ARTIFICIAL INTELLIGENCE STUDIES
YAPAY ZEKA ÇALIŞMALARI
2018, 1(1):26-29

www.aistudies.org

Sahte Twitter Hesaplarının Yapay Sinir Ağları ile Tespiti

Mehmet ŞİMŞEK[*,a], Oğuzhan YILMAZ [,a], Asena Hazal KAHRİMAN [,a], Levent SABAH [,b]

[a,*] *Düzce Üniversitesi Bilgisayar Mühendisliği Bölümü, DÜZCE 81620, TÜRKİYE*
[b.] *Düzce Üniversitesi Bilgi İşlem Daire Başkanlığı, Rektörlük, DÜZCE 81620, TÜRKİYE*

MAKALE BİLGİSİ

Alınma: 26.04.2018
Kabul:15.06.2018

Anahtar Kelimeler:
Online Sosyal Ağlar
Twitter
Sahte Hesap
Yapay Sinir Ağları
Sınıflandırma

Sorumlu Yazar:
e-posta:
mehmetsimsek@duzce.edu.tr

ÖZET

Online Sosyal Ağlar (OSA), bilgi paylaşımı, haber takibi, ürün tanımıtı gibi amaçlar için oldukça elverişli ortamlardır ve bu ağlar insanlar tarafından yaygın olarak kullanılmaktadırlar. OSA'ların bu avantajlarına rağmen, bir OSA'daki bir hesabın gerçek bir kişiye ya da kuruluşa ait olduğunu anlamak zordur. Oluşturulan sahte hesaplar üzerinden istenmeyen içerikler ağ üzerinde yayılabilir. Bu nedenle, sahte hesapların tespiti önemli bir problemdir. Bu çalışmada, bir Yapay Sinir Ağı (YSA) sınıflandırıcısı bu probleme uygulanmış ve farklı aktivasyon fonksiyonları için deneysel sonuçlar değerlendirilmiştir. Deneysel sonuçlara göre, YSA sınıflandırıcısı sahte hesap sınıflandırmada başarılı sonuçlar vermiştir. Farklı aktivasyon fonksiyonlarının YSA'nın farklı katmanlarında kullanımı, sonuçları anlamlı biçimde etkilemektedir. Literatürde, diğer sınıflandırma yöntemleri, OSA'larda sahte hesap ve spam içerik yayan hesapların tespitinden yagın olarak kullanılmıştır. Bildiğimiz kadarıyla, yapay sinir ağlarını farklı aktivasyon fonksiyonları ile sahte hesap tespiti probleminde bu kadar detaylı kullanan bir çalışma bulunmamaktadır.

DOI: 10.30855/AIS.2018.01.01.03

Detecting Fake Twitter Accounts with using Artificial Neural Networks

ARTICLE INFO

Received:
26.04.2018
Accepted:
15.06.2018

Keywords:
Online Social Networks
Twitter
Fake Account
Artificial Neural Networks
Classification

Corresponding Authors
e-mail:
mehmetsimsek@duzce.edu.tr

ABSTRACT

Online Social Networks (OSNs) are great environments for sharing ideas, following news, advertising products etc., and they have been widely using by people. Although these advantages of OSNs, it is difficult to understand whether an account in OSNs really belongs to a person or organization. Through created fake accounts, unwanted content can spread over the social network. Therefore, the identification of fake accounts is an important problem. In this study, we applied Artificial Neural Network (ANN) classifier to this problem and we evaluated performances of different activation functions. According to the experimental results, use of artificial neural networks in detecting fake accounts yielded successful results. The use of various activation functions in different layers on the ANN significantly affects the results. In the literature, other classification methods have been widely used for detecting fake accounts and spammers on OSNs. To the best of our knowledge, there is no detailed study that classifies fake accounts using ANNs with different activation functions.

DOI: 10.30855/AIS.2018. 01.01.03

1. INTRODUCTION *(GİRİŞ)*

On the Online Social Networks (OSN), people share the ideas; follow news; influence each other's. Essentially, OSNs are reflections of real social networks. However, OSNs have a problem which real social networks don't have it: fake accounts. Fake accounts are great problem for OSNs. Fake accounts can spread false news; can manipulate the follower numbers of a person; can send spam contents to many users. All this is an obstacle to the main functioning of OSNs. It is therefore important to identify fake accounts in an OSN. Although there are many OSNs, Twitter is the most prominent one. Twitter is distinguished from other OSNs, because of the number of users; frequency of use; the use of a lot of important people and organizations etc. For this reason, many studies in the literature focused on identifying fake accounts on Twitter. The features of an account can provide information about its authenticity or its fraud. By the same way, the tweets of and account or its relations an provide information about its authenticity or its fraud. For this reason, fake account detection are categorized as follows: Detecting with using account based features, detecting with using tweet based features, and detecting with relationship between users [1].

Main approach in the literature is that classifying users or tweets according to their features with using classification methods [2]–[5]. Lee et. al. have constructed a honeypot for collecting information about fake accounts' interactions [6]. They have collected the average tweets per day, the ratio of the number of following and followers etc. Then, they have applied some machine learning classification techniques for classifying users. Similarly, Lin and Huang have used the ratio of user's tweets that contain URLs, and user's interactions as features; and they have classified users with using Decision Tree as spammer or non-spammer [7]. Some of the studies have analyzed the content of tweets to classify them. [8]–[10] have analyzed the URLs which are posted in tweets and have made a classification according to whether the content is malicious or not.

Besides, there are hybrid approaches in the literature [11]–[14]. Hybrid approaches aim to classify accounts based on account based features, tweet based features, and graph based features. These studies such as the others have classified the accounts as spammer or non-spammer by using the classification algorithms and the features, too. According to the literature review, Artificial Neural Networks (ANNs) has not been used commonly. For this purpose, we applied an ANN classifier to this problem and we evaluated performances of different activation functions. Also, we only used account based features because of their lightweight nature and that they allow to real-time detection [1], [15].

The rest of paper is organized as follow: Section 2 gives the materials and methods. The results are given in Section 3. Section 4 concludes the paper and discusses the results.

2. MATERIAL and METHODS *(MATERYAL VE YÖNTEMLER)*

In this section we only gave the some account based features. For further and deeply information about other types of features, [1] and [16] should be examined. The dataset used in this study has 100.000 account's features and their classes as spammer and non-spammer [17]. In the dataset, 95.000 accounts are non-spammers and 5.000 accounts are spammers. We used the following 10 numeric features from dataset: Account age, Number of following, Number of follower, Number of user favorites, Number of lists, Number of tweets, Number of retweets, Number of hashtags, Number of user mention, and URL.

We performed account classification with using an ANN. We used four different activation functions and compared the efficiencies of different combinations of these activation functions. These activation functions are Softmax Function, Sigmoid Function, Rectifier (ReLu), and Hyperbolic Tangent (Tanh). The constructed ANN has 1 input layer with 10 neurons, 1 hidden layer with 10 neurons, and 1 output layer with 1 neuron. Briefly, it may be useful to summarize the activation functions used in this study.

Softmax function calculates probability distributions of an event over other events. This function is used in various multiclass classifications. The main advantage of Softmax is the output probabilities range. The range changes 0 to 1. Softmax, returns the probabilities of each class on multi-class classification problems, and target class have the highest probability.

Sigmoid function takes a real number and returns a value between 0 and 1 as result. The Sigmoid function is used for binary classification in logistic regression.

Hyperbolic Tangent function is sigmoidal as Sigmoid function, but it gives values between -1, 1 as output.

Rectifying activation function was first introduced by Hahnloser et al [18]. This activation function gives better results for training deeper neural networks. Compared to sigmoid or similar activation functions, it allows for faster and effective training of deep neural networks on large and complex datasets [19].

3. EXPERIMENTAL RESULTS and DISCUSSION *(DENEYSEL SONUÇLAR ve TARTIŞMA)*

When we trained our artificial neural network twice, we notice the second time we obtained a lower accuracy both on the training set and the test set than the first time. The reason of this is The Bias-Variance Tradeoff. The Bias-Variance Tradeoff is the fact that we are trying to train the model that will not only accurate but also that should not have too much variance of accuracy, when we trained several times. For avoiding from the variance problem we used k-Fold Cross Validation.

We performed 2 groups of experiments. In the first group, we used same activation function at all layers (input layer, hidden layer and output layer). Table 1 shows the performances of all activation functions in the first group of experiments.

Table 1. Performances of Softmax, Sigmoid, Rectifier (ReLu), and Hyperbolic Tangent (Tanh) activation functions on the problem

Activation Function	**Mean**	**Variance**
Softmax	0,0498	-
Sigmoid	0,9727	0.0015
Rectifier	0,9550	0,0104
Tanh	0,9120	0,1145

According to the first experiments, Sigmoid function has given the best results. Hyperbolic Tangent's results not as good as the results of sigmoid function, the variance is too high and the mean of accuracies are too low. Rectifier Function's results are better than Hyperbolic Tangent function's results but still not good as Sigmoid Function's results. Softmax Function has the worst results because Softmax function also has been used for the output layer. As an output layer Softmax function is not suitable. Variance value has not been given for Softmax experiment because of it was very low.

In the second group of experiments, we used different activation functions at different layers. Table 2 shows the performances of all activation functions in the first group of experiments. We named the experiments with using the layers which the activation functions has been implemented in.

Table 2. Performances of the different activation functions at different layers

Activation Functions (input layer, hidden layer, output layer	**Mean**	**Variance**
Softmax-Softmax-Sigmoid	0,9772	0,0015
Tanh-Tanh-Sigmoid	0,9768	0,0039
Rectifier- Rectifier- Sigmoid	0,9721	0,0023
Rectifier- Rectifier-Tanh	0,8593	0,1846
Tanh-Tanh-Rectifier	0,9594	0,0113
Softmax-Softmax-Tanh	0,9524	0,0091
Softmax-Softmax-Rectifier	0,9601	0,0124

According to the second experiments, Softmax-Softmax-Sigmoid has given the best results. The performances of Tanh-Tanh-Sigmoid and Rectifier- Rectifier- Sigmoid are closer to the performance of Softmax-Softmax-Sigmoid.

4. CONCLUSION *(SONUÇ)*

In this study, we dealt with fake account detection problem on Twitter with using artificial neural networks, and we have done comprehensive experiments with different activation functions and their combinations. According to the experimental results, the use of artificial neural networks in detecting fake accounts yielded successful results. The use of different activation functions in different layers significantly affects the results. In addition, the training times of artificial neural networks were short. This shows that artificial neural networks can be used in detecting fake accounts using fast-changing tweet-based and graph-based features.

In the literature, other classification methods have been widely used for detecting fake accounts and spammers on OSNs. To the best of our knowledge, there is no detailed study that classifies fake accounts using artificial neural networks with different activation functions. With the expansion and development of the concept of deep learning, the use of artificial neural networks will become more widespread.

REFERENCES *(KAYNAKLAR)*

[1] A. Talha and R. Kara, “A Survey of Spam Detection Methods on Twitter,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, 2017.

[2] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “Fame for sale: Efficient detection of fake Twitter followers,” *Decis. Support Syst.*, vol. 80, pp. 56–71, Dec. 2015.

[3] T. Wu, S. Wen, Y. Xiang, and W. Zhou, “Twitter spam detection: Survey of new approaches and comparative study,” *Comput. Secur.*, vol. 76, pp. 265–284, Jul. 2018.

[4] D. Ramalingam and V. Chinnaiah, “Fake profile detection techniques in large-scale online social networks: A comprehensive review,” *Comput. Electr. Eng.*, vol. 65, pp. 165–177, Jan. 2018.

[5] P. V. Bindu, R. Mishra, and P. S. Thilagam, “Discovering spammer communities in twitter,” *J. Intell. Inf. Syst.*, Jan. 2018.

[6] K. Lee, J. Caverlee, and S. Webb, “Uncovering social spammers,” in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’10*, 2010, p. 435.

[7] P.-C. Lin and P.-M. Huang, “A Study of Effective Features for Detecting Long-surviving Twitter Spam Accounts,” in *2013 15th International Conference on Advanced Communications Technology (ICACT)*, 2013, pp. 841–846.

[8] D. K. McGrath and M. Gupta, “Behind phishing: an examination of phisher modi operandi,” in *Usenix Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2008, p. 4.

[9] S. Lee and J. Kim, “WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream,” *IEEE Trans. Dependable Secur. Comput.*, vol. 10, no. 3, pp. 183–195, May 2013.

[10] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09*, 2009, p. 1245.

[11] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC ’10*, 2010, p. 1.

[12] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, “Who is tweeting on Twitter,” in *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC ’10*, 2010, p. 21.

[13] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, “6 million spam tweets: A large ground truth for timely Twitter spam detection,” in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 7065–7070.

[14] A. A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, and C. Yang, “CATS: Characterizing automation of Twitter spammers,” in *2013 Fifth International Conference on Communication Systems and Networks (COMSNETS)*, 2013, pp. 1–10.

[15] P. Kaur, A. Singhal, and J. Kaur, “Spam detection on Twitter: A survey,” in *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016.

[16] J.-V. Cossu, V. Labatut, and N. Dugué, “A review of features for the discrimination of twitter users: application to the prediction of offline influence,” *Soc. Netw. Anal. Min.*, vol. 6, no. 1, p. 25, Dec. 2016.

[17] “Twitter Spammer Dataset.” [Online]. Available: https://raw.githubusercontent.com/YIHE1992/Convolutional-Neural-Network/master/95k-continuous.csv. [Accessed: 01-Mar-2018].

[18] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, “Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit,” *Nature*, vol. 405, no. 6789, pp. 947–951, Jun. 2000.

[19] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” *AISTATS ’11 Proc. 14th Int. Conf. Artif. Intell. Stat.*, vol. 15, pp. 315–323, 2011.