

NAİVE BAYES TƏSNİFAT MODELİNƏ ƏSASLANAN SPAM E-POÇTLARIN TƏSNİFATI

Aytac Nizamova

Azərbaycan Texniki Universiteti, Bakı

nizamovaaytac@gmail.com

Müasir dövrdə spam e-poçt adlanan arzuolunmaz e-poçtlar cəmiyyət üçün böyük problemə çevrilir. Həm şirkətlər, həm də şəxslər bu spam e-poçtların səbəb olduğu vaxt və pul itkisindən əziyyət çəkirlər. Bir hesabatda qeyd olunur ki, spam e-poçt təhlükəsi artmaqda davam edir və spam e-poçtların miqdarı bütün e-poçtların 75%-dən çoxunu təşkil edir. Spam süzgəcində maşın öyrənməsinin ən çox istifadə edilən və qabaqcıl metodu məzmunə əsaslanan filtrləmə üsulu olan *Naive Bayes* Təsnifatıdır. Bu üsul adətən e-poçtla-

rın məzmununda sözləri, söz və ifadələrin ardıcılığını və paylanması təhlil edir və sonra daxil olan e-poçt spamlarını filtrləmək üçün yaradılan qaydalardan istifadə edir [1]. Maşınöyrənmə alqoritmləri e-poçt məlumatlarında xüsusi qaydaları öyrənmə və təsnifat yaratmaq üçün onlardan istifadə edə bilər [2]. Bu alqoritmlərə əlavə olaraq, stoxastik optimallaşdırma üsulları da spamların süzülməsində iştirak edir. Onlar mürəkkəb riyazi hesablamalara malik olmayan məsələlərin optimal həll yollarını tanıya bilərlər [3]. Bu məqalə maşınöyrənmə və süni intellektin təkmilləşdirilməsinə əsaslanan spam filtrasiyasının bir fərziyyəsini nümayiş etdirir.

Metodologiya

Spamları fərqləndirmək üçün həmişə spam e-poçtlarını ehtiva edən böyük verilənlər bazası lazımdır. Dəqiqlik spam filtri üçün yeganə performans deyil. Spam filtrinən səhv təsnif edilməməsi də vacibdir. Bir spam e-poçt səhv olaraq normal e-poçt olan ham e-poçta təsnif edildikdə, insanlar sil düyməsini klikləməklə bu problemi asanlıqla həll edə bilərlər. Buna görə də, maşınöyrənmə alqoritminin fəaliyyətini yalnız onun dəqiqliyi ilə qiymətləndirmək qeyri-adekvatdır. Dada klasterləşdirmə *Naive Bayes* Təsnifatlayıcıları [5] kimi spam filtrasiyası üçün istifadə olunan ümumi metodları topladı və müqayisə etdi. Spam filtrasiyası üçün istifadə edilən iki qruplaşma üsulu sıxlığa əsaslanan və K-ən yaxın qonşulardır (KNN). K-ən yaxın qonşuların paylanması spamın təsnifatı üçün çox vacibdir.

Məhdudiyyətlər və Perspektivlər

Aydın ki, spam təsnifatının dəqiqliyi 100%-ə çatmayıb. Bəzi normal e-poçtlar spam kimi təsnif edilir, bəzi spam e-poçtlar isə normal e-poçtlar kimi fərqlənir. Qeyd edildiyi kimi, normal e-poçtları spam e-poçtlara qoymaq böyük səhvdir, çünki istifadəçilər e-poçtları gözlədikləri kimi qəbul etməyəcəklər. *Naive Bayes* Təsnifatı faydalıdır, lakin bəzi məhdudiyyətlərə malikdir. Kəşf göstərir ki, *Naive Bayes* Təsnifatının performansı spamları süzməkdə 100% düzgünlüyə nail ola bilməz. Buna görə də, süni intellektin inkişafı ilə birləşmiş bəzi qabaqcıl alqoritm *Naive Bayes* Təsnifat alqoritmə ortaya çıxacaq. Süni intellektin bir aspekti neyron şəbəkələridir. Neyron şəbəkəsi əksər maşınöyrənmə problemi üçün işləyən alqoritmədir. Neyroşəbəkələrin və *Naive Bayes* Təsnifatlayıcısının birləşməsi ilə dəqiqlik demək olar ki, 100%-ə çatacaq ki, bu da Paul Grahamın "Spam üçün plan" [6] əsərində qeyd olunur. Bu bir-biri ilə yaxşı əlaqəli neyronlar spam e-poçtları ayırmaq kimi müəyyən problemi həll etmək üçün birlikdə işləyəcəklər. *Naive Bayes* Təsnifatı və neyron şəbəkələrinin birləşməsi əksər e-poçt filtrlərində işləyir.

Nəticə

Xülasə, spam e-poçtlar global təhlükəsizlik və iqtisadiyyatın virusuna çevrilib. Poçt şirkətləri spamları təsnif etmək üçün müxtəlif faydalı və qabaqcıl üsullar təklif etsələr də, istifadəçilər hələ də dünyadakı çoxlu sayda spamın təsirinə məruz qalırlar. Bu yazıda Google, Yahoo və Outlook.com kimi poçt şirkətlərinin indiki spam filtrasiyası ilə müqayisə edilərək, *Naive Bayes* Təsnifatına və klasterləşdirmə alqoritmə əsaslanan bir alqoritm təqdim etdi. Bu alqoritm istifadəçilərə həm spam e-poçtlar, həm də normal e-poçtlar üçün öz lüğətlərini dəyişdirməyə və spam e-poçtları 90%-dən çox dəqiqliklə müəyyən etməyi imkan verə bilər. Bununla belə, bu araşdırmada sınaqdan keçirilmiş e-poçtlar nisbətən kiçikdir və dəqiqlik ideal məqsədə çatıb. Hələ təkmilləşdirilməli olan məqamlar var. Ümumilikdə, bu nəticələr spamı filtrləmək üçün təlimat təklif edir.

Ədəbiyyat

- [1] Kristina, V., Karpaqavalli, S., Suqanya, G. Nəzarət olunan maşın öyrənməsindən istifadə edərək e-poçt spam filtrasiyası texnikalar. Kompüter Elmləri və Mühəndisliyi üzrə Beynəlxalq Jurnal (JCSE) 2.09, 3126-3129 (2010).
- [2] Guzella, T.S., Walmer M. C. Spam filtrasiyası üçün maşın öyrənmə yanaşmalarının nəzərdən keçirilməsi. Tətbiqləri olan Ekspert Sistemləri 36.7, 10206-10222 (2009),
- [3] Qasem, S.N., Siti, M.S., Azlan, M.Z. Radial əsas funksiyası neyron şəbəkəsinin dizaynı üçün çoxməqsədli hibrid təkamül alqoritmləri. Biliyə əsaslanan sistemlər 27, 475-497 (2012).
- [4] Stuart, I., Cha S.H., Charles T. Lazımsız e-poçtlar üçün neyron şəbəkə təsnifatı. mövzusunda Beynəlxalq Seminar Sənədlərin Təhlili Sistemləri. Springer, Berlin, Heidelberg, 2004.
- [5] Whissell, J.S., Charles L.A.C. Yarı nəzarət edilən spam filtrasiyası üçün klasterləşdirmə. 8-ci illik Əməkdaşlıq, Elektron mesajlaşma, Sui-istifadə və Spamlar Mübarizə Konfransının materialları. 2011.
- [6] Graham, Paul. Multimodal optimallaşdırma üçün Firefly alqoritmləri. Stokastik alqoritmələr üzrə beynəlxalq simpozium. Springer, Berlin, Heidelberg, 2009.