Neurocomputing 205 (2016) 463–471

Contents lists available at ScienceDirect

Neurocomputing

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

Selecting discriminative features in social media data: An unsupervised approach

Elham Hoseini, Eghbal G. Mansoori

Shiraz University, Iran

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form
19 February 2016
Accepted 27 March 2016
Communicated by Jitao Sang
Available online 11 May 2016

Keywords:
Unsupervised feature selection
Social media
Link information
Graph partitioning

ABSTRACT

The usage of high-dimensional data complicates data processing in social network area. Accordingly, the researchers are motivated to propose some novel approaches to overcome this challenge. One of the best solutions is extracting the effective information from data pool and discarding the unnecessary ones. Feature selection is a known technique which aims to distinguish the discriminative features. Because of the unlabeled nature of datasets in social network, an unsupervised feature selection algorithm might be a good scenario. In addition to features, we try to confront the inherently linked users in social network datasets. This is because a stronger unsupervised feature selection technique is needed to ignore the independent and identically distributed assumption of data. Hence, by optimizing a novel objective function in this paper, the top-ranked features are extracted for further processing. This objective function incorporates both the inter-relationship of users in addition to their features. An efficient iterative algorithm is also designed to optimize the proposed objective function. We compare our method with two supervised and unsupervised evaluation criteria on real-world social network datasets. The experimental results demonstrate the effectiveness of our proposed approach.

1. Introduction

By rapid growing of social network services such as Facebook [49], Flickr [52] and Twitter [50] in recent years, millions of users participate in online social activities. A social network is a website that brings people together to connect with friends, share ideas and interests (e.g., textual posts, photos, videos, music and other personal information), or make new friends [42]. This type of collaboration and sharing of information is often referred to as social media.

In social media [40], the existing user information includes biographic facts (e.g., age, gender, location and marriage status), personal interest (e.g., politics, entertainment, and sports), occupation information (e.g., researcher, student, software engineer, and musician), social media activities (comments, follow, tags, like) and etc. This personal information is usually referred as user attributes (features) in social media. Inferring such user features can benefit many applications in information retrieval, personalization and recommendation.

Since the number of user features is very high in most social networks, this leads to high- dimensional data whose processing would be an undeniably challenge. Additionally, user features are not always available. This is because the users are likely to provide the easy-to-fill basic information such as name and gender, but seldom introduce their interests and other detailed information. Also, due to the privacy issues, most sites of social network limit the access to some personal information.

One effective approach to handle high-dimensional and sparse user attributes is feature selection [1,2]. This tries to select a subset of high discriminative features from a pool. Feature selection aims to minimize the redundancy of features while maximizing the relevancy to the target (class label). This is done to improve the performance of learning models by alleviating the curse of dimensionality and speeding up the learning process. Also, it improves the generalization capability of a learning model [3,4,17].

Feature selection algorithms are categorized into supervised and unsupervised models. In supervised methods which the training data has labels, the correlated features are accessed based on distinguishing distinct classes. These methods generally suffer from high complexity [11]. On the other hand, unsupervised feature selection algorithms use unlabeled data. These methods are particularly difficult since the definition of relevancy in features becomes unclear [5–7,10]. Furthermore in high-dimensional data, it is likely to find many sets of features that seem equally good without considering additional constraints [5,6]. Most of existing feature selection algorithms work with independent and identically distributed data. However, social media data is inherently linked which adds further challenges to feature selection process.

E-mail addresses: hoseini-e@shirazu.ac.ir (E. Hoseini), mansoori@shirazu.ac.ir (E.G. Mansoori).

http://dx.doi.org/10.1016/j.neucom.2016.03.078
0925-2312/

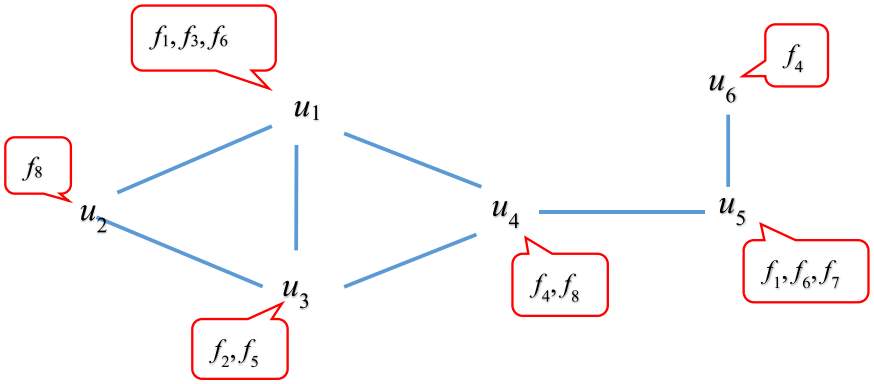

Fig. 1. A sample of linked social media data with six users and eight features.

As stated, social network datasets consist of some users which are probably connected to some like-minded participates. Each user has its own features which might be composed of its provided tags, personal information, etc. Fig. 1 shows a sample of linked data in social network where the edges represent the inter-user relations. It consists of six distinct users with eight features. These features might represent age, gender, marriage status, religious, and number of followers, followings, comments and likes, respectively.

Clearly, in most social networks, only some features are available for a single user [22]. This is because the users usually fill in their basic fields such as name and gender, but seldom introduce their interests and other detailed information. Additionally, due to the privacy issues, most social network sites limit the access to some personal information. Accordingly, in the example shown in Fig. 1, user u_1 has features f_1, f_3 and f_6 but not features f_2, f_4 and f_5.

In social media, two linked instances are more likely to have similar interests than two randomly picked ones. However, most existing feature selection algorithms only work on data features and seldom, they are capable to consider the relationship among data. This is due to lack of information about the relationship among instances (users in social media) in most datasets related to various fields. However in social media data, the relationship between users is available. Incorporating this information in feature selection algorithm is a suggestive motivation for selecting most discriminative features. In this paper, by considering linked property of social network data, a novel feature selection algorithm is developed. We have tried to exploit and model the relations among data instances and to employ them for feature selection in an unsupervised scenario.

In supervised learning model, label information plays the role of constraint. Without labels, alternative constraints such as data variance and separability can be used [1,8,9]. Because of the importance of discriminative information in data analysis, it is beneficial to exploit this discriminative information for feature selection. However, selecting discriminative features in unsupervised scenarios is a significant but hard task due to lack of labels. In this regard, to consider both relationship among users and discriminative information, we have proposed an unsupervised feature selection framework for social network data (called UFSS). It tries to take into account these contributions: (i) users partitioning by applying their inter-relationship information, and (ii) seeking for a subset of features with high users' discrimination.

The rest of this paper is organized as follows. The related work is presented in Section 2. Our new framework of unsupervised feature selection for social network, UFSS, is introduced in Section 3, including approaches to capture link information, optimization, and convergence analysis. The experimental results with discussion are presented in Section 4. We conclude this work in Section 5.

2. Related work

Feature selection methods are categorized into three main schemes: filters, wrappers and embedded models. Filter methods apply a statistical measure to assign a scoring value to each feature individually [15]. Then, features are ranked according to their scores and either selected to be kept or removed from the dataset. These methods are often univariate and consider the features independently. Also, feature selection is performed as a preprocessing stage. The well-known filter models are ReliefF [30], multi-class ReliefF [30], mRMR [31], SPEC [19], Laplacian score [24] and its extensions [19].

On the other hand, wrapper methods consider the selection of a set of features as a search problem. That is, different combinations are prepared, evaluated by a predictive model and compared to each other according to score of model accuracy [9,11]. In the embedded models, however, feature selector is a combination of both filter and wrapper. These methods learn which features best contribute to the accuracy of the model while creating the model. Since the wrapper methods can adapt themselves to the machine learning algorithm, they are likely to have better results than filter methods. However, the wrappers are computational costly [33]. Nevertheless, because of their inherent advantages, most of the feature selection algorithms use wrapper models.

In other ways, the feature selection scheme is divided into supervised and unsupervised methods according to whether the training data is labeled or unlabeled [1]. Supervised methods like Fisher score [32] and ReliefF [30] usually are more reliable than unsupervised ones. However, these methods suffer from two main restrictions. First, since they evaluate each feature independently, they ignore the correlation between features. Second, access to labeled training data in real world is too expensive. Nevertheless, much attention has been paid to unsupervised feature selection in recent years.

The lack of constraints in unsupervised feature selection model may lead to the creation of several subsets of similar features. Also, how to evaluate them is a major challenge. In this regard, maximum variance criterion (MaxVar) [44] can be the most effective and easiest method to evaluate selected features. This criterion projects all data instances on the dimension with maximum variance. Though it looks for features that are useful for display, no assurance is given that these characteristics of different classes are separated effectively. Recently, some filter methods like Laplacian score [24] and its extensions [19] have been proposed for unsupervised learning. Laplacian score is a filter method which uses the nearest neighbor graph to model the local geometrical structure of the data and to extract the best features based on the structure of the graph. Also in [19], a spectral feature selection algorithm (SPEC) is proposed. Its basis is on sparse multi-output regression by considering $l_{2,1}$ norm. This algorithm performs well in both