Learning to Classify Users in Online Interaction Networks

Georgios Rizos, Symeon Papadopoulos, Yiannis Kompatsiaris

Information Technologies Institute, CERTH
Thessaloniki, Greece
{georgerizos,papadop,ikom}@iti.gr

ABSTRACT

We study the problem of user multi-label classification in settings where two types of information are available: a) a set of seed users with known labels, b) the online interactions among the users of interest. User labels may refer to topics of different granularities (e.g. broad themes, news stories, etc.), user types (e.g. person, news agency, etc.), political stance (e.g. liberal, conservative) and others. To tackle the problem, we propose a semi-supervised learning framework that represents users by means of network-based features. We propose the use of Absorbing Requlatized Commute-Time Embedding (ARCTE) as a means to extract local graph features and devise a computationally more efficient scheme (compared to existing ones) for their extraction. We then compare the results of this representation with a number of previously proposed alternatives on a Twitter dataset of 534K users. We also discuss a few key practical issues as well as the repercussions of the proposed approach with respect to privacy in online networked environments.

1. INTRODUCTION

Our research focuses on multi-label user classification in the context of Online Social Networks (OSNs). This problem has profound applications in the fields of online search, recommender systems and user profiling, and it is particularly challenging since most OSN users do not explicitly state their interests, leading to sparsely labelled, noisy data and also due to the massive size of the raw data involved. When considering potentially informative signals for this problem, user generated content such as text messages is a strong cue. However, it may prove to be noisy or insufficient due to brevity (e.g. Twitter posts), ambiguity or multi-languality. Our approach relies on the structure of interaction networks and has its basis in social science. According to the principle of *homophily* [3], people sharing the same beliefs and interests tend to interact with each other and as such we expect them to form denser than average communities.

We formalize user classification as a semi-supervised learning problem: given a set of interactions among the OSN users of interest, we first build an interaction graph, then we extract graph-based features and using the known labels of a small number of seed nodes for training, we predict the labels for the rest of nodes. The framework is illustrated in Figure 1.

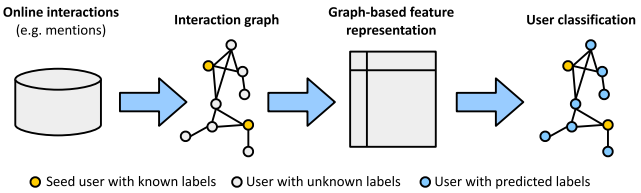

Figure 1: Overview of user classification framework

2. APPROACH OUTLINE

As a first step, a graph embedding or community detection method is employed to extract the feature-based representation by projecting the nodes to a latent space. The coordinates of the instances in this latent space are denoted by their matrix form $\mathcal{X}$. The coordinates of node i, $\mathcal{X}_{i:} \in \Re^k$, where k is the dimensionality of the latent space. The graph includes both the labelled $\mathcal{X}_l$ and unlabelled $\mathcal{X}_u$ nodes. The framework qualifies as semi-supervised learning regardless of whether a supervised or semi-supervised approach is selected to train a classification model $\mathcal{M}$ based on $\mathcal{X}$, because the full graph is used to extract $\mathcal{X}$. To deal with the multi-label nature of the problem, any multi-label classification scheme may be used [5], such as One-vs-All.

To construct $\mathcal{X}$, our approach is to search locally around each node in the graph to find similar nodes. We divide our approach into two steps to get two sets of communities: $\mathcal{X} = \text{embed}(\mathcal{C}_1 \cup \mathcal{C}_2)$. The first set of output communities is the set of base communities. Each base community around node j is defined as its immediate neighbourhood plus the node itself: $c_j = \mathcal{N}(j) \cup j, \forall j \in \mathcal{N}$. As for the second set, we want to identify sets of nodes around each user by searching locally for highly probable random walk destinations. Specifically, we first compute the Regularized Commute-Time kernel [2] (a variant of the well-known PageRank), then sort nodes according to the degree-normalized ranking, and select the fewest possible highest ranking nodes as the local community such that they comprise a strict superset of the corresponding base community. A key contribution of our approach involves a much more efficient means of computing the Regularized Commute-Time kernel, by avoiding self-transitions of probability, making it applicable to very large interaction networks.

3. EXPERIMENTS & KEY RESULTS

We tested our approach on the interaction graph (comprising replies/mentions) extracted from the SNOW dataset [4]. The largest weakly connected component of this graph consists of 533,874 users and 965,821 mention/reply edges. This graph was used to generate the feature-based representations

$\mathcal{X}$ for all nodes. User labels were then generated for 13,000 users of this graph, by computing the PageRank on the mention graph and selecting the top ranking users. For each of those, we used the Twitter API to collect up to 500 lists that these users belong to. The list names and descriptions were tokenized, stop words were removed, and the remaining tokens were lemmatized. With manual inspection, some lemmas were removed and others were merged into a final list of labels, which is illustrated in Figure 2. Using TF-IDF scoring, we computed a user-label matrix and then selected as "correct" labels, those that were associated with a user with a score equal to, or higher than the 75th percentile of normalized frequencies. We then used 1%, 2%, 3%, 5% and 10% of nodes for training and the remaining for testing, using an SVM to learn the feature-based representation. Classification accuracy was measured in terms of the F-score, micro-averaged over the F-scores derived for each label. To test the effectiveness of our proposed representation (ARCTE), we compare it with the F-score achieved using previously proposed representations, namely the Laplacian Eigenmaps (LapEig) [1], the Multiple Resolution Overlapping Communities (MROC) [6], and the base communities (BaseCom) defined above. The results of Table 1 were computed based on the six `stories` labels of Figure 2.

The results demonstrate that the proposed representation is more effective with respect to multi-label user classification leading to improved accuracy over competing methods, especially when the percentage of labelled nodes is very low. Furthermore, the proposed approach is the fastest to compute compared to previously proposed ones (with the exception of the BaseCom representation, which is trivial to extract). Overall, it appears that ARCTE offers an effective and at the same time scalable method to perform multi-label user classification.

Practical issues: The fact that the feature representation of users is generated based on their online interactions has two profound issues. First, online interactions appear in a streaming fashion, hence the structure of the corresponding interaction networks and ultimately the values of the extracted features will depend on the period and time of observation. In the case of the SNOW dataset, for instance, the interactions were recorded over the course of 24 hours. One could therefore doubt about the validity or stability of labels for those users. To address this concern, longer observational studies of a longitudinal nature would be required. Second, several of the observed interactions are not motivated by interest on a topic, but are the result of other social mechanisms (e.g. casual chat) or even spam (e.g. attempts to maximize one's reach). To mitigate the confounding impact of such interactions on the proposed analysis, appropriate filtering mechanisms should be devised. A related issue appears in the context of interactions between users that "agree" on a certain label, but disagree on others, necessitating the development of appropriate label-dependent homophily characterization mechanisms.

Table 1: Classification accuracy (micro-averaged F-score).

	1%	2%	3%	5%	10%
ARCTE	**0.4532**	**0.4959**	**0.5290**	**0.5586**	**0.5966**
LapEig [1]	0.4398	0.4632	0.4869	0.5199	0.5668
MROC [6]	0.4156	0.4461	0.4674	0.4937	0.5236
BaseCom	0.4019	0.4210	0.4295	0.4479	0.4792

Figure 2: User labels in the SNOW dataset

Privacy implications: It is noteworthy that purely based on interactions (i.e. without looking into the content of posts), an analyst is in a position to infer a variety of labels/attributes about OSN users. Given that several of the inferred labels are of sensitive nature (e.g. religion, political opinion) and that individuals are likely not aware that such profiling mechanisms are available, one should apply them with care and caution.

Acknowledgments: This work is supported by the Reveal and USEMP FP7 projects, partially funded by the EC under contract numbers 610928 and 611596 respectively.

4. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[2] F. Fouss, K. Francoisse, L. Yen, A. Pirotte, and M. Saerens. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural Networks*, 31:53–72, July 2012.

[3] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[4] S. Papadopoulos, D. Corney, and L. Aiello. SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media. In *Proceedings of the SNOW 2014 Data Challenge co-located with (WWW 2014)*, pages 1–8, 2014.

[5] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[6] X. Wang, L. Tang, H. Liu, and L. Wang. Learning with multi-resolution overlapping communities. *Knowledge and information systems*, 36(2):517–535, 2013.