Addressing Task Heterogeneity in Social Media Analytics

Arun Reddy Nelakurthi; Jingrui He

Arizona State University, Tempe AZ 85281

1 Introduction

In the recent years, social media has gained significant popularity and has become an essential medium of communication between individuals. The rise in social media usage both horizontally in the number of unique social media accounts, and vertically in the number of people using it led to an explosion of data. Popular social networks like Facebook, Instagram, and Twitter, manage tens of petabytes of information, with data flows of hundreds of terabytes per day, and continually expanding numbers of users. Such user-generated content provides an excellent scenario for applying the metaphor of mining any information. In a social media context, users create a massive amount of data where we can look for valuable nuggets of knowledge by applying different search (information retrieval) and mining techniques (data mining, text mining and web mining).

Unlike the traditional mining settings where the data is considered to be homogeneous for most of the mining tasks, the user generated social media data is intrinsically heterogeneous and thus pose a set of challenges. The health related social media websites, that are similar but cater users from different geographical locations can suffer from distributional shift in the user generated data, either the features or the class labels. Also, in opinion and sentiment related data like user reviews, the user behavior in one social network could be different to the user behavior in other network. The shift may be due to the user bias or the personal preferences of the user. The shift makes it a challenge to model the user behavior for cross-network classification tasks. Transfer learning addressed the problem of distribution shift in data, but there was not much research on modeling user behavior in transfer learning settings. In scenarios where parts of data in one social network is hidden, missing or not available, leveraging only parts of the information that is available for data mining tasks is also very challenging and not well studied. Finally, if successful in extracting the relevant information from social media analysis, does it impact the behavior or improve the lives of individuals using social media is another challenging problem. Motivated by the applications of task heterogeneity, in this work, I present my research proposal on techniques for addressing task heterogeneity and the underlying challenges in social media analytics.

Figure 1 illustrates a high level overview of my research. The core research problems addressed in my research proposal are as follows: (1) To develop a framework to model user-bias in social networks with heterogeneous data; (2) To leverage the user-bias to improve the classification accuracy under domain adaptation settings; (3) Recommend similar actors across social networks; (4) Adapt and leverage information from existing black-box models to improve classification accuracy in new domains; and (5) Explaining task heterogeneity models.

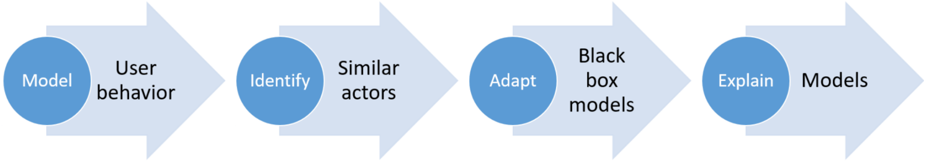

Fig. 1. An Overview of my Research.

2 Major Contributions

User-guided Cross-domain Sentiment Classification[6]: In sentiment analysis when addressing the problem of limited labeled data from the target domain, transfer learning, or domain adaptation, has been successfully applied, which borrows information from a relevant source domain with abundant labeled data to improve the prediction performance in the target domain [7]. The key to transfer learning is how to model the relatedness among different domains. For sentiment analysis, a common practice is to assume similar sentiment polarity for the common keywords shared by different domains. However, existing methods largely overlooked the human factor, i.e., the users who expressed such sentiment [1, 2]. We addressed this problem by explicitly modeling the human factor related to sentiment classification. In particular, we assume that the content generated by the same user across different domains is biased in the same way in terms of the sentiment polarity. In other words, optimistic/pessimistic users demonstrate consistent sentiment patterns, no matter what the context is.

Similar Actor Recommendation[5]: In disease-dedicated social networking sites and online support groups, it is common for the users to stick to one disease-specific social network, although their desired resources are spread over multiple social networks, such as patients with similar questions and concerns. Motivated by this application, we focused on cross-network link recommendation, which aims to identify similar actors across multiple heterogeneous social networks. The problem setting is different from existing work on cross-network link prediction, which either tries to link accounts of the same user from different social networks [4], or aims to match users with complementary expertise or interest [10].

Impact of Social Media on Diabetes self-management: This study aimed to evaluate the use of social media of individuals with diabetes mellitus (DM). Both web-based and in-clinic surveys were collected from individuals with DM. Questions included reasons for visiting the website, and ascertainment of DM reported self-care practices. Descriptive and correlation analyses were employed to evaluate respondents diabetes-specific social networking site usage behaviors.

3 Future Work

Most research in the past focused on semi-supervised learning techniques based on transfer learning and domain adaptation that assumed there exists a relevant source domain with abundant data. With increasing popularity in using machine learning for solving real-world issues, a wide range of machine learning tools were employed to build statistical models for data prediction, forecasting, and analysis. Often these tools are trained on large labeled datasets and with extensive human and computational resources. Later such labeled datasets were sparingly reused to building new models. In our work, we plan to bridge this gap by addressing the problem of domain adaptation using an off-the-shelf tool without source domain data. The objective is to propose a generic framework using manifold learning and gradient boosting for adapting the outputs from an off-the-shelf tool to accommodate the changes in the learning task. It considers two major types of changes, i.e., label deficiency and distribution shift, and aims to maximally boost the performance of the off-the-shelf tool in the target domain, with the help of a limited number of target domain labeled examples.

In the last few years, explainability in machine learning is gaining significant popularity. Decision critical machine learning applications like in health-care call for explaining the decision - Why they do what they do? [8, 3]. In our work, we plan to extend the model-agnostic rule-based approach from [9] to transfer learning settings. The objective of the approach is to identify the set of high-precision rules based on the data features that drive the classification. The process involves identifying the rules in the source domain, rules in the target domain and efficiently learning a bridge between these set of rules that explain the transfer learning model.

References

1. Blitzer, J., Kakade, S., Foster, D.P.: Domain adaptation with coupled subspaces. In: AI Statistics. pp. 173–181 (2011)
2. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: ICML. pp. 513–520 (2011)
3. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. arXiv (2017)
4. Kong, X., Zhang, J., Yu, P.S.: Inferring anchor links across multiple heterogeneous social networks. In: CIKM. pp. 179–188. ACM (27 Oct 2013)
5. Nelakurthi, A.R., He, J.: Finding cut from the same cloth: Cross network link recommendation via joint matrix factorization. In: AAAI. pp. 1467–1473 (2017)
6. Nelakurthi, A.R., Tong, H., Maciejewski, R., Bliss, N., He, J.: User-guided cross-domain sentiment classification. In: Proceedings of the 2017 SIAM International Conference on Data Mining. pp. 471–479. SIAM (2017)
7. Pan, S.J., Yang, Q.: A survey on transfer learning. Knowledge and Data Engineering, IEEE **22**(10), 1345–1359 (2010)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: SIGKDD. pp. 1135–1144. ACM (2016)
9. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations (2018)
10. Tang, J., Wu, S., Sun, J., Su, H.: Cross-domain collaboration recommendation. In: SIGKDD. pp. 1285–1293. SIGKDD, ACM, New York, NY, USA (2012)