types of applications, data arrives sequentially and novel features or instances may appear incrementally. For example, in online social networks such as Twitter, in the case of presenting a new hot topic, a set of new keywords appears which leads to increase the dimensionality of the data over time. Traditional feature selection methods need to load the entire training dataset in the memory, which leads to exceeding the memory capacity for many real-world applications. These limitations make the traditional batch feature selection techniques impractical for emerging big data applications. To overcome these problems, online streaming feature selection methods (OSF) have been recently proposed to provide a complementary algorithmic methodology to addresses high dimensionality in big data analytics by choosing the most informative features [15–18].

Considering the fact that the whole data is unavailable, a successful OSF method needs an efficient incremental update rule in its search process. To this end, several methods have been recently proposed to select a best feature subset from online data streams. These methods can be classified into two categories: instance-based and feature-based OSF methods. In instance-based OSF methods, the number of instances increases over the time, while the number of features is assumed to be fixed [16,19–21]. This type of methods can be employed in some applications such

* Corresponding author.
  E-mail addresses: ma.rahmaninia@gmail.com (M. Rahmaninia),
p.moradi@uok.ac.ir (P. Moradi).

as traffic network monitoring, financial analysis of stock data streams and Internet query monitoring, where all feature space is available from the beginning but the number of instances increase over time. For example, the method proposed in [22] uses an incremental learning algorithm to select prominent features as new instances arrive. Therefore, the scope of these methods is limited to the problems where all features are given before the learning process. On the other hand, feature-based OSF methods assume that the feature space is unavailable or is infinite before starting the feature selection process [17,23–27]. In some real-world applications, the features are often expensive to generate (e.g., a lab experiment), and thus may appear in streaming manner. Generally, in feature-based OSF methods a criterion is defined to decide whether or not a newly arrived feature should be added to the model. For example, in [23] a statistical analysis is performed to evaluate the importance of a so-far-seen feature. This method requires prior knowledge about the entire feature space. Also, it does not have any strategy to remove those features that later found to be redundant. In [26] a conditional independent test is used to evaluate both relevancy and redundancy of a newly arrived feature. Although this method considers both relevancy and redundancy analysis in its online process, in each iteration it compares a new feature with all subsets of previously selected features, and thus it requires a high computational cost to process a so-far-seen feature. Recently, in [24] the authors proposed a method called SOALA that uses a pairwise mutual information to evaluate the relevance and redundancy of features in an online manner. In this method, deciding to include or ignore the features depends on some predefined parameters that setting their precise value requires to know the whole feature space. Also, the authors of [17] proposed an online feature selection method based on rough set theory to assess the importance of features in an online manner. Although their method does not require any domain knowledge, it only operates effectively with datasets containing discrete values, and therefore it is necessary to perform a discretization step for real-valued attributes. This method also suffers from high computational complexity when dealing with large-scale datasets.

In this paper, two novel feature-based OSF methods are proposed which aims to achieve high classification accuracy with a reasonable running time. We suppose that the features appear incrementally over time while the number of instances is considered to be fixed. The first method selects a prominent subset of feature with the minimal size, while the other method chooses a set of prominent features in constant size k. These methods employ mutual information concept to evaluate the relevancy and redundancy of features without using any learning model. Thus, the proposed methods can be classified into filter-based feature selection methods. A number of mutual information based methods have been successfully applied to feature selection tasks, see a review in [4]. However, most of these algorithms consider batch feature selection problem and cannot be applied for online feature selection scenarios. For example, in [28] mutual information is used to choose relevant features and eliminate redundant ones in an offline manner, and one needs to know the whole feature space at the beginning of the search process. The framework of the proposed methods consists of two steps. In the first step, the relevancy of each newly arrived feature with the target class is evaluated. The relevant features are included to the selected subset and the others are ignored. In the second step, the goal is to identify and eliminate those of ineffective features through several iterations. In this step, a specific strategy is used to eliminate redundant features. Using this strategy, if a previously selected feature is identified as a redundant feature, it is removed from the selected subset. To this end, a measure is introduced to evaluate the effectiveness of a newly arrived feature considering both of relevancy and redundancy concepts. Using this measure, those of ineffective features compared to

a so far seen feature are eliminated from the feature set and the process is continued iteratively until there is no more ineffective in the feature set. Using this process, one more chance is given to a new arrived feature to be processed in the further steps. Also, by removing a feature, it is ensured that there are exists some other effective features that has higher relevancy and lower redundancy value than it. This process also leads to decreasing the risk of discounting a feature. The proposed methods have several novelties compared to the previous feature-based OSF methods [17,23–27], as:

1. The proposed methods employ mutual information to analysis relevancy and redundancy of features. Compared to [29,30] which uses rough set theory in their processes, using mutual information has several advantages. While those based on rough sets require $O\left(n^2 m\right)$ time steps (n and m show the number of instances and the number of features, respectively), it is only $O(nm)$ for the proposed methods. Moreover, mutual information can be used for both discrete and continues features [31], while rough sets can only be applied on data with discrete variables.
2. The proposed methods do not employ any adjustable user-defined parameters. Thus, compared to [23,26], they can generate more robust results over various information sources.
3. The proposed methods employ an elimination strategy to remove redundant features in further steps, even if they have been previously selected. Compared to [23], this strategy results in returning a set of features with minimal redundancy.
4. The redundancy analysis step of the proposed methods is only performed on relevant features, while in [26,27], in each step the redundancy of so-far-seen feature is computed with all other previously selected features that needs a high computational cost. Also, [26,27] use a k-greedy search strategy to eliminate redundant features by checking all subsets of selected features. Thus, their complexity for evaluating the redundancy and relevancy of each feature is $O(|S_t|2^{|S_t|})$, where S_t denotes the selected feature at time t. Our algorithms take only $O\left(|S_t|^2\right)$ time steps to identify and eliminate redundant features.
5. Compared with the algorithm proposed in [24], the proposed methods calculate the redundancy of features considering the target class. [24] uses the pairwise mutual information to discover redundant features without considering the target class. Two features may be dependent on each other, while each sharing different information about the target class, and thus cannot not considered as redundant features.
6. Compared to [22], the proposed methods are filter-based feature-based OSF methods and do not use any learning model in their processes. In [22] the data samples arrive one-by-one and a learning method is used to evaluate features, and thus it is a wrapper and instance-based OSF method. Therefore, it is much slower than the proposed methods.
7. Although the method proposed in [28] uses the mutual information concept for relevancy and redundancy analysis, it is an offline feature selection method and needs to access the whole feature space and cannot be used online streaming feature selection.

The efficiency of the proposed methods is evaluated on two complex scenarios over 29 datasets in different categories. Our experiments show superiority of the proposed methods over others.

2. Related work and background

2.1. Related work

The aim of feature selection is to select a set of prominent features to improve interpretability and efficiency of the learning