Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbours

Faiyaz Zamal
faiyaz.zamal@mail.mcgill.ca

Wendy Liu
wendy.liu@mail.mcgill.ca

Derek Ruths
derek.ruths@mcgill.ca

Network Dynamics Lab (www.networkdynamics.org) / School of Computer Science / McGill University / Montreal, Canada

Introduction

Latent attribute inference is the mechanised study of the content generated by a person in order to infer information about that person. Past work in this field has made use of the microblogging network Twitter to generate feature vectors from users' microblog and profile content and then train and test with a classifier.

In this study, we extend existing work by leveraging the principle of homophily: a user's friends should also provide some information about the user herself. We use the features introduced in previous studies and evaluate the inference accuracy gained by using friend features to supplement or replace the user's features, for various neighbourhood subsamples. We consider three attributes with varying degrees of assortativity: age, gender and political affiliation.

Methods

For each attribute, we obtained profile data and up to 3000 microblog posts ("tweets") for approximately 200 users per binary label (Figure 1). Users were manually curated to account for misclassification and private accounts. For each user, the same data for the user's first 1000 friends were gathered.

Attribute	Label 1	Label 2	Labelling method
Gender	Male	Female	Common male/female names (US)
Age	18-23	25-30	Tweet mining: "Happy Xth birthday to me"
Politics	Democrat	Republican	Wefollow.com (Twitter user listings)

Figure 1. Labels and labelling methods for each attribute.

We then built a feature vector for each of the ~400 labelled users for each attribute, making use of content available in the user's profile and/or microblog (Figure 2). Subsequently, various ways of subsampling the users' neighbourhoods were tested, producing 13 configurations in total (Figure 3). Finally, support vector machines (SVM) and gradient boosted decision trees (GBDT) were used (separately) as binary classifiers for each attribute.

k-top discriminating features (k=20)
hashtags ('#hashtags'), words, digrams, trigrams, stems ('emerg'), costems ('ency')

Frequency features (per day)
tweets, retweets, hashtags, links, mentions

Ratio features
tweet:retweet, followers:followees (friends)

Figure 2. Features studied per user.

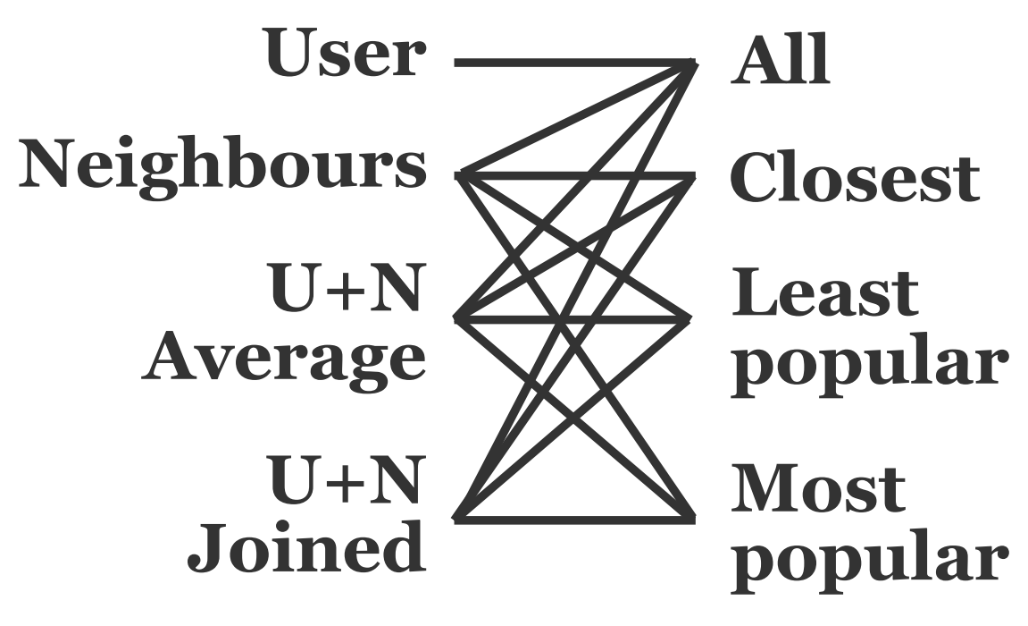

Figure 3. The 13 possible neighbourhood configurations.

Results

The results of the 10-fold cross validation for the SVM are shown in Figure 4 (GBDT was consistently 5–10% less accurate than SVM, but followed the same trends). The best results are reported for each group, with the relevant neighbourhood configuration for Neighbours and User+Neighbours highlighted in Figure 5. The sources for Prior work are listed under References.

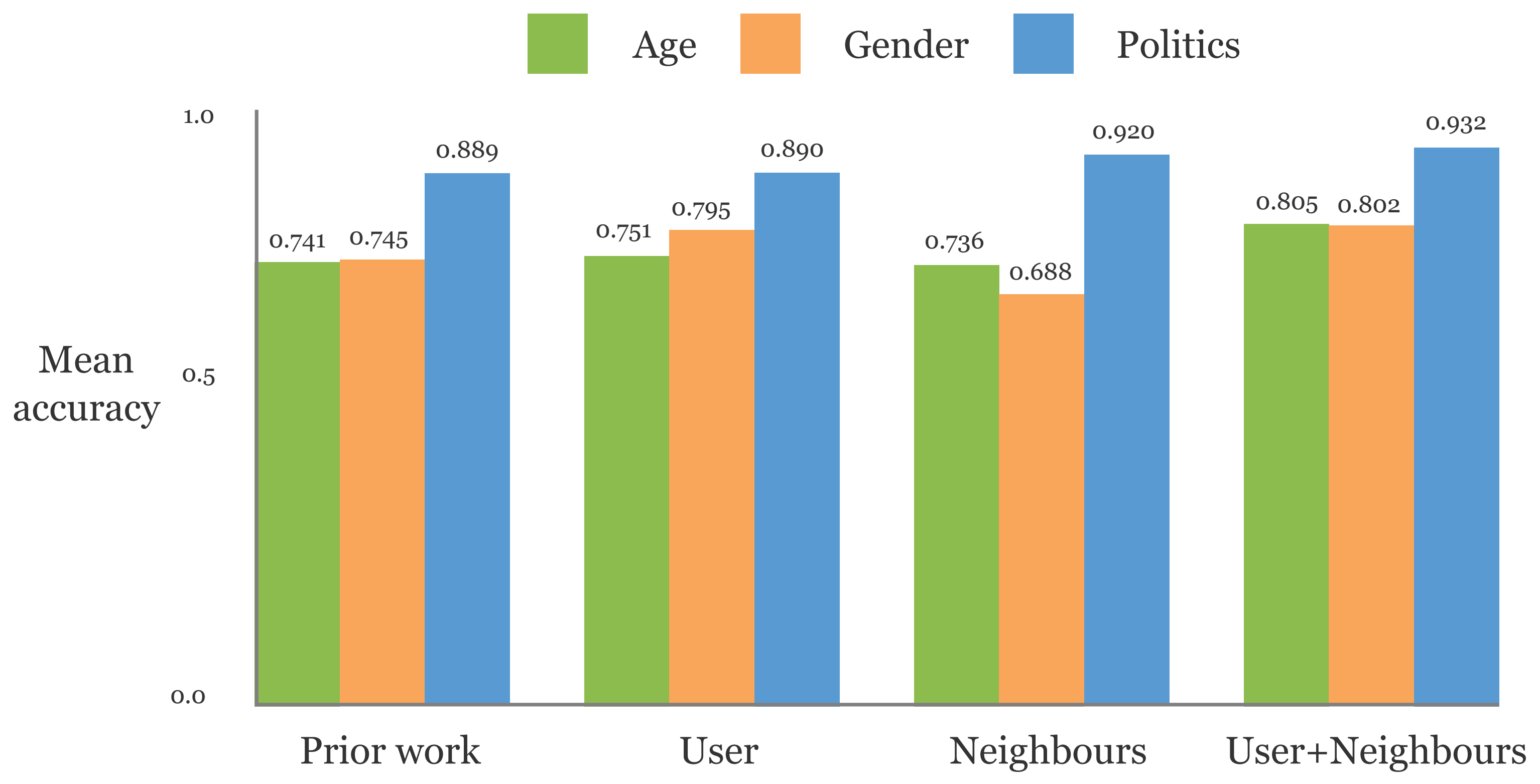

Figure 4. SVM classifier results for all attributes.

Configuration	Age	Gender	Politics
UserOnly	0.751	0.795	0.890
Nbr-All	0.736	0.669	0.920
Nbr-Most	0.619	0.688	0.777
Nbr-Least	0.691	0.560	0.725
Nbr-Closest	0.716	0.598	0.895
Avg-All	0.795	0.750	0.918
Avg-Most	0.739	0.749	0.885

Configuration	Age	Gender	Politics
Avg-Least	0.805	0.758	0.878
Avg-Closest	0.779	0.674	0.909
Join-All	0.764	0.799	0.932
Join-Most	0.741	0.755	0.889
Join-Least	0.782	0.774	0.873
Join-Closest	0.772	0.802	0.915

Figure 5. Accuracy rates for all configurations.

The results show that using a combination of user and neighbour information consistently leads to better classification when compared to prior work or the user-only configuration. However, each attribute behaved quite differently. For age, augmenting the user's information with that of her "least popular" friends—those who are likely to be personal acquaintances—resulted in the best performance. For gender, when only neighbours were considered, accuracy dropped greatly. For politics, neighbourhood data proved to be especially useful, with some neighbourhood-only configurations resulting in better accuracy than the user-only configuration.

Conclusions

Neighbourhood context can improve inference accuracy. For age and political affiliation, adding neighbourhood information resulted in statistically significant ($p < 0.002$) accuracy gains, with 21% and 38% improvement toward perfect inference, respectively. (Gender will be discussed below.)

Attribute assortativity influences accuracy gain. Age and political affiliation have been shown to be highly assortative in online communities by prior studies, explaining the accuracy gain when including neighbour information. Gender, on the other hand, has been shown in previous studies to have minimal assortativity, which explains the lack of improvement made by our method.

Choice of neighbourhood influences accuracy gain. For political affiliation, using all the neighbours resulted in the most accurate classification, suggesting that a random sampling of a user's neighbourhood provides the best reflection of the user's own political affiliation. For age, the least popular neighbours, who are likely to be the user's personal friends, offered the most signal. Gender did not have a neighbourhood configuration that significantly outperformed the others.

Neighbourhood data can be comparable to user data. In the cases of age and political affiliation, we find that when user features are omitted, the neighbourhood features alone are sufficient to obtain inference accuracy that is statistically on par with or better than when only user features are used. This does not, however, hold with gender, due to the limited assortativity of this attribute. Although the usefulness of neighbourhood information is limited for attributes with low assortativity, such information can still carry significant signal for other attributes, which is particularly useful when the profile and microblog data of the user in question are not accessible.

References

Age: Rao, D., and Yarowsky, D. 2010. Detecting latent user properties in social media. In Proceedings of the NIPS workshop on Machine Learning for Social Networks.
Gender: Burger, J.; Henderson, J.; and Zarrella, G. 2011. Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.
Politics: Pennacchiotti, M., and Popescu, A. 2011. A machine learning approach to twitter user classification. In Proceedings of the International Conference on Weblogs and Social Media.