HODISALARNI BASHORAT QILISHDA NAÏVE BAYES KLASSIFIKATORIDAN FOYDALANISH.

Nabiyev Xolbek Tolib o'g'li

"Mirzo Ulug'bek" nomidagi O'zbekiston Milliy Universiteti magistranti
E-mail: nabiyevxolbek2000@gmail.com

K E Y W O R D S

Axborot xavfsizligi, konfidentsial ma'lumot, tarmoq monitoringi, mashinaviy o'qitish

A B S T R A C T

Tabiiy tilni qayta ishlash jarayonlarining muhim vazifalaridan biri bu-matnni klassifikatsiya qilishdir . Matnni tasniflash orqali undagi sodir bo'lishi mumkin bo'lgan ehtimolliklarni yuqori aniqlikda aniqlash imkoni paydo bo'ladi . Ushbu muammoga javob topishda mashinali o'qitishga asoslangan usullar keng qo'llaniladi. Mazkur maqolada matnni tasniflash algoritimiga misol tariqasida Naïve Bayes algoritimi haqida ma'lumot va ushbu algoritimga misol keltiriladi va uning ishlash jarayoni haqida ma'lumot berilgan.

Bugungi kunda zamonaviy axborot texnologiyalarisiz hech bir sohani tasavvur qilaolmaganimiz kabi , matnni klassifikatsiya qilish jarayoni ham ba'zi sohalarning asosiy elementi hisoblanadi. Ularga misol qilib:

-Spam xabarlarni aniqlash: elektron pochtaga keladigan xabarlarni spam yoki spam emasligini aniqlash,

-Sentiment tahlili: Matnnda ifoda etilgan subyektiv fikrlar (Masalan: Biror mahsulot haqida istemolchilar fikri. Ijobiy,salbiy va h.k.)

-Chatbotlar va virtual yordamchilar: Foydalanuvchi maqsadini tushinish va unga kerakli javoblarni berish,

-Ilmiy maqolalarni tasniflash: Matn tasniflagichlar maqolalarni avtomatik tarzda tegishli kategoriyalarga ajratadi(Masalan: tibbiyot , fizika , kompyuter va tarmoq sohalari v.h.k [1]

Albatta matn klassifikator muhim ro'l o'ynaydigan va yana bir nechta sohalar mavjud va eng asosiysi ushbu sohalarning har birida matn klassifikatori muhim ahamiyatga ega va u aynan jarayonni avtomatlashtirishga , vaqtni tejashga va aniqlikni oshirishga katta hissa qo'shadi.

Biz yashayotgan XXI-asr ,axborot asrida dunyo bo'ylab har soniyada millionlab axborotlar almashinuvi yuz bermoqda , va ularning ulkan qismi kompyuter tizimlari va tarmoqlarida faol tarzda amalga oshiriladi . Ushbu jarayonda tarmoq va tizimdagi ma'lumotlar orasidan aynan konfidentsial axborotlani tez va samarali aniqlash maqsadida bir necha mashhur hisoblangan matnni tasniflovchi algoritimlardan boshqacha qilib aytganda matn klassifikatorlaridan foydalanish mumkin .Bularga misol sifatida Naive Bayes, Decision trees, Random Forests va yana bir qanchasini misol keltrish mumkin

Naive Bayes klassifikatori – ehtimolik nazariyasiga asoslangan oddiy, ammo samarali tasniflovchi modellardan biridir [2][1]. Shuningdek ko'plab tadqiqotlarda keltrilishicha **"**Naïve Bayes klassifikatori , matnni yoki hujjatlarni o'qitishda hodisalarning yuz berish ehtimolligi asosida avvaldan belgilangan turlarga ajratadi va ehtimollikning eng yuqori natijasini tanlaydi.Ushbu algoritim asosan mashinani o'qitishda klassifikatsiya muammolarini hal etishda muhim ahamiyatga ega hisoblanadi".Ta'kidlash kerakki

[1] Andrew K. McCallum and Kamal Nigam. 1998. Employing EM in pool-based active learning for text classification. Proceedings of ICML-98, 15th International Conference on Machine Learning

© *Nabiyev Xolbek Tolib o'g'li*

ushbu algoritim o'zining oddiyligi va samaradorligi bilan mashhurdir [3]. Uning yordamida modellar yaratish va bashorat qilish tezroq amalga oshadi [4].Naive Bayes algoritimini matnni tasniflash va spam xabarlarni aniqlash va yana bir qancha turli vazifalarda qo'llash mumkin. [5]

Naive Bayes klassifikatori ishlov beradigan ma'lumotlar turiga qarab 4 toifaga bo'linadi va ular quyidagilar:

- **Gaussion Naive Bayes-** bu asosan doimiy (raqamli) atributlar bilan ishlashga moslashgan.Shuningdek ushbu klassifikatorda hisoblash jarayoni juda sodda va katta ma'lumotlar hajmi bilan oson ishlaydi.Xotiraga bo'lgan ehtiyoji ham pas. [6]
- **Optimal Naive Bayes –**bu sodir bo'lish mumkin bo'lgan barcha gipotezalar va ehtimollar sinfni tanlaydi va bu klassifikator nazariy jihatdan eng yaxshi natijalani beradi Lekin u juda sekin, shuningdek ko'p vaqt talab qiladigan shartlardan o'tadi.
- **Multinomial Naive Bayes-**ushbu klassifikator hujjatlarni tasniflashda masalalarida va so'z chastotasi bilan ishlashda samarali hisoblanadi.
- **Bernoulli Naive Bayes-**ushbu algoritim asosan ikkilik (binary) yoki mantiqiy tipdagi ma'lumotlar bilan ishlashga mo'ljallangan hisoblanadi.Biroq u uzliksiz ma'lumotlarga mos emas [7]

Bayes teoremasi:

$$P(A|B)=\frac{P(B|A)*P(A)}{P(B)}$$

P(B)≠0

Bu yerda:

- P(A|B)- shartli ehtimollik. B hodisa berilgan A hodisaning sodir bo'lish ehtimolligi (posterior)
- P(B|A)- shartli ehtimollik. A hodisasi sharti bilan berilgan B hodisaning sodir bo'lish ehtimolligi (likelihood)
- P(A)- A ning yuzaga kelish ehtimoli (prior);
- P(B)- B ning yuzaga kelish ehtimoli (evidence); [8]

B o'zgaruvchisi xususiyatlarni ifodalaydi va u quyidagicha berilgan :

$$B=(b1,b2,b3,.....,bn)$$

Bu orqali formulani quyidagi ko'rinishda yozib olamiz:

$$P(A|b_1,b_2....b_n)=\frac{(P(b1|A)(P(b2|A)....(P(bn|A)*P(A)}{(P(b1)(P(b2)....(P(bn)} \text{ [9]}$$

Naive Bayes klassifikatorining ishlash jarayoni:

1. **Malumotlarni tayyorlash-** bu jarayonda barcha atributla ikkilik formatga o'tkazilidi [10]
2. **Ehtimollarni hisoblash:**har bir sinf uchun atributlarning mavjudligi yoki yo'qligi hisoblanadi
3. **Tasniflash:** Ehtimoli eng yuqori sinf hisoblanadi [5]

Ehtimollikni aniqlash algoritimlari orqali ob-havo bashorati kabi taxminiy fikrlarni ham ilgari surish mumkin. Ob-havo pragnozi avvalgi va hozirgi atmosfera holati to'g'risidagi miqdoriy ma'lumotlarni to'olash orqali amalga oshiriladi. [11]

Naive Bayes klassifikatoriga misol tariqasida ushbu misolni ko'rib chiqamiz.

Muammo bayoni: Naive Bayes klassifikatoridan foydalanib, futbolchilar 14 kun davomida ob-havoning ko'rinishi. tempraturasi, namligi va shamol bo'lganda ma'lum kombinatsiyalardan foydalanib o'yin o'ynashadimi yoki yo'qligin taxmin qilish.

Day	Kun	Temprature	Namlik	Shamol	Futbol o'ynamoq
1.	Quyoshli	Issiq	O'rtacha	Kuchsiz	Yo'q
2.	Quyoshli	Issiq	O'rtacha	Kuchli	Ha
3.	Bulutli	Salqin	Yuqori	Kuchli	Ha

© Nabiyev Xolbek Tolib o'g'li

4.	Bulutli	Salqin	O‘rtacha	Kuchli	Ha
5.	Yomg‘ir	Iliq	Yuqori	Kuchli	Yo‘q
6.	Bulutli	Iliq	Yuqori	Kuchsiz	Yo‘q
7.	Yomg‘ir	Salqin	O‘rtacha	Kuchli	Ha
8.	Yomg‘ir	Salqin	Yuqori	Kuchli	Yo‘q
9.	Quyoshli	Iliq	Yuqori	Kuchsiz	Ha
10.	Bulutli	Iliq	O‘rtacha	Kuchsiz	Ha
11.	Quyoshli	Issiq	O‘rtacha	Kuchli	Yo‘q
12.	Yomg‘ir	Iliq	Yuqori	Kuchsiz	Ha

Ushbu jadvaldan quyidagi qiymatlarni hosil qilamiz.

P(Futbol o‘ynamoq=Ha)=7/12

P(Futbol o‘ynamoq=Yo‘q)=5/12

Navbatdagi qadamda berilgan 4 ta kiritishni o‘z ichiga olgan ob havo ko‘rinishi , temperaturasi, namlik,shamol -har bir atribut uchun chastotalar jadvalini to‘ldiramiz

Tempratura	**Ha**	**Yo‘q**
Issiq	**1/7**	**2/5**
Salqin	**3/7**	**1/5**
Iliq	**3/7**	**2/5**

Kun	**Ha**	**Yo‘q**
Quyoshli	2/7	2/5
Bulutli	3/7	1/5
Yomg‘ir	2/7	2/5

Shamol	**Ha**	**Yo‘q**
Kuchli	4/7	3/5
Kuchsiz	3/7	2/5

Namlik	**Ha**	**Yo‘q**
O‘rtacha	3/7	3/5
Yuqori	3/7	3/5

Ehtimollik jadvallari ob-havo bir vaqtda quyoshli , tempratura salqin namlik o‘rtacha va shamol kuchsiz kombinatsiyasida o‘yinchilar futbol o‘ynashii yoki yo‘qligini hisoblash uchun ishlatilishi mumkin. B teng bo'lgan quyidagi omillarning kombinatsiyasini ko'rib chiqing:

- Kun – quyoshli
- Tempratura –salqin
- Namlik-o‘rtacha
- Shamol-kuchsiz

Yuqoridagi shartlarga asoslanib o‘yinchilarning futbol o‘ynash ehtimolini hisoblaymiz

A= futbol o‘ynamoq(ha)
B= futbol o‘ynamoq(yo‘q)

P(A|B)=P(futbol o‘ynamoq|Kun=Quyoshli,Tempratura=salqin,namlik=o‘rtacha,Shamol= Kuchsiz)

P(B|A)=P(Ha)P(Quyoshli|Ha) P(Salqin|Ha) P(O‘rtacha|Ha) P(Kuchsiz|Ha)=

=(7/12)(2/7)*(3/7)*(3/7)*(3/7)≈0.013*

Va ushbu ehtimollikni teskari holatdagi natijasini ham hisob topamiz

Ya’ni:

P(A|B)=P(yo‘q)P(Quyoshli| yo‘q) P(Salqin| yo‘q) P(O‘rtacha| yo‘q) P(Kuchsiz| yo‘q)=

=(5/12)*(2/5)*(1/5)*(3/5)*(2/5)≈0.008

© *Nabiyev Xolbek Tolib o‘g‘li*

Ushbu natijalardan foydalanib quyidagi qiymatni hisoblaymiz:

$$P(A|B)= 0.013/(0.013+0.008)=0.619$$

$$P(B|A)= 0.008/(0.013+0.008)=0.380$$

Natija :

$$P(A|B)> P(B|A)$$

Ya’ni Ushbu gipotezaning sodir bo‘lish ehtimoli uning sodir bo‘lmaslik ehtimolidan katta . Bu natijaga asoslanib, biz kiritgan atributlarning sodir bo‘lishi ehtimoli mavjuddir.

Naive Bayes algoritimi aynan shunga o‘xshash va turli xil atributlarga asoslanib sinflarning ehtimolini bashorat qilish uchun shunga o‘xshash sinflardan foydalanadi.

Naive Bayes algoritimining ustunlik va kamchiliklari

Ustunligi	Kamchiligi
Ushbu algoritim tez va foydalanishda oson hisoblanib, ma’lumotlar to‘plami sinfini tezda taxmin qilishda ko‘maklashadi [12]	Katta ma’lumotlar to‘plami bilan ishlaganda juda oddiy modellash sekin ishlashi mumkin
Ushbu algoritim kamroq o‘qitish ma’lumotlari bilan yaxshiroq natija ko‘rsatadi	Aniqlikni yo‘qotisj ehtimoli mavud
Naïve Bayes algoritimini tushunish va amal qilish oson , uni o‘rganish jarayoni ko‘p vaqt talab qilmaydi	Xususiyatlarning mustaqil bo‘lishi haqidagi gipoteza real hayotdagi NLP ilovalariga mos kelmaydi
Bir vaqtning o‘zida bir qancha sinflarni prognoz qilishda ham samarali natijalar beradi	O'zgaruvchilarning ma'lum bir ma'lumotlarga qanchalik eski yoki yangi bo'lishi natijalarga ta'sir qilishi mumkin.

Xulosa

Ushbu maqolada kelajakda sodir bo‘lishi kutilayotgan hodisalar ehtimolligini aniqlash ustuvor vazifa sifatida qaraladi va unga misol sifatida ob-havoni oldindan taxmin qilish uchun Naïve Bayes algoritimi modeli taklif qilingan va prognoz qilingan ob- havo holatlari yomg‘ir, tuman va quyosh shuningdek shamoldir. Modelning samaradorligiga qaramay uni kengaytrilishi va hatto yanada takomillashtrilishi mumkin.Ma’lumotlar to‘plami ko‘proq elemntlarni qamrab olishi shuningdek ma’lumotlarni qayta ishlashda takomillashgan texnikalardan foydalanish mumkin.

Foydalanilgan adabiyotlar va internet saytlari

1. X. Zhou, X. Tao, R. Gururajan and G. Bargshady, "A Survey on Text Classification and Its Application," Web Intelligence, pp. 2-5, 2020.
2. Andrew, K. McCallum and K. Kamal Nigam., "Employing EM in pool-based active learning for text classification," in 15th International Conference on Machine Learning, 1998.
3. E. B. Boltayevich and Axmedova Xolisxon Ilxomovna, "SO‘Z MA’NOSINI ANIQLASHDA NAIVE BAYES ALGORITMIDAN FOYDALANISH," Journal of Science and Innovative Development, pp. 2-5, 2023.
4. M. Murata, M. Utiyama, K. Uchimoto, Q. Ma and H. Isahara, "Japanese word sense disambiguation using the simple Bayes and support vector machine methods," in : The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, Online, 2001.
5. P. Domingos, Michael and J. Pazzani, "On the optimality of the simple bayesian classifier

© Nabiyev Xolbek Tolib o‘g‘li

under zero-one loss," in Machine Learning, Netherland, 1997, pp. 100-125.
6. S. Xu, "Bayesian Naïve Bayes Classifiers to Text Classification," Journal of Information Science, pp. 1-8, 2016.
7. Y. Gao, Y. Xu and Y. Li, "Pattern-based topics for document modelling in information filtering," IEEE, 2014.
8. Z. H, "The optimality of naive Bayes," in Proceedings of the 17th International Florida Artificial, 2004.
9. S.-B. Kim, K.-S. Han, H.-C. Rim and a. S. H. Myaeng, "Some Effective Techniques for Naive Bayes," IEEE Transactions on Knowledge and Data Engineering, pp. 1-7, 2006.
10. Z. Wan, "TEXT CLASSIFICATION: A PERSPECTIVE OF DEEP LEARNING," JOUR, 2023.
11. M. Das and S. Ghosh, "A Probabilistic Approach for Weather Forecast Using Spatio-Temporal Interrelationships Among Climate Variables," in In 2014 9th International Conference on Industrial and Information, Online, 2014.
12. M. D, "Feature Subset Selection in Text-Learning," in Proc.10th European Conf. Machine Learning, 1998.
13. X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in neural information processing systems, 2015, 2020, pp. 650-660.
14. Y. C. Yew, T. G. Cheng, D. C. W. Sen, B. J. Wei and Z. A. A. Salam, "Text classification with Naïve Bayes," Journal of Applied Technology and Innovation, pp. 1-4, 2021.
15. J. T, "Text Categorization with Support Vector Machines:Learning with Many Relevant Features," in 10th European Conf. Machine Learning, 1998.
16. H.-C. S. a. H.-C. R. Sang-Bum Kim, "Poisson Naive Bayes for Text Classification with Feature Weighting," Research gate, pp. 2-5, 2003.
17. L. P, " Induction of recursive Bayesian classifiers," in Proceedings of the Eighth European Conference, Austria , Vienna, 1993.
18. Y. Li, X. Zhou, P. Bruza, Y. Xu and R. Lau, "A two-stage text mining model for information filtering," in Proceedings of the 17th ACM conference on Information and knowledge management, 2008.
19. R. Kohavi, B. Becker and D. Sommerfield, "Improving simple Bayes," ftp://starry.stanford.edu/pub/ronnyk/-, pp. 5-12, 1997.
20. L. K., "Learning to filter netnews.," in Proceedings of the 12th International Conference on Machine L, 1995.
21. D. Jurafsky and J. H. Martin, "Naive Bayes, Text Classification, and Sentiment," in Speech and Language Processing, 2024, p. Chapter 4.
22. R. I, "An empirical study of the naive Bayes classifier," IJCAI Workshop on Empirical Methods in AI, pp. 1-5, 2001.
23. D. Heckerman, E. Horvitz, M. Sahami and S. T. Dumais, "A Bayesian approach to filtering junk e-mail," AAAI-98 Workshop on Learning for Text Categorization., pp. 5-15, 1998.
24. H. DJ and Y. K, "Idiot’s Bayes–Not so stupid after all," International Statistical Review, pp. 385-395, 2001.
25. I. K. Oshodi, "Machine Learning-Based Algorithms for Weather Forecasting," International Journal of Artificial Intelligence and Machine Learning ·, pp. 1-5, July 2022.
26. Muminov B., Egamberdiyev E. OPTIMIZATION OF FUZZY INFERENCE SYSTEMS WITH GENETIC ALGORITHMS //DTAI–2024. – 2024. – T. 1. – №. DTAI. – C. 250-252.

© *Nabiyev Xolbek Tolib o‘g‘li*