Knowledge-Based Systems 136 (2017) 187–199

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Online feature selection for high-dimensional class-imbalanced data

Peng Zhou[a], Xuegang Hu[a,*], Peipei Li[a,*], Xindong Wu[b,*]

[a] Hefei University of Technology, Hefei 230009, China
[b] University of Louisiana, Lafayette, LA 70504, USA

ARTICLE INFO

Article history:
Received 9 February 2017
Revised 5 September 2017
Accepted 6 September 2017
Available online 8 September 2017

Keywords:
Online feature selection
High dimensional
Class imbalance
Neighborhood rough set

ABSTRACT

When tackling high dimensionality in data mining, online feature selection which deals with features flowing in one by one over time, presents more advantages than traditional feature selection methods. However, in real-world applications, such as fraud detection and medical diagnosis, the data is high-dimensional and highly class imbalanced, namely there are many more instances of some classes than others. In such cases of class imbalance, existing online feature selection algorithms usually ignore the small classes which can be important in these applications. It is hence a challenge to learn from high-dimensional and class imbalanced data in an online manner. Motivated by this, we first formalize the problem of online streaming feature selection for class imbalanced data, and then present an efficient online feature selection framework regarding the dependency between condition features and decision classes. Meanwhile, we propose a new algorithm of Online Feature Selection based on the Dependency in K nearest neighbors, called K-OFSD. In terms of Neighborhood Rough Set theory, K-OFSD uses the information of nearest neighbors to select relevant features which can get higher separability between the majority class and the minority class. Finally, experimental studies on seven high-dimensional and class imbalanced data sets show that our algorithm can achieve better performance than traditional feature selection methods with the same numbers of features and state-of-the-art online streaming feature selection algorithms in an online manner.

1. Introduction

Feature selection aims to select a subset of features, which can be used to derive a mapping function from samples to classes that is "as good as possible" according to some criterion [1]. There are many representative algorithms for traditional feature selection, such as ReliefF [2], Fisher Score [3], MI(Mutual Information)[4], mRMR (minimal Redundancy and Maximal Relevance) [5], Laplacian score [6,7], LASSO (least absolute shrinkage and selection operator) [8] and so on [9]. With the increasing of the scale of data, traditional batch feature selection can not meet the efficiency demand any more. For instance, the Web Spam Corpus 2011, a collection which has approximately 330,000 spam web pages and 16,000,000 features (attributes) [10]. Besides, all aforementioned approaches assume that all candidate features are available before learning takes place. However, in many real-world applications, features are generated dynamically, and arrive one by one over time. For example, in image analysis [11], multiple descriptors are extracted dynamically to capture various visual information of images, including HOG (Histogram of Oriented Gradients), color histogram and SIFT (Scale Invariant Feature Transform). Each of these descriptors is generated independently and image features are often expensive to generate and store and therefore may exist in a streaming format. Another example is the Mars crater detection from high resolution planetary images [12]. Tens of thousands of texture-based features, in different scales and different resolutions, can potentially be generated for high resolution planetary images. It is infeasible to acquire the entire feature set which means to have a near global coverage of the Martian surface. In order to deal with this challenge, many online streaming feature selection methods have been proposed [13].

Online feature selection with streaming features has attracted much attention in recent years and played a critical role in dealing with extremely high-dimensional problems [14–18]. Streaming features are defined as features that flow in one by one over time whereas the number of training examples remains fixed [15]. More specifically, Perkins and Theiler [19] considered the problem of online feature selection and proposed the Grafting algorithm based on a stagewise gradient descent approach. Grafting treats feature selection as an integral part of learning a predictor within a regularized framework. Zhou et al. [20] proposed two algorithms of information-investing and alpha-investing, based on

[*] Corresponding authors.
E-mail addresses: doodzhou@hotmail.com (P. Zhou), jsjxhuxg@hfut.edu.cn (X. Hu), peipeili@hfut.edu.cn (P. Li), xwu@louisiana.edu (X. Wu).

http://dx.doi.org/10.1016/j.knosys.2017.09.006
0950-7051/© 2017 Elsevier B.V. All rights reserved.

streamwise regression for online feature selection. Alpha-investing does not need a global model and it is one of the penalized likelihood ratio methods. Wu et al. [15] presented an online streaming feature selection framework with two algorithms called OSFS (Online Streaming Feature Selection) and fast-OSFS. OSFS contains two major steps, including online relevance analysis and online redundancy analysis. Yu et al. [18] proposed the SAOLA approach (a Scalable and Accurate Online feature selection Approach) for high dimensional data. SAOLA employs novel online pairwise comparison techniques and maintains a parsimonious model over time in an online manner. Eskandari et al. [21] proposed a Rough Set based method for online streaming feature selection. The proposed algorithm uses classical significance analysis concepts in Rough Set theory to control an unknown feature space in online streaming feature selection problems.

Meanwhile, in many real-world applications, such as fraud and intrusion detection, text classification and medical diagnosis, in addition to high dimensionality, class imbalance is also very common [22,23]. For example, the number of fraud users is obviously much lower than that of normal users. In fact, the ratio of the small to the large classes can be drastic such as 1 to 100, 1 to 1,000, or 1 to 100,000 [24,25]. All these online streaming feature selection approaches mentioned above were proposed to deal with data sets with normal class distributions, thus, they cannot handle the class imbalance data effectively. It is hence a challenge for existing online streaming feature selection approaches.

To handle the issue of class imbalance in data sets, existing solutions mainly focus on two levels: the data level and the algorithmic level [26]. The former includes many different forms of re-sampling and the latter involves adjusting the costs of various classes, the probabilistic estimation and the decision threshold. In recent years, one class learning and feature selection for class imbalance data have also attracted much attention [26]. Feature selection can be very helpful for imbalanced data sets [27]. The aim of feature selection for imbalance data is to select features that can get higher separability between the majority class and the minority class. Existing works in feature selection for imbalance data are mostly batch algorithms [27–31]. For example, Zheng et al. [28] proposed a feature selection framework which selects positive features and negative features separately, and then explicitly combines them to improve the classification accuracy in the handling of class imbalance data. Hulse et al. [27] gave detailed comparisons of six commonly-used filters and three filters using classifier performance metrics on high-dimensional imbalance data. The biggest finding from this paper is that feature selection is beneficial to handle most high-dimensional imbalanced data sets. Wasikowski et al. [30] presented a first systematic comparison of three types of methods developed for imbalanced data classification problems (re-sampling, new algorithms and feature selection) and of seven feature selection metrics evaluated on small sample data sets from different applications. Results showed that the signal-to-noise correlation coefficient (S2N) and Feature Assessment by Sliding Thresholds (FAST) are great candidates for feature selection in most imbalanced applications, especially in the case of selecting a very small number of features. Maldonado et al. [29] proposed a backward elimination approach based on successive holdout steps, whose contribution measure is based on a balanced loss function obtained over an independent subset. Nevertheless, all aforementioned algorithms were proposed for traditional feature selection. To the best of our knowledge, there is no work relevant to online streaming feature selection for high-dimensional class imbalance data so far.

Rough set theory, proposed by Pawlak, has been proven to be an effective tool for feature selection, rule extraction and knowledge discovery [32]. Pawlak's rough sets were originally proposed to deal with categorical data. There are many works of using rough sets for attribute reduction and feature selection [33–36]. However, in real-world applications, there are many numerical features in data sets. Then, a neighborhood rough set that supports both continuous and discrete data was proposed to deal with this challenge [37]. There are some works using the neighborhood rough set for feature selection [38–42] and it has been proved as an effective approach in the handling of feature selection problems. We aim to apply neighborhood rough set theory in the handling of online streaming feature selection. This is because rough set based data mining does not require any domain knowledge [21]. In addition, we focus on class imbalance data where instances of the minority class are rare in data sets. We refine the neighborhood rough set method in this paper and use neighbors's class information for feature selection. The proposed neighborhood rough set based method does not need to consider the global class distribution of a data set which makes the impact of class imbalance be relatively small. However, all these neighborhood rough set based methods mentioned above were designed for traditional batch feature selection and there is no work of using a neighborhood rough set for feature selection in an online manner.

We would like to distinguish online streaming feature selection in this paper from previous studies of dynamic information systems in [43–47]. A complete information system is defined as $S = (U, C \cup D, V, f)$, where U is a non-empty finite set of objects, C is the set of condition attributes and D is the set of decision attributes. $V = \bigcup_{a \in A} V_a$, where V_a is a domain of attribute a and $A = C \cup D$. f: $U \times A \to V$ is an information function such that $f(x, a) \in V_a$ for every $x \in U$, $a \in A$. Nowadays, the dynamic information system learning approaches based on Rough Set theory mainly focus on the following two cases. (1) The object set in an information system evolves over time while the attribute set remains constant. (2) The attribute set in the information system evolves over time while the object set remains constant. Most existing efforts consider the situation where objects or features are available in the information system [48–51]. It is different from online streaming feature selection where the number of objects is fixed and the feature set grows with time. At each time stamp, we can just get one feature from the streaming features and the full feature space is unknown or inaccessible.

Motivated by this, in this paper, we first formalize the problem of online streaming feature selection for class imbalance data and then present an Online Feature Selection framework based on the dependency between the condition features and decision classes, named OFSD. Our contributions are as follows:

- We formally define the problem of online streaming feature selection for class imbalance data.
- We propose an online feature selection framework based on the dependency of either a single feature or a selected feature set to decision classes. To the best of our knowledge, all existing online feature selection methods measure features by a certain criterion individually and there is no related work considering the selected feature set as a group.
- A new Online Feature Selection algorithm based on the dependency between condition features and decision classes in K nearest neighbors (called K-OFSD) is proposed to handle the class imbalance data in an online manner. In order to select features which can get high separability between the small class and the large class, we make a full use of the neighbors' class information near to the target object. For the class imbalance problem, we refine the neighborhood rough set theory and use the information of a fixed number of nearest neighbors to select features, which can promote the dependency between condition features and decision classes.
- Extensive experimental studies show that our proposed algorithm can get better performance than traditional imbalanced