Information Fusion 40 (2018) 1–17

Contents lists available at ScienceDirect

Information Fusion

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

A Social-aware online short-text feature selection technique for social media

Antonela Tommasel*, Daniela Godoy

ISISTAN, UNICEN-CONICET. Campus Universitario, Tandil (B7001BBO), Argentina

ARTICLE INFO

Article history:
Received 30 December 2016
Revised 4 May 2017
Accepted 16 May 2017
Available online 17 May 2017

Keywords:
Online feature selection
Classification
Micro-blogging communities

ABSTRACT

Large-scale text categorisation in social environments, characterised by the high dimensionality of feature spaces, is one of the most relevant problems in machine learning and data mining nowadays. Short-texts, which are posted at unprecedented rates, accentuate both the importance of learning tasks and the challenges posed by such large feature space. A collection of social media short-texts does not only provide textual information but also topological information given by the relationships between posts and their authors. The linked nature of social data causes new complementary data dimensions to be added to the feature space, which, at the same time, becomes sparser. Additionally, in the context of social media, posts usually arrive simultaneously in streams, which hinders the deployment of efficient traditional feature selection techniques that assume a feature space fully known in advance. Hence, efficient and scalable online feature selection becomes an important requirement in numerous large-scale social applications. This work presents an online feature selection technique for high-dimensional data based on the integration of two information sources, social and content-based, for the real-time classification of short-text streams coming from social media. It focuses on discovering implicit relations amongst new posts, already known ones and their corresponding authors to identify groups of socially related posts. Then, each discovered group is represented by a set of non-redundant and relevant textual features. Finally, such features are used to train different learning models for classifying newly arriving posts. Extensive experiments conducted on real-world short-texts demonstrate that the proposed approach helps to improve classification results when compared to state-of-the-art and traditional online feature selection techniques.

1. Introduction

Large-scale text categorisation in social environments is one of the most relevant problems in machine learning and data mining nowadays. With social media data growing at unprecedented rates, this problem becomes a matter of paramount importance for numerous real-world applications. For example, tweets could be classified aiming at discovering breaking news or events (such as natural disasters) helping to understand the impact of incidents, or assisting in emergency management and crisis coordination. Additionally, trending topics or social trends could be discovered by analysing clusters of related tweets.

The pervasive use of social media offers research opportunities for analysing user behaviour and how they interact with their friends. Unlike social connections formed by people in the physical world, social media users are free to connect with a wider number of people for a variety of reasons. The low cost of link formation might lead to networks with heterogeneous relationship origin or strength. For example, in *Twitter*, a user might follow others because they publish interesting information, they have the same interests, or even because they share some common friends, amongst other possible explanations. In addition to social information indicating friendship or simply user interaction, there are other information sources that might implicitly define connections between users in social media. For example, whether two users use the same terms, hashtags, or post about the same topic. Moreover, the social media experience of users is no longer limited to a unique site, as users use social media for different purposes [49]. As a result, each social media site provides heterogeneous and complementary information sources for describing a particular user, their interests and social relations.

A task that can greatly benefit from the integration of multiple information sources is text categorisation. Such task is characterised by the high dimensionality of their feature space where

* Corresponding author.
E-mail addresses: antonela.tommasel@isistan.unicen.edu.ar (A. Tommasel), daniela.godoy@isistan.unicen.edu.ar (D. Godoy).

http://dx.doi.org/10.1016/j.inffus.2017.05.003
1566-2535/© 2017 Elsevier B.V. All rights reserved.

most terms have low frequencies. This situation is commonly known as the curse of dimensionality, which refers to the increasing computational complexity of learning problems as the volume of data grows exponentially regarding the underlying space dimension. This problem worsens when considering short-texts, such as tweets, *Facebook* posts or even blogs' social annotations. Nonetheless, a collection of short-texts in social media does not only provide textual information but also topological information due to the relationships between posts and users. In turn, the linked nature of social media data causes new dimensions (such as friendship relations between users) to be added to the feature space [34]. The increasing amount of data does not only affect the computational complexity of algorithms, but also poses new challenges regarding how to represent and process new data, and how to effectively leverage on such data for improving the performance of text learning tasks [7].

Feature selection (FS) [3] is one of the most known and commonly used techniques to diminish the impact of the high-dimensional feature space by removing redundant and irrelevant features. The standard FS setting assumes the existence of instances, and therefore a feature space, fully known in advance. Thus, FS consists in finding a small subset of the most relevant features according to certain evaluation criterion. This setting is known as batch FS. However, in real-world applications, and particularly social media ones, such assumptions might not hold as either training examples could arrive sequentially, or it could be difficult to collect the full training set [44]. For example, in the context of social media data, posts usually arrive simultaneously in streams, hindering the deployment of efficient and scalable batch FS techniques. Thus, traditional batch FS techniques are not suited for emerging big data applications. In these situations, online feature selection (OFS) in which instances and their corresponding features arrive in a continuous stream, needs to be performed. This process involves choosing a subset of features and the corresponding learning model at different time frames. Thereby, OFS is particularly important in real-world systems in which traditional batch FS techniques cannot be applied.

Motivation

Although FS techniques have received considerable attention during the last decades, most studies focus on developing batch techniques instead of facing the challenging problem of OFS. The majority of FS techniques are designed for data containing uniform features, which are typically assumed to be independent and identically distributed. However, this assumption might not hold in social media since measuring the relevance of features in isolation possibly ignores dependencies amongst them given by the social context. Interestingly, most algorithms only focus on content-based information sources, even though social media content might be topically diverse and noisy, which hinders the effective identification of relevant and non-redundant features. It is worth noting, linked data has become ubiquitous in social networks, as in *Twitter* (in which not only tweets can be linked, but also their authors might be socially related) or *Facebook* (in which users share friendship relationships), providing additional information sources such as correlations between instances. For example, posts from the same user or two linked users are more likely to have similar topics. As the different information sources provide complementary views of data, when assessing them independently, algorithms may fail to account for important data characteristics. Instead, FS techniques should be capable of combining multiples information sources. In this context, the availability of link information enables advanced research in FS techniques, which needs to address two challenges: how to exploit relations amongst data instances, and how to leverage those relations for FS.

Efficient and scalable OFS is an important requirement for numerous large-scale social applications. Despite presenting significant advantages in efficiency and scalability, existing OFS techniques do not fully leverage on the multiple information sources available. Instead, they mainly focus on textual information. Potentially, the performance of such approaches could be improved by including additional information sources in social media data. Furthermore, most of the approaches that claim to be applicable in OFS, might fail when used in the context of social media data, due to the need of knowing either all data instances or features in advance, making them unsuitable for data streams. In consequence, novel approaches for efficiently selecting and updating the selected subset of features need to be developed.

Considering that different information sources in social media can provide multiple and possibly complementary views about data, this paper aims at addressing the OFS task for high-dimensional short-text data arriving in a stream. The hypothesis behind this work is that more accurate OFS techniques could be developed by effectively integrating multiple information sources. The main goal of this work is to define and evaluate a new intelligent technique for short-text mining to enhance the process of knowledge discovery in social media. To that end, an OFS technique for leveraging on social information to complement commonly used content-based information is presented. The technique is based on the integration of social network structures into the process of OFS [42].

Unlike other works found in the literature, the focus of the presented technique is to analyse different types of social relationships between posts and their authors. Particularly, this work aims at performing real-time classification of continuously generated short-texts in social networks by exploring the combination of multiple relations amongst data instances in the social environment and how to leverage such multiple relations for enhancing FS techniques. The goal is to discover implicit relations between new posts and already known ones, based on a network comprising the individual posts and the users who have written them. Then, the content in the discovered groups of socially related posts is analysed to select a set of non-redundant and relevant features to describe each group of related posts. Finally, such features are used for training different learning models to categorise newly arriving posts.

Contributions

The expected contributions of this work are described as follows. First, it tackles the problem of how to exploit social relations amongst data instances by studying the linked nature of social media data. Second, it proposes a technique for leveraging on those social relations. Third, it combines social information with the content of posts for effectively and efficiently performing FS. Fourth, the technique is scalable, and thus appropriate for real-time environments in which neither features nor instances are known in advance. Furthermore, it allows the process of data instances as they are generated in a reasonable amount of time. Finally, the presented technique could help in the development of new and more effective models for personalising and recommending content in social environments.

The rest of this paper is organised as follows. Section 2 discusses related research on OFS. Section 3 presents the proposed OFS technique combining two heterogeneous and complementary information sources: social and content-based information. Section 4 describes the experimental settings and results obtained for two social media datasets. Finally, Section 5 summarises the conclusions drawn from this study, and presents future lines of work.