TAYANCH VEKTORLAR USULI YORDAMIDA SPAM HUJJATLARNI ANIQLASH ALGORITMI

Xamdamov Rustam Xamdamovich
Raqamli texnologiyalar va sun'iy intellektni rivojlantirish ilmiy-tadqiqot instituti
elshodhaydarov1881@gmail.com

Ibrohimov Aziz Ravshanbek o'g'li
"Kiberxavfsizlik markazi" DUK, Axborot tizimlari va rusurslari bo'limi
aribrohimov@gmail.com

Haydarov Elshod Dilshod o'g'li
Muhammad al-Xorazmiy nomidagi TATU
elshodhaydarov1881@gmail.com

ANNOTATSIYA

Ushbu maqolada elektron pochtadagi hujjatlarni tayanch vektorlar usuli asosida filtrlash hamda spam xabarlarni aniqlash algoritmi va dasturi ishlab chiqilgan. Algoritmni ishlab chiqishda hujjatlarni sinflashtirish masalasi yechilgan. Hujjatlarni toifalarga ajratishda tayanch vektorlar usuli qo'llanilgan. Bunda optimallashtirish masalasini yechishda moslanuvchan tasodifiy qidiruv usulidan foydalanilgan. Ishlab chiqilgan algoritm yordamida dastur yaratilgan hamda algoritmni ishlash aniqligi natijalari keltirilgan, shuningdek, boshqa usullar bilan solishtirish natijalari ham keltirilgan.

Kalit so'zlar: spam, tayanch vektorlar usuli (SVM), moslanuvchan tasodifiy qidiruv usuli, vazn koeffitsiyentlari, gipertekislik.

ALGORITHM FOR DETECTING SPAM DOCUMENTS USING THE METHOD OF BASE VECTORS

ABSTRACT

This article has developed an algorithm and program for filtering e-mail documents based on the method of base vectors and identifying spam messages. When developing the algorithm, the issue of document classification is solved. When categorizing documents, the method of base vectors was used. This used an adaptive random search method to solve the optimization issue. Using the developed algorithm, a program is created and the results of the accuracy of the algorithm's operation are presented, as well as the results of comparison with other methods.

Keywords: spam, base vectors method (SVM), configurable random search method, weight coefficients, hypertexity.

KIRISH

Bugungi kunda elektron pochta xabarlarini filtrlash natijasida spam xabarlarni aniqlashda qo‘llanib kelinayotgan klassik usullar yoki mashinali o‘qitishga asoslangan usullarning afzalliklari bo‘lganligi bilan bir qatorda kamchiliklar ham mavjud. Yangi ishlab chiqiladigan va amaliyotga joriy qilinadigan usullar mavjud kamchiliklarni qisman darajada ham bartaraf etsa, bu elektron pochta xabarlarini filtrlash tizimning samaradorligini oshiradi.

Elektron pochta xabarlarini spam xabarlardan himoyalash uchun tashkilotning elektron axborot aylanish tizimidagi axborotni sinflashtirish va undagi hujjatlarni spam yoki spam emasligini aniqlashtirish zarur. Tayanch vektorlar asosida matnlarni sinflashtirish usulida foydalanuvchi hujjatlar orasidan spam deb hisoblaganlarini belgilaydi, ya’ni o‘qitish to‘plamini quradi. Bu axborot asosida chiziqli tengsizliklar tizimi ko‘rinishida ko‘rinuvchi qobiqlar hosil qilinadi. Birinchi qadamda tizim spam kabi aniqlanuvchi hujjat-nuqtalari to‘plamini beradi. Ikkinchi qadamda tizim – spam kabi aniqlanmagan hujjat-nuqtalari to‘plamini quradi.

ADABIYOTLAR TAHLILI VA METODOLOGIYASI

Tasniflash muammosini yechish uchun tayanch vektorlar usulini (*Support Vector Machine - SVM*) qo‘llash mumkin. Algoritmning asosiy g‘oyasi, uning vazn koeffitsiyentlari(noma’lum parametrlari)ni aniqlashdan iborat va so‘ngra algoritmning oddiy amalga oshirilish jarayoni taqdim etiladi.

Masalani yechimi sifatida, tushuntirishlar sodda bo‘lishi uchun oddiy xoldagi, ikkilik (faqat ikkita sinf mavjud bo‘lganda) tasniflash muammosini hal qilish ko‘rib chiqiladi. Birinchidan, algoritm o‘quv tanlanmasidagi obyektlar bo‘yicha o‘qitiladi, ular uchun sinf belgilari oldindan ma’lum. Bundan tashqari, oldindan o‘qitilgan algoritm nazorat tanlanmasidan har bir obyekt uchun sinf yorlig‘ini bashorat qiladi. Sinf yorliqlari *Y = {-1, +1}* qiymatlarini olishi mumkin. Obyekt - R^n o‘lchovli fazoda n ta xususiyatga ega $x = (x_1, x_2, \dots, x_n)$ vektor yordamida ifodalanadi. O‘rganish davomida algoritm R^n fazodan obyektni olib, y sinf belgisini hosil qiluvchi x argumentini qabul qiladigan *F(x)=y* funksiyasini qurishi kerak.

Tasniflash vazifasi o‘quv tanlanmasi bilan o‘qitishga taalluqli. SVM – o‘quv tanlanmasi bilan o‘qitish algoritmi hisoblanadi.

Klassifikator sifatida SVMning asosiy maqsadi ikkala sinfni optimal tarzda R^n fazoda ajratuvchi gipertekislikning

$$\omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n + \omega_0 = 0 \qquad (1)$$

tenglamasini topish hisoblanadi. *x* obyektining *F* funksiyasini *Y* sinf yorlig‘iga aylantirishning umumiy ko‘rinishi:

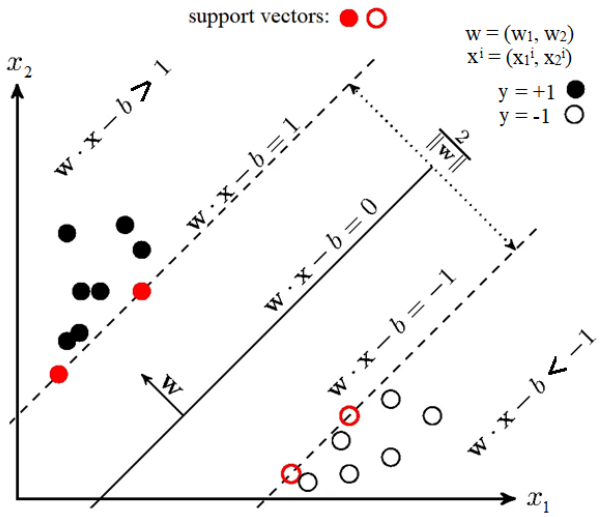
$$F(x) = sign(\omega^T x - b). \quad (2)$$

Bu yerda

$$\omega = (\omega_1, \omega_2, \ldots, \omega_n), \quad b = -\omega_0 \quad (3)$$

belgilashlar kiritilgan. ω va b o‘qitish algoritm vazn koeffitsiyentlari(noma’lum parametrlari) sozlangandan so‘ng, qurilgan gipertekislikning bir tomoniga tushgan barcha obyektlar birinchi sinf, ikkinchi tomoniga tushgan obyektlar esa ikkinchi sinf sifatida belgilanadi.

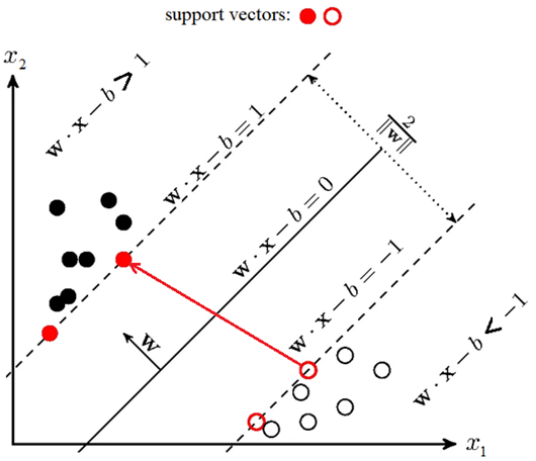
sign() funksiyasi ichida algoritm vazn koeffitsiyentlari(noma’lum parametrlari) bilan obyekt xususiyatlarining chiziqli birikmasi mavjud, Shuning uchun SVM chiziqli algoritmlarga taalluqli hisoblanadi. Ajratuvchi gipertekislikni ko‘p usullar bilan qurish mumkin, lekin SVMda ω va b vazn parametrlari shunday sozlanadiki, sinf obyektlari ajratuvchi gipertekislikdan imkon qadar uzoqroqda yotadi. Boshqacha qilib aytganda, algoritm gipertekislik va unga eng yaqin bo‘lgan sinf obyektlari orasidagi bo‘shliqni maksimal darajada oshirishga yo‘naltirilgan. Bunday obyektlar tayanch vektorlar deb ataladi (1-rasm). Algoritmning nomi shundan kelib chiqqan.



1-rasm. Tayanch vektorlar usulining geometrik talqini.

SVM vazn koeffitsiyentlari(noma’lum parametrlari)ni sozlash qoidalari quyidagicha amalga oshiriladi:

Ajratuvchi gipertekislikning tanlov nuqtalaridan iloji boricha uzoqroqda bo‘lishi uchun tayanch vektorlar orasidagi masofa (kenglik) imkon darajada maksimal bo‘lishi kerak. ω vektori ajratuvchi gipertekislikning normal vektoridir. Bu yerda ikkita a va b skalyar vektorlarning ko‘paytmasini <a,b> yoki $a^T b$ ko‘rinishda belgilaymiz. Uchlari turli sinflardagi tayanch vektorlar bo‘lgan vektorning ω vektoriga proyeksiyasini topamiz. Ushbu proyeksiya ajratuvchi kenglikning enini ko‘rsatadi (2-rasm):



2-rasm. Vazn koeffitsiyentlari(noma'lum parametrlari)ni sozlash qoidalarining natijasi.

$$\langle (x_+ - x_-), \omega/||\omega|| \rangle = (\langle x_+, \omega \rangle - \langle x_-, \omega \rangle)/||\omega|| = \big((b+1) - (b-1)\big)/||\omega||$$
$$= 2/||\omega||$$
$$2/||\omega|| \to max$$
$$||\omega|| \to min$$
$$\frac{(\omega^T \omega)}{2} \to min. \quad (4)$$

Sinflar chegarasidan *x* obyektining chetlanishi deb quyidagi kattalikga aytiladi:

$$M = y(\omega^T x - b). \quad (5)$$

Agar chetlanish kattaligi M faqat manfiy bo'lsagina (y va $(\omega^T x - b)$ har xil, ya'ni musbat va manfiy son qiymatlar qabul qilganda), algoritm obyektda xatolikka yo'l qo'yadi. Agar $M \in (0, 1)$ bo'lsa, u holda obyekt ajratuvchi kenglik ichiga tushadi. Agar $M > 1$ bo'lsa, u holda x obyekti to'g'ri tasniflangan va ajratuvchi kenglikdan ma'lum masofada joylashgan bo'ladi. Agar quyidagi shart bajarilsa, algoritm obyektlarni to'g'ri tasniflaydi, ya'ni to'g'ri sinflarga ajratadi:

$$y(\omega^T x - b) \geq 1 \quad (6)$$

Agar ikkita olingan ifodalar, ya'ni (4) va (6) birlashtirilsa, qat'iy chegara (hard-margin SVM) bilan standart SVM sozlamasi hosil qilinadi, bunda hech bir obyektga ajratish kengligi ichiga kirishga ruxsat berilmaydi. Masala Kuna-Taker teoremasidan foydalangan xolda analitik tarzda yechiladi. Hosil bo'lgan masala Lagranj funksiyasining egar nuqtasini topish ikkilamchi masalasiga ekvivalentdir

$$\begin{cases} \frac{(\omega^T \cdot \omega)}{2} \to \min\limits_{\omega, b} \\ y(\omega^T x - b) \geq 1 \end{cases} \quad (7)$$

(7) optimallashtirish masalasini quyidagi ko'rinishda yozib olamiz:

$$q(\omega) = (\omega^T \cdot \omega) = \omega_0^2 + \sum_{j=1}^{n} \omega_j^2 \to \min(\omega_j, \omega_0) => (\omega_0^*, \omega_1^*, \dots, \omega_n^*), \quad (8)$$

$$\begin{cases} \sum_{j=1}^{n} \omega_j \cdot x_{1i}^j + \omega_0 \geq 1 & i = 1,2,\dots,k_1 \\ \sum_{j=1}^{n} \omega_j \cdot x_{2i}^j + \omega_0 \leq -1 & i = 1,2,\dots,k_2 \end{cases} \quad (9)$$

bu yerda

$$\omega = (\omega_0, \omega_1, \dots., \omega_n) \quad (10)$$

$b = -\omega_0$.

Masalaning yechimi sifatida tanlanmani(o‘quv hamda nazorat) ikkita sinfga (ham va spam) ajratuvchi quyidagi gipertekislikni

$$R(\omega^*, x) = \sum_{j=1}^{n} \omega_j \cdot x^j + \omega_0 \quad (11)$$

topish masalasi turadi.

(9) tengsizliklar sistemasini (6) formuladan foydalanib, ikkita sinf obyektlari uchun quyidagi bitta tengsizliklar sistemasi ko‘rinishida yozib olish mumkin:

$$y_i(\sum_{j=1}^{n} \omega_j \cdot x_i^j + \omega_0) \geq 1 \qquad i = 1,2,\dots,k$$

U xolda optimallashtirish masalasi quyidagi ko‘rinishda ifodalanadi:

$$\mathrm{q}(\omega) = \omega_0^2 + \sum_{j=1}^{n} \omega_j^2 \rightarrow \min(\omega_j, \omega_0) => (\omega_0^*, \omega_1^*, \dots., \omega_n^*), \quad (12)$$

$$y_i(\sum_{j=1}^{n} \omega_j \cdot x_i^j + \omega_0) \geq 1 \qquad i = 1,2,\dots,k \quad (13)$$

Hosil bo‘lgan (12),(13) optimallashtirish masalasini stoxastik tasodifiy qidiruv usullarining asoschisi L.A.Rastrigin tomonidan taklif etilgan moslashuvchan(adaptiv) tasodifiy qidiruv usulidan foydalanib yechamiz [1].

NATIJALAR

Moslashuvchan tasodifiy qidiruv usuli. Umimiy holda quyidagi optimallashtirish masalasi qo‘yilgan bo‘lsin

$$q(\omega) \rightarrow min => \omega^* \quad (14)$$

$$\omega \in D$$

bu yerda $q(\omega)$ – umumiy holda nochiziqli ko‘p o‘zgaruvchili funksiya

$\omega = (\omega_1, \omega_2, \dots., \omega_n)$,

$\omega^* = (\omega_1^*, \omega_2^*, \dots., \omega_n^*)$ – (14) masalaning yechimi.

D – optimallashtirish masalasining aniqlanish sohasi, odatda tenglik va tengsizliklar ko‘rinishida berilishi mumkin.

Moslashuvchan tasodifiy qidiruv usulida quyidagi rekkurent formuladan foydalaniladi:

$$\omega^{k+1} = \omega^k + \Delta\omega^{k+1}, \quad (15)$$

$$\Delta\omega^{k+1} = \begin{cases} a^{k+1}\Delta\omega^k, & agarda \ \ q(\omega^k) < q(\omega^{k-1}) \\ a^{k+1} \cdot \xi^{k+1}, & agarda \ \ q(\omega^k) \geq q(\omega^{k-1}) \end{cases} \quad (16)$$

a^{k+1} – (k+1) qadamning uzunligini xarakterlovchi parametr. U optimallashtirish jarayonining qay darajada ketishiga qarab hozirgi holatga moslashtiriladi – agarda

bundan oldingi qadam muvofaqqiyatli bo‘lsa, u holda bu parametr kattalashtiriladi, aks holda kichiklashtiriladi:

$$a^{k+1} = \begin{cases} \delta_1 a^k, agarda\ \ q(\omega^k) < q(\omega^{k-1}) \\ \delta_2 a^k, agarda\ \ q(\omega^k) \geq q(\omega x^{k-1}) \end{cases} \qquad (17)$$

Bu yerda $\delta_1 > 1,\ \delta_2 < 1$ koeffitsiyentlar hisoblash jarayonining ijobiy natijadorligini taminlash maqsadidan kelib chiqib tanlanadi.

Bu parametrlarni to‘g‘ri tanlash masalaning ijobiy yechimini tanlashda alohida o‘rin tutadi.

ξ^{k+1} – (k+1) qadamdagi n-o‘lchovli birlik sferada barcha yo‘nalishlar bo‘yicha tekis taqsimlangan birlik tasodifiy vektordir.

U quyidagi algoritm yordamida hosil qilinadi:

$\xi = (\xi_1, \xi_{2,} \dots., \xi_n)$

$$\xi_i = \frac{\gamma_i}{\sqrt{\sum_{j=1}^{n} \gamma_j^2}}; \qquad i=1,2,.....,n \qquad (18)$$

Bu yerda γ_i $(i=1,2,....,n)$ $\gamma_i \in [-1,1]$ intervalda tekis taqsimlangan tasodifiy sonlar ketma-ketligi.

U quyidagi

$$\theta: \quad \sqrt{2};\ \frac{\sqrt{2}}{2};\ \sqrt{3};\ \frac{\sqrt{3}}{2}; \qquad (19)$$

irratsional sonlar qatnashgan ifodalardan ixtiyoriy birortasidan foydalangan xolda quyidagi hisoblash formulasi asosida hosil qilinadi.

$$\eta_j =]j \times \theta[, \quad j = 1,2, \dots \dots., ... \qquad (20)$$

Bu yerda $]\cdot[- j \times \theta$ ifoda hisoblangandan so‘ng, uning kasr qismi olinishi funksiyasini bildiradi. Bu η_j sonlar [0,1] intervalda tekis taqsimlangan tasodifiy sonlar ketma-ketligi hisoblanadi.

$$\gamma_j = (s-d)\eta_j + d, \quad j=1,2,\dots..... \qquad (21)$$

bu yerda d=-*1*, *s=1*.

Hosil bo‘lgan γ_j sonlar [-1,1] intervalda tekis taqsimlangan tasodifiy sonlar ketma-ketligi hisoblanadi.

Hosil qilingan formulalar yordamida A.P. Karpenkoning “Sovremenniye algoritmi poiskovoy optimizatsiii” o‘quv qo‘llanmasidagi misollarni yechimini topishda ham yaxshi samaraga erishsa bo‘ladi[2].

Bir o‘lchamli funksiyalar, Maslananing berilishi, **“volna”** funksiyasi

$$f(x) = e^{-x^2} + 0.01\cos(200x) \to max$$

$$x^* = 0$$

$$f^* = 2$$

$$[-\alpha; \alpha]^| \ \ \alpha = 2$$

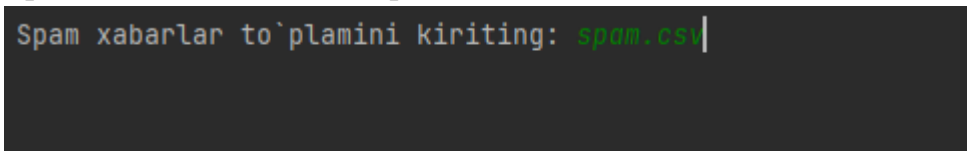
Yechimi:

0.013062675502308826
0.3727513179215124
1.01
0.3727513179215124
0.013062675502308826

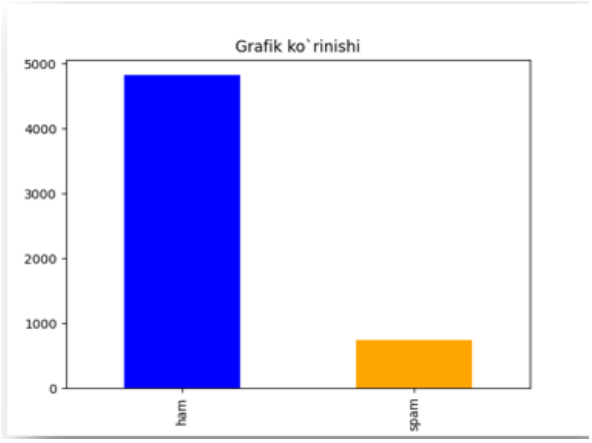
Dasturiy ta’minot spam habarlarni klasifikatsiya qilish orqali aniqlash imkoniyatini beruvchi Multi-NB,SVM, KNN, RF, AdaBoost modellar o‘rganib chiqildi va yuqori aniqlik berganliki sababli SVM(Support Vector Machine) tanlab olindi va unga dasturiy maxsulot ishlab chiqildi:

Bizda spam va ham habarlar to‘plami mavjud bo‘lib uchbu to‘plamni yuqoridagi modellarga tekshirib ko‘ramiz va aniqlilik darajasini hisoblaymiz

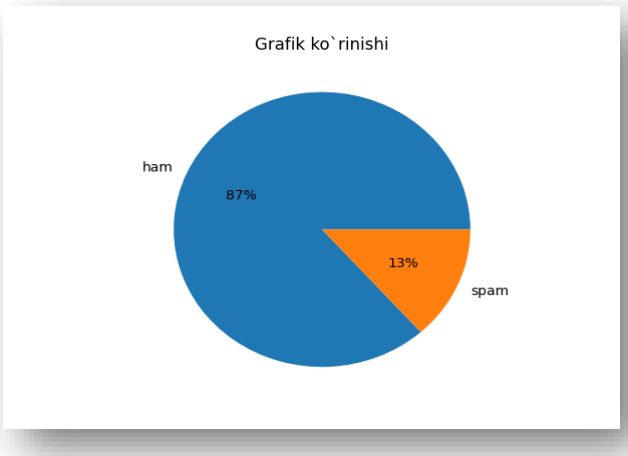
1. Spam va ham xabarlar to‘plamini kiritamiz.



2. Birinchi spam va ham habarlarni ajratib olamiz.



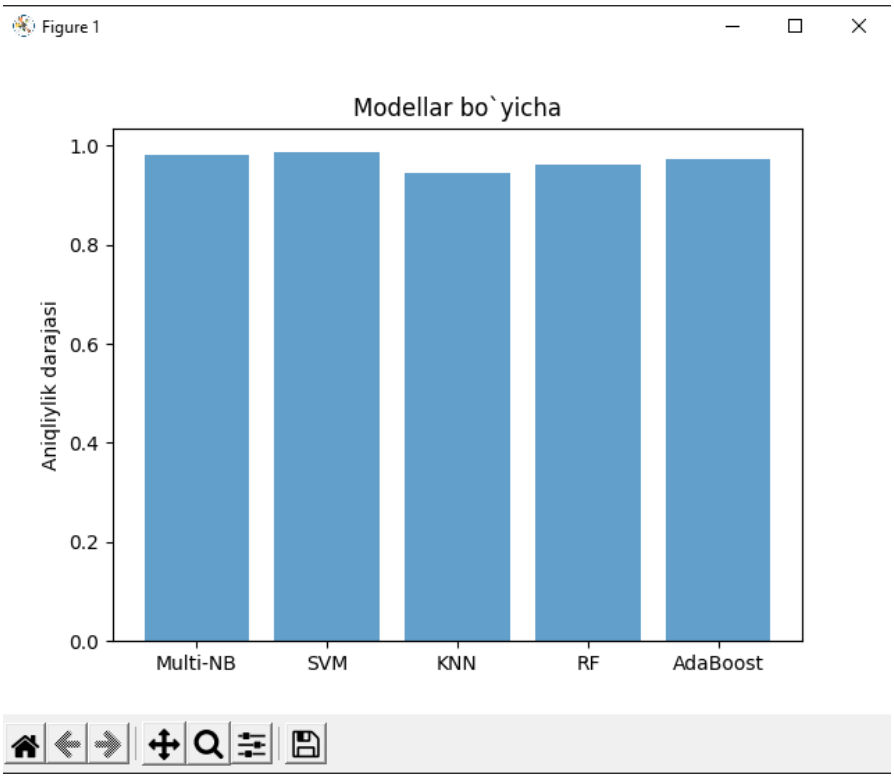
3. Foizda spam va ham habarlar.



4. Quyidagi natijalarni olamiz.

```
Xabarlarni klasslarga ajratish:  ['ham' 'spam']
Ham xabarlar soni : 4825
Spam xabarlar soni: 747


O`qitishlar soni:  3900
Testlar soni:  1672


Multi-NB
Aniqliylik darajasi %:
98.20574162679426
F1 Score
0.9344978165938864


SVM
Aniqliylik darajasi %:
98.6244019138756
F1 Score
0.9468822170900693


KNN
Aniqliylik darajasi %:
94.4377990430622
F1 Score
0.7452054794520547
```

5. Grafik shaklda ko‘rinishi bu yerda ko‘rib turganingizdek SVM eng yuqori natijani bergan 98.62 %.



Masalani nazorat tanlanmasi k=5572ta obyektlar uchun hisoblash tajribasini o‘tqazamiz. Bunda nazorat tanlanmasidagi obyektlar(xabarlar)ni ikkita sinf(ham yoki spam)ga toifalash maqsadida quyidagi produksion model yordamida ifodalangan qoidadan foydalanamiz:

$$x_i = \begin{cases} \text{spam}, & agar \ \sum_{j=1}^{6} x_i^j \geq b_q \ yoki \ (x_i^7 > 0.06 \ yoki \ x_i^8 > 0.03) \\ \text{ham}, & boshqa \ hollarda \end{cases}$$

i=1,2, . . . , k.

b_q - belgilar yig‘indisi uchun bo‘sag‘aviy parametr, ko‘rilayotgan masala uchun xususiy xolda b_q = 2 olish tavsiya etiladi. Ushbu *k=5572* nazorat tanlanmasi uchun to‘g‘ri toifalash ko‘rsatkichlari “spam” sinfi obyektlari uchun 4825ta obyektdan 4785tani, ya’ni 99,17%ni, “ham” sinfi obyektlari uchun 747ta obyektdan 736tani, ya’ni 98,5%ni tashkil etdi.

Ikkinchi turdagi xatolik bo‘yicha 25 ta “spam” xabarni 0.51% va 6 ta “ham” xabarni 0.8% noaniqlik bilan topdi. Umumiy natija alohida “spam” xabarlar uchun 98.65% va ham xabarlar uchun 97.72%ni hamda birgalikdagi “spam” va “ham” xabarlar uchun 98.6%ni tashkil etdi.

1-jadval.

Testlash natijalari.

obyekt	Nazorat tanlanmasi (summ - sinov xabarlar soni)	Birinchi turdagi xatolik (FRR - False Rejection Rate) - noto‘g‘ri rad etishlar soni (% yolg‘on rad etish)	To‘g‘ri toifalash natijasi, % da	Ikkinchi turdagi xatolik (FAR - False Acceptance Rate) - noto‘g‘ri aniqlashlar soni (noto‘g‘ri aniqlash%)	Identifikatsiya (umumiy) natijasi % da
spam	4825	40 (0,83%)	99,17%	25 (0.51%)	98.65%
ham	747	11 (1,5%)	98,5%	6 (0.8%)	97,72%
Jami (e mail = spam + ham)	5572	51 (1%)	99%	31 (0,6%)	98,6%

Birinchi turdagi xatolik (FRR - False Rejection Rate) – “spam” xabarni “spam” yoki ham xabarlarni “ham” deb aniqlamaslik holati (yoki% noto‘g‘ri rad etish).

Ikkinchi turdagi xatolik (FAR - False Acceptance Rate) - “spam” xabarni “ham” yoki “ham” xabarlarni “spam” deb aniqlash holati (yoki% noto‘g‘ri aniqlash).

I turdagi xatolik: (FRR / SUMM)*100
II turdagi xatolik: (FAR / SUMM)*100
Identifikatsiya (umumiy) Natija: 100-((FRR + FAR)*100)/ SUMM

XULOSA

Ishlab chiqilgan algoritm yordamida elektron pochta xabarlarini toifalarga tasniflash mumkin. Tasodifiy qidiruv usullarini qo‘llagan holda algoritmning ishlash samaradorligini yanada oshirish imkoniyati tug‘iladi. Spam-hujjatlarga va qonuniy sinflashtirishni samarali algorimtlarini qurish uchun spamni tavsifini va xususiyatlarini aniqlash imkonini beradi. Bu yaratilgan algoritmni nafaqat spamlarni aniqlashda balkim toifalarga tasniflashning qolgan masalalarida ham qo‘llasa bo‘ladi.

ADABIYOTLAR RO‘YXATI

1. *L.A.Rastrigin.* Sovremenniye prinsipi upravleniya slojnimi obyektami. *M.*: Sovetskoye radio. 1980. - 232s.
2. *A.P. Karpenko.*Sovremenniye algoritmi poiskoviy optimizatsii. M.:Izdatelstvo MGTU im. N.E.Baumana, 2014. - 446s.
3. *Burkov Andrey*, Mashinnoye obucheniye bez lishnix slov. — SPb.: Piter, 2020. — 192 s.
4. Khamdamov R.Kh., Khaydarov E.D. “Detecting spam messages using the naïve Bayes algorithm of basic machine learning” International Conference on Information Science and Communications Technologies(ICISCT) 2021, Tashkent.
5. Khamdamov R.Kh., Khaydarov E.D. “Pre-processing of primary spam classification data from email messages” Sovremennoye sostoyaniye i perspektivi razvitiya sifrovix texnologiy i iskusstvennogo intellekta Sbornik dokladov respublikanskoy nauchno-texnicheskoy konferensii Samarkand, 26-27 oktabrya 2022 g.
6. *T. Savita, B. Santoshkumar*, Effective spam detection method for email, international conference on advances in engineering & technology - 2014 (ICAET2014), OSR J. Comp. Sci. (IOSR – JCE) (2014).