Gender Prediction Through Synthetic Resampling of User Profiles Using SeqGANs

Munira Syed, Jermaine Marshall, Aastha Nigam, and Nitesh V. Chawla(✉)

University of Notre Dame, Notre Dame, IN, USA
{msyed2,jmarsha5,anigam,nchawla}@nd.edu

Abstract. Generative Adversarial Networks (GANs) have enabled researchers to achieve groundbreaking results on generating synthetic images. While GANs have been heavily used for generating synthetic image data, there is limited work on using GANs for synthetically resampling the minority class, particularly for text data. In this paper, we utilize Sequential Generative Adversarial Networks (SeqGAN) for creating synthetic user profiles from text data. The text data consists of articles that the users have read that are representative of the minority class. Our goal is to improve the predictive power of supervised learning algorithms for the gender prediction problem, using articles consumed by the user from a large health-based website as our data source. Our study shows that by creating synthetic user profiles for the minority class with SeqGANs and passing in the resampled training data to an XGBoost classifier, we achieve a gain of 2% in AUROC, as well as a 3% gain in both F1-Score and AUPR for gender prediction when compared to SMOTE. This is promising for the use of GANs in the application of text resampling.

Keywords: Gender prediction · Resampling · Adversarial · Topic modeling

1 Introduction

Demographic prediction based on browsing behavior has applications in content recommendation, targeted advertising, and personalized news feeds among others. A few studies on the problem of gender prediction use content-based features (e.g. words, tagged categories, learned topics), and click-based features such as time and sequence of clicks [9,11]. Some studies characterize and predict users' demographic information based on various attributes of their browsing and search behavior [12,15]. For example, Hu et al. [9] utilized web page views and clicks to make predictions on the users' gender and age, Kabbur et al. [11] predicted gender using the content and structure of webpages, and Phuong et al. [18] reported that topic-based features worked better than time and sequential features for gender prediction. To make predictions about individual topic interests,

A. Tagarelli and H. Tong (Eds.): CSoNet 2019, LNCS 11917, pp. 363–370, 2019.
https://doi.org/10.1007/978-3-030-34980-6_38

Nigam et al. [16] observed and collected users' health-seeking behavior, i.e., user demographics, temporal features, and socio-economic community variables. Similarly, in our analysis, we use topic features derived from the text data of the articles read by users. By only using the content, there is a potential to generalize and create user profiles across website platforms and other domains. However, using bag-of-words representation leads to high dimensionality so instead, we use topic modeling to represent the user profiles as topic profiles. In our data set, there is a gender imbalance problem because women tend to search and read more health-based articles online than men. In addition, the preferences and health seeking behavior of females is very different from male users [16]. Furthermore, online article content is generated and expires quickly, so learning article-specific content does not generalize well. In our analysis, we use learned topics as features to mitigate this short-lived nature of articles, with the added benefit of topics being generalizable and transferable to other domains. By concatenating all of the articles a user reads, we can build a user profile. This representation of users would be beneficial because user interests do not change as quickly as the content they consume on a website. While most websites can have varying distributions of demographic representation, it is necessary to understand how content is consumed and interest varies based on the variety of demographic features [16]. Since we want to be able to identify the reading/consumption patterns of these under-represented users accurately, we use resampling techniques that can better represent the minority class. Imbalance can be tackled at the data level through various techniques such as oversampling (data augmentation), where we duplicate some of the minority samples, and undersampling, where we discard some of the majority samples. Undersampling techniques have the drawback of losing potentially valuable data whereas random oversampling may lead to a higher weight for the minority samples [8]. To mitigate the bias from duplicating the minority samples, Synthetic Minority Oversampling Technique (SMOTE) was introduced by Chawla et al. [5] for generating synthetic samples of the minority class. Other SMOTE variants have also been proposed since then [8].

Most of the previously mentioned popular resampling techniques exist for resampling real, continuous data. However, when this is applied to numerical representations of text data, it could lead to the generation of noisy samples. For example, in a bag-of-words representation of text where the text samples are represented by counts of words in the vocabulary, synthetic resampling methods could generate non-integral number of words. Thus, to avoid the percolation of noise from the numerical representation of text, we can resample the minority text data using synthetic text generation techniques. LSTMs and RNNs have been used for generating text in various applications such as generating lyrics [19] and fake reviews [2]. Adversarial methods such as SeqGANs are similar to these techniques in that they use RNNs in their generator for generating text data [21]. Generative adversarial networks have been successfully used for generating synthetic samples of the minority class to augment the training set [14]. Zhu et al. focused on solving a class imbalance problem with GANs in the domain of emotion classification using images with relative success [22]. In the text domain, Anand et al. [1] used text-GANs to generate synthetic URLs for phishing detection. While these techniques exist for synthetic text generation, their

application for the task of resampling minority text data for classification, and specifically the use of GANs to do so, is under-explored. Existing resampling methods for text classification either rely on bag-of-words through term weighting [10] or generating synthetic text data using probabilistic topic models [6]. In fact, Sun et al. [20] systematically explored the effect of popular resampling strategies on tf-idf represented imbalanced text classification problems, and found that in most cases basic SVM performs better without resampling. Our goal therefore, is to explore the use of SeqGANs for generating text data for imbalanced binary classification. For this, we propose a pipeline that represents users with user profiles and topic modeling.

2 Dataset Description

The browsing data used in this paper was generated on a health-based website which collects users' demographic information from their subscribers and receives the browsing activity of these users. The data was collected from user clicks on articles from 2006 through 2015. Over this time frame, data from 263 topics related to health was collected [16]. The content of the URLs accessed by users was crawled from the website and processed by removing stop words. We experimented with a varying number of topics and decided to use 200 topics uniformly in all of our experiments. Since different age groups have varying topic interests [16] we split the dataset by age groups: 18–24 (32% of the data), 25–34 (33.3%), 35–44 (21.6%), 45–54 (14.4%), 55–64 (11.6%), and 65–80 (5.6%). For our experiments, we used the age group of 65–80 because they are at higher risk for health issues. This portion is small enough to avoid scalability issues with SeqGAN. There are 17,499 users in this age group with 13,021 females and 4,478 males (25.59% of users are male). We also discovered a long right tailed distribution with a steadily decreasing number of users as the number of article clicks per user increases.

3 Model

3.1 Steps I and II - User Representation

Users read various articles, which is the input to the model (click level representation, Fig. 1). At the user level, we create user profiles by concatenating all the articles read by the individual to generate a single text document correspond to each user.

3.2 Step III - User Representation Using Topics

We next represent a user in a structured format using topic modeling (topic profile at user level, Fig. 1). A topic model is trained on the corpus of the individual articles accessed by all of the users in the training and testing sets. While many topic modeling techniques such as SVD, LDA and their many variants exist, we use NMF because it is well-suited for the task of topic modeling and relies on

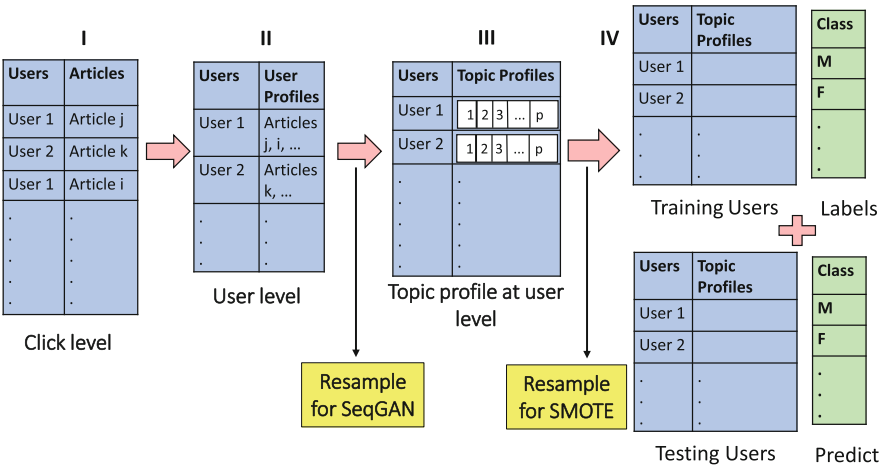

Fig. 1. Classification pipeline

matrix factorization. In our case, the vocabulary of the corpus is huge even after filtering stopwords. Thus, a bag-of-words representation would be infeasible for representing each user. The NMF topic model [17] is trained on all of the individual articles that appear in the corpus. The topic representation for each user is generated by transforming their profiles into the topic space.

Let the corpus D consist of articles $d_1, d_2, ..., d_m$. Each document d_j is represented by a vector of w words. Thus, the document matrix $\mathbf{D}$ has the dimensions $m \times w$. We generate an NMF topic model in the topic space of p dimensions by decomposing matrix $\mathbf{D}$ into factors $\mathbf{W}$ and $\mathbf{H}$. Thus, $\mathbf{D} = \mathbf{WH}$, where $\mathbf{W}$ has the dimensions $m \times p$ and $\mathbf{H}$ has the dimensions $p \times w$. Here $\mathbf{W}$ can be interpreted as representation of documents in the topic space and $\mathbf{H}$ is the representation of topics in the word space. NMF optimizes the objective function

$$\frac{1}{2}||\mathbf{D} - \mathbf{WH}||_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(D_{ij} - (WH)_{ij}\right)^2 \tag{1}$$

where $\mathbf{D} = \mathbf{WH}$ and $\mathbf{W}$ and $\mathbf{H}$ are minimized alternately. A user U_i is a linear combination of all the documents in D. We represent the user by concatenating the articles read by the user. Thus, U_i is given by $\sum_{j=1}^{m}(c_j d_j)$ where c_j is the number of times user i read article d_j. The topic representation $\mathbf{W}'$ of a matrix of n users $\mathbf{U}$ (dimensions $n \times m$) consisting of $U_1, U_2, ..., U_n$, whether training or testing, would be given by $\mathbf{U} = \mathbf{W}'\mathbf{H}$, where $\mathbf{W}'$ is obtained by minimizing the same objective function in Eq. 1 while $\mathbf{H}$ is kept constant.

3.3 Step IV - Split into Training and Testing Sets

The topic representation of the s users in the training set W'_{train} is the input to train a classifier, and those of the $n - s$ users in the testing set given by W'_{test} are used to predict the gender of the users as the output.

3.4 Resampling

In the case of the minority class, we use two resampling approaches for improving the performance of the classifier (i.e. SMOTE-based and text-based).

The data can be resampled after Step III in which topic profiles of the users are generated. This is where we will apply SMOTE and its variants. However, the text-based resampling would occur after Step II by generating synthetic user profiles of the minority class from the text of real users of the minority class. The intuition is that by applying resampling at an earlier stage, we avoid biases introduced through the conversion of text to numerical representation. Resampling at this stage can be done by Random Oversampling of the minority texts (ROS), Random Undersampling of the majority texts (RUS), and SeqGAN to reduce the imbalance between the classes.

SeqGAN. We formulate the sequence generation problem for gender classification as shown below to produce a sequence of tokens $X_{1:T} = (x_1, x_2, ..., x_T)$, $x_T \in Y$ where Y is the vocabulary of the set of candidate tokens. We train a Discriminator model D in order to guide a Generator model G. The discriminator's goal is to predict how likely a sequence X_{1T} is to be from the real sequence information. G is then updated by a policy gradient from the expected reward received from D. The formulation for the policy gradient is shown in Eq. 2 below:

$$J(\Theta) = E[R_T|s_0] = \sum_{y_1 \in Y} G_\Theta(y_1|s_0) \times Q_{D_\Theta}^{G_\Theta}(s_0, y_1) \tag{2}$$

"where R_T is the reward for a complete sequence and $Q_{D_\Theta}^{G_\Theta}(s, a)$ is the expected cumulative reward starting from state s taking an action a following policy G_Θ" [21].

4 Experiments

We compare the resampling methods in our gender prediction task. Specifically, we report on two sets of experiments to compare the performance of SMOTE-based and SeqGAN text-based methods for resampling. We use 5-Fold cross-validation for evaluation with AUROC, AUPR, and F1-Score as the performance metrics. The 65–80 age group has an imbalance of approximately 25% male.

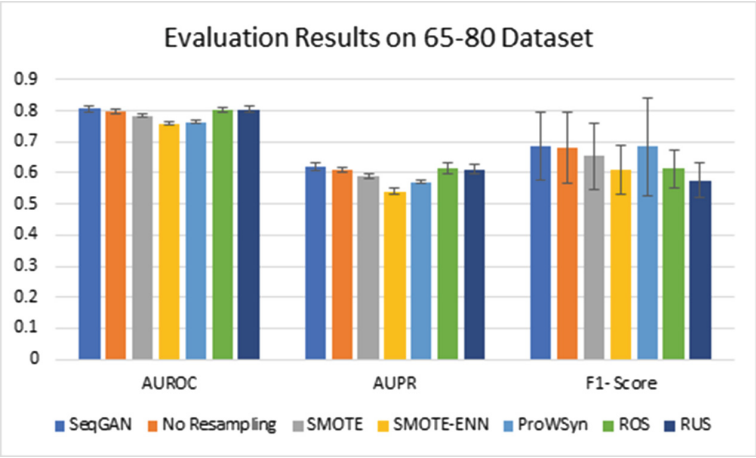

Fig. 2. Evaluation results of different resampling methods 65–80 dataset

4.1 Experiment 1: SMOTE-Based Resampling

We evaluate the capability of some SMOTE-based resampling techniques. SMOTE Edited Nearest Neighbor Rule (SMOTE-ENN) handles class imbalance by removing samples from both the majority and minority class [4]. SMOTE-Out considers the nearest majority and minority example to create synthetic data [13], and ProWSyn generates weights for the minority samples based on boundary distance [3]. We generated synthetic samples so as to balance the two classes.

4.2 Experiment 2: Text-Based Resampling

From each male user profile, we sampled 20 words with high TF-IDF values to represent the individual male user as input to the SeqGAN. Using SeqGAN, we generated 500 sequences of 20 words each which is the same sequence length used in [21]. Thus, we generated 500 synthetic male profiles for each fold of the 5-fold cross-validation. We used the implementation of SeqGAN with a CNN in the discriminator network and an LSTM in the generator network.[1]

We utilized XGBoost (with parameters set to a learning rate of 1, estimators of 9, a max depth of 5, subsample of 0.99, min-child-weight of 5, scale-pos-weight of 3, seed of 3, and gamma of 3) [7] after testing multiple configurations. We used XGBoost instead of a neural network such as DNN for this problem because XGBoost performed well and is a powerful ML algorithm also used in many papers and competitions. In the health domain, interpretability of models is vital to their practical usage and so we use XGBoost instead of neural networks which are not as easily interpreteble. Though there has been recent work

[1] https://github.com/bhushan23/Transformer-SeqGAN-PyTorch/blob/master/seq_gan/.

on explainable neural networks, that is beyond the scope of this paper. We compare SeqGAN against baselines with no resampling, resampling with SMOTE, SMOTE-Out, ProWSyn, and SMOTE-ENN as shown in Fig. 2. WE did not find significant differences when parameters were varied for the SMOTE-based baselines. In Fig. 2, we see that SeqGAN does not suffer from the sub-class problem and outperforms SMOTE and SMOTE-ENN in terms of AUROC, AUPR, and F1-Score. Text-based resampling methods of ROS and RUS perform very similarly to SeqGAN. XGBoost without resampling is second only to the text-based resampling methods. However, we expected this as XGBoost has a parameter known as 'scale-pos-weight' that varies the ratio of positive and negative examples. This allows the algorithm to better control for imbalance than many classic supervised learning algorithms.

5 Conclusion

We report on a new application for SeqGANs which synthetically resamples minority text data to improve imbalanced classification. The experimental results show that SeqGAN outperforms SMOTE-based resampling techniques when combined with the predictive power of XGBoost. In the future, we will explore other resampling techniques through the use of GANs, better text-summarization strategies to reduce the length of the input to SeqGAN, and more efficient methods of using higher sequence lengths with SeqGAN.

Acknowledgements. We thank Trenton Ford for helpful discussions. This work was supported in part by the National Science Foundation (NSF) Grant IIS-1447795.

References

1. Anand, A., Gorde, K., Moniz, J.R.A., Park, N., Chakraborty, T., Chu, B.T.: Phishing URL detection with oversampling based on text generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 1168–1177. IEEE (2018)
2. Bartoli, A., De Lorenzo, A., Medvet, E., Tarlao, F.: Your paper has been accepted, rejected, or whatever: automatic generation of scientific paper reviews. In: Buccafurri, F., Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-ARES 2016. LNCS, vol. 9817, pp. 19–28. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45507-5_2
3. Barua, S., Islam, M.M., Murase, K.: ProWSyn: proximity weighted synthetic oversampling technique for imbalanced data set learning. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) PAKDD 2013. LNCS (LNAI), vol. 7819, pp. 317–328. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37456-2_27
4. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. **6**, 20–29 (2004)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)

6. Chen, E., Lin, Y., Xiong, H., Luo, Q., Ma, H.: Exploiting probabilistic topic models to improve text categorization under class imbalance. Inf. Process. Manage. **47**(2), 202–214 (2011)
7. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: KDD (2016)
8. Fernández, A.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61**, 863–905 (2018)
9. Hu, J., Zeng, H.J., Li, H., Niu, C., Chen, Z.: Demographic prediction based on user's browsing behavior. In: WWW (2007)
10. Jindal, R., Malhotra, R., Jain, A.: Techniques for text classification: literature review and current trends. Webology **12**(2), 1–28 (2015)
11. Kabbur, S., Han, E.H., Karypis, G.: Content-based methods for predicting web-site demographic attributes. In: 2010 IEEE International Conference on Data Mining, pp. 863–868 (2010)
12. Kim, D.Y., Lehto, X.Y., Morrison, A.M.: Gender differences in online travel information search: implications for marketing communications on the internet. Tourism Manage. **28**(2), 423–433 (2007)
13. Koto, F.: SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE: an ehancement strategy to handle imbalance in data level. In: 2014 International Conference on Advanced Computer Science and Information System, pp. 280–284 (2014)
14. Lee, S.K., Hong, S.J., Yang, S.I.: Oversampling for imbalanced data classification using adversarial network. In: 2018 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1255–1257. IEEE (2018)
15. McMahan, C., Hovland, R., McMillan, S.: Online marketing communications: exploring online consumer behavior by examining gender differences and interactivity within internet advertising. J. Interact. Advertising **10**(1), 61–76 (2009)
16. Nigam, A., Johnson, R.A., Wang, D., Chawla, N.V.: Characterizing online health and wellness information consumption: a study. Inf. Fusion **46**, 33–43 (2019)
17. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
18. Phuong, T.M., et al.: Gender prediction using browsing history. In: Huynh, V., Denoeux, T., Tran, D., Le, A., Pham, S. (eds.) Knowledge and Systems Engineering, pp. 271–283. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-02741-8_24
19. Potash, P., Romanov, A., Rumshisky, A.: Ghostwriter: using an LSTM for automatic rap lyric generation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1919–1924 (2015)
20. Sun, A., Lim, E.P., Liu, Y.: On strategies for imbalanced text classification using SVM: a comparative study. Decis. Support Syst. **48**(1), 191–201 (2009)
21. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: sequence generative adversarial nets with policy gradient. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
22. Zhu, X., Liu, Y., Qin, Z., Li, J.: Data Augmentation in Emotion Classification Using Generative Adversarial Networks. ArXiv abs/1711.00648 (2017)