Integrating Social Network Structure into Online Feature Selection

Antonela Tommasel
ISISTAN Research Institute, CONICET-UNCPBA
antonela.tommasel@isistan.unicen.edu.ar

Abstract

Short-texts accentuate the challenges posed by the high feature space dimensionality of text learning tasks. The linked nature of social data causes new dimensions to be added to the feature space, which, also becomes sparser. Thus, efficient and scalable online feature selection becomes a crucial requirement of numerous large-scale social applications. This thesis proposes an online feature selection technique for high-dimensional data based on both social and content-based information for the real-time classification of short-text streams coming from social media. The main objective of this thesis is to define and evaluate a new intelligent text mining technique for enhancing the process of knowledge discovery in social-media. This technique would help in the development of new and more effective models for personalisation and recommendation of content in social environments.

1 Introduction

With social media data growing at unprecedented rates, the problem of large-scale text learning in social environments becomes a matter of paramount importance. Text learning tasks are characterised by the high dimensionality of their feature space where most terms have a low frequency. Indeed, they are susceptible to the "curse of dimensionality", which refers to the increasing computational complexity as the volume of data grows exponentially. This problem worsens when considering short-texts, such as tweets, *Facebook* posts or blogs' social annotations. Feature selection (FS) is one of the most used techniques to reduce the impact of this problem by removing redundant and irrelevant features.

Although FS has recently received considerable attention, most studies focus on batch techniques instead of facing the challenging problems of online feature selection (OFS). Moreover, they are designed for data assumed to be independent and identically distributed, not fully leveraging on linked data. Linked data has become ubiquitous in social networks, e.g. *Twitter* and *Facebook* users are socially related, providing extra information such as correlations between instances. Hence, two challenges need to be addressed: how to exploit relations among data instances, and how to leverage those relations for FS. Interestingly, FS for link data is rarely addressed, and approaches considering it are generally unsuitable for online environments. Most techniques claiming to be suitable for online settings might fail when considering social media data as they need to know all features or instances in advance, resulting unsuitable for data streams. Also, they have a high computational complexity and lack of updates of the selected feature set. Thus, new techniques for efficiently selecting and updating feature sets need to be developed.

This thesis presents a technique for leveraging on social information to complement content-based information for performing OFS. It aims at the real-time classification of continuously generated short-texts in social networks. Unlike other works in the literature, the focus is on analysing diverse social relationships between posts and their authors. The goal is to discover implicit relations between new posts and already known ones, based on a network comprising 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 of them. Finally, features are used for training learning models to categorise newly arriving posts.

2 Social-based OFS

The proposed approach aims at addressing the massive-scale OFS task for high-dimensional short-text data arriving in a continuous stream, in which neither features nor data instances are fully known in advance. The rest of this section summarises the most important aspects of the approach's general methodology, which can be described as follows:

1. Data Modelling. Data is modelled as a graph representing the social posts and their relations.

2. Social Analysis Step. Social relationships between posts are analysed to find groups of socially related posts.

3. Content Analysis Step. An optimal feature set to describe each group of socially related posts is found.

4. Model Learning. Content features are used for training learning models for classifying newly arriving posts.

5. Arrival and classification of new posts. When a new post arrives, its relations with the known posts are exploited to find its most similar posts and thus, its content-based representation. Based on such representation, the post is classified.

6. Re-run of the Social Analysis Step. After new posts are classified, the feature space is updated.

2.1 Social Analysis

The focus of the approach is to analyse the social relationships between social posts and their authors to detect groups of socially related posts. In this regard, a social graph is created to define implicit relations between new posts and already known ones based on a network comprising the individual posts and the users who have written them. In social media data, both the graph topological structure (i.e. social relations between users) and the vertex properties (i.e. posts characteristics) are important. Hence, besides the relations between posts derived from the social relations between their authors, additional content-related relations could be defined. For example, the resemblance of posts' content or metadata could be used to reinforce (i.e. weighting) the social relations found among them. Thus, it is important to adequately compute the similarity among nodes in order to fully leverage the information convened by the network relations.

2.2 Content Analysis

Once the sets of socially related posts are discovered they are individually analysed to find the optimal feature set to describe them. An optimal feature subset should include all relevant and none redundant features. Then, the feature sets are used for training specialised classifiers for further distinguishing posts that might also belong to the set. Traditionally, the focus has been only on identifying relevant features. However, only assessing feature relevance cannot identify redundant features as they are likely to have similar rankings, and thus will also be selected.

2.3 Classification of newly arriving posts

When a new post to be classified arrives, the group of socially related posts it belongs to is first determined by considering its social and content-based relations with the posts in the social graph. Such group defines the textual features to represent the new post and the trained classifier to use. Once the most similar community or communities are found, several alternatives for assigning the post to a category are proposed. After posts are classified, they are added to the social graph to update the underlying community structure. Thus, the feature space describing communities is periodically updated for coping with the continuous evolution of topics and the newly discovered posts. Finally, social media is a highly dynamic environment in which new topics not only constantly emerge, but also become obsolete and tend to disappear. As a result, strategies for removing old post from the social graph after a certain time they were added could be also implemented.

3 Current State

Preliminary evaluations conducted on two real-world short-texts datasets achieved promising results (Figure 1) when compared to traditional and state-of-the-art (e.g. [Wang *et al.*, 2013; 2014; Zubiaga *et al.*, 2015]) baselines specifically defined for social media, in batch and online settings. The obtained results exposed the limitations of pure content-based techniques for classifying social media short-texts. Hence, they evidenced the need of considering social information, and its advantages for selecting the most relevant feature set. Results highlighted the difficulty of the task and the need of continuing to develop new techniques. In summary, results allowed to confirm that purely content-based strategies might not be sufficient for classifying social media texts. Thus, leveraging on social information becomes crucial for OFS.

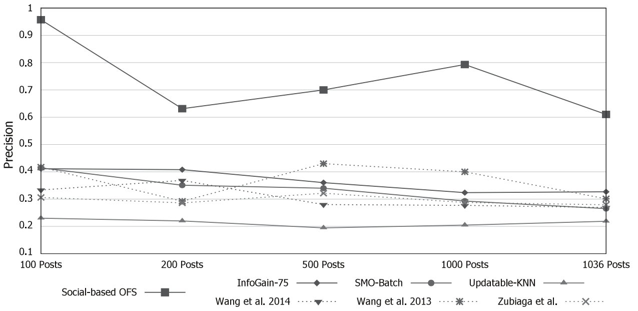

Figure 1: Precision Results - *Twitter* dataset

4 Conclusions

This thesis tackles the challenging problem of online feature selection, which is an important requirement in numerous large-scale social applications. The main contributions of this work are described as follows. First, it addresses the problem of how to exploit the linked nature of social media data. Second, it proposes a technique for leveraging on social relations. Third, it combines social information with content for effectively and efficiently performing feature selection. Fourth, the technique is scalable, hence appropriate for real-time environments in which neither features nor instances are known in advance. Furthermore, it allows to process data instances as they are generated in a reasonable amount of time. Finally, although the approach is designed to be applied to multi-class classification of social posts, it can be also applied in binary-class settings, and even in other learning tasks, such as unsupervised and semi-supervised environments, in which none or only a small labelled dataset is available.

As regards future work, an extensive experimental evaluation must be performed. New alternatives for further exploiting the topology of social relations and communities are under consideration. Thus, the social graph could be enriched by defining additional relations between posts. Moreover, the performance of several methods for assessing the redundancy and relevance of features will be analysed. Finally, the usefulness of overlapping communities will be explored.

References

[Wang *et al.*, 2013] J. Wang, Z. Zhao, X. Hu, Y. Cheung, M. Wang, and X. Wu. Online group feature selection. In *Proceedings of the 23rd IJCAI*, pages 1757–1763. AAAI press, 2013.

[Wang *et al.*, 2014] J. Wang, P. Zhao, S.C.H. Hoi, and R. Jin. Online feature selection and its applications. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):698–710, 2014.

[Zubiaga *et al.*, 2015] A. Zubiaga, D. Spina, R. Martínez, and V. Fresno. Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473, 2015.