HUJJATLARNI TASNIFLASH ALGORITMLARINI QIYOSIY TAHLILI

Davletov A.Y.

Alfraganus Unversity nodavlat oliy talim tashkilotining Raqamli texnologiyalar kafedrasi dotsenti, texnika fanlari boyicha falsafa doktori (PhD).

Sabirova S.T.

Alfraganus Unversity nodavlat oliy talim tashkiloti Raqamli texnologiyalar fakulteti “Kompyuter injiniringi” mutaxassisligi 1-bosqich magistranti.

https://doi.org/10.5281/zenodo.13325064

Annotatsiya. *Hujjatlarni tasniflash tabiiy tilni qayta ishlash (NLP) sohasidagi hal qiluvchi vazifa bo'lib, matn hujjatlarini oldindan belgilangan sinflar yoki mavzularga ajratishni o'z ichiga oladi. Ushbu jarayonni avtomatlashtirish uchun har xil algoritmlar ishlab chiqilgan bo'lib, ularning har biri o'ziga xos kuchli va cheklovlarga ega. Ushbu maqolada biz hujjatlarni tasniflashning mashhur algoritmlarini qiyosiy tahlil qilib, ularning metodologiyasi, afzalliklari va muammolari haqida tushunchalarni ochib beramiz.*

Kalit so'zlar: *hujjatlar, tasniflash, algoritm, tahlil qilish, avtomatlashtirish.*

COMPARATIVE ANALYSIS OF DOCUMENT CLASSIFICATION ALGORITHMS

Abstract. *Document classification is a critical task in the field of natural language processing (NLP), which involves classifying text documents into predefined classes or topics. Various algorithms have been developed to automate this process, each with its own strengths and limitations. In this article, we provide a comparative analysis of popular document classification algorithms, revealing insights into their methodology, advantages, and challenges.*

Key words: *documents, classification, algorithm, analysis, automation.*

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ КЛАССИФИКАЦИИ ДОКУМЕНТОВ

Аннотация. *Классификация документов является важной задачей в области обработки естественного языка (NLP), которая включает в себя классификацию текстовых документов по заранее определенным классам или темам. Для автоматизации этого процесса были разработаны различные алгоритмы, каждый из которых имеет свои сильные стороны и ограничения. В этой статье мы даем сравнительный анализ популярных алгоритмов классификации документов, раскрывая их методологию, преимущества и проблемы.*

Ключевые слова: *документы, классификация, алгоритм, анализ, автоматизация.*

Kirish. Axborotning haddan tashqari yuklanishi davrida katta hajmdagi matnli ma'lumotlarni avtomatik ravishda tasniflash va tartibga solish qobiliyati birinchi o'ringa chiqdi.

Hujjatlarni tasniflash algoritmlari bu ishda muhim rol o'ynaydi va samarali ma'lumot olish, his-tuyg'ularni tahlil qilish va kontentni tavsiya qilish imkonini beradi. Turli tasniflash algoritmlarining nuanslarini tushunib, tadqiqotchilar va amaliyotchilar ma'lumotlarni qayta ishlash quvurlarini optimallashtirishlari va qaror qabul qilish jarayonlarini yaxshilashlari mumkin.

Adabiyotlar tahlili va tadqiqot metodologiyasi. Naive Bayes - Bayes teoremasiga asoslangan, xususiyat mustaqilligi faraziga asoslangan ehtimolli algoritm. U soddaligi va hisoblash samaradorligi tufayli matnlarni tasniflash vazifalari uchun keng qo'llaniladi. Naive Bayes kichik va o'rta o'lchamdagi ma'lumotlar to'plamlari bilan yaxshi ishlaydi va siyrak ma'lumotlar bilan ishlashga mohir. Biroq, xususiyatlar haqiqatan ham mustaqil bo'lmaganda yoki ma'lumotlar ichidagi o'ta murakkab munosabatlar bilan shug'ullanganda uning ishlashi yomonlashishi mumkin.

Yordam vektor mashinalari (SVM) ko'p qirrali nazorat qilinadigan o'rganish algoritmlari bo'lib, ular yuqori o'lchamli ma'lumotlarni tasniflashda ustunlik qiladi. SVMlar xususiyat maydonida turli sinflarni ajratib turadigan optimal giperplanni topish orqali ishlaydi. Ular mustahkam, yaxshi umumlashtirish qobiliyatiga ega va chiziqli va chiziqli bo'lmagan tasniflash vazifalarini bajara oladi. Shunga qaramay, SVMlar giperparametrlarni sinchkovlik bilan tanlashni talab qiladi va hisoblash intensiv bo'lishi mumkin, ayniqsa katta ma'lumotlar to'plamlari bilan [2].

Tasodifiy o'rmon - bu bir nechta qaror daraxtlaridan iborat ansambl usuli. U haddan tashqari moslashishga chidamliligi, katta ma'lumotlar to'plamlari uchun miqyosi va shovqinga chidamliligi bilan mashhur. Tasodifiy o'rmonlar yuqori o'lchamli xususiyatlar bo'shliqlarini boshqarishda yaxshi ishlaydi va tashqi ko'rsatkichlarga nisbatan kamroq sezgir. Biroq, ular muvozanatsiz ma'lumotlar to'plami uchun eng yaxshi tanlov bo'lmasligi mumkin, chunki ko'pchilik sinf bashoratlarda ustunlik qiladi.

Muhokama va natijalar. Konvolyutsion neyron tarmoqlari (CNN) hujjatlar tasnifi sohasida inqilob qilgan chuqur o'rganish modellari. CNN konvolyutsion filtrlar va birlashtirish operatsiyalarini qo'llash orqali matn ma'lumotlaridan ierarxik xususiyatlarni avtomatik ravishda o'rganadi. Ular fazoviy munosabatlarni va matn ma'lumotlaridagi murakkab naqshlarni egallashga mohir. Biroq, CNNlar o'qitish uchun katta hajmdagi etiketli ma'lumotlarni va keng hisoblash resurslarini talab qiladi. Tasodifiy o'rmon algoritmi ansambl usullari oilasiga tegishli bo'lgan kuchli va ko'p qirrali mashinani o'rganish texnikasi. U tasniflash va regressiya vazifalari uchun keng qo'llaniladi va turli sohalarda ajoyib samaradorlikni namoyish etadi. Ushbu maqolada biz

"Tasodifiy o'rmon" algoritmining ichki ishlashi, uning asosiy xususiyatlari, afzalliklari va ilovalari bilan tanishamiz. Tasodifiy o'rmon algoritmi o'z mohiyatiga ko'ra qaror daraxtlari ansamblidir.

Har bir qaror daraxti ma'lumotlarning bir qismi bo'yicha mustaqil ravishda o'qitiladi va individual bashorat qiladi. Tasodifiy o'rmonning yakuniy bashorati keyin barcha alohida daraxtlarning bashoratlarini yig'ish orqali aniqlanadi. Ushbu yig'ish jarayoni haddan tashqari moslashishni kamaytirishga yordam beradi va modelning umumiy bashorat aniqligini yaxshilaydi [1].

Tasodifiy o'rmon algoritmining asosiy xususiyatlari:

Bootstrap Aggregating (Bagging): Tasodifiy o'rmon paketlash deb ataladigan texnikadan foydalanadi, bu har bir qaror daraxtini o'quv ma'lumotlarining tasodifiy kichik to'plamiga o'rgatishni o'z ichiga oladi. O'quv namunalarini tanlashda bunday tasodifiylik turli xil va mustahkam modellarni yaratishga yordam beradi. Ta'lim ma'lumotlarini tanlab olishdan tashqari, Random Forest qaror daraxtining har bir bo'linishi uchun xususiyat tanlashda tasodifiylikni ham taqdim etadi. Bu ansambldagi turli xil daraxtlarning turli xil kichik xususiyatlar to'plamiga e'tibor qaratishini ta'minlaydi, bu esa yaxshiroq umumlashtirishga olib keladi va ortiqcha moslashish xavfini kamaytiradi. Tasodifiy o‘rmondagi alohida qaror daraxtlari sayozdir, ya’ni ular juda chuqur o‘sishi mumkin emas. Bu modelning ma'lumotlardagi shovqinni ushlashiga yo'l qo'ymaydi va izohlashni yaxshilaydi [3].

Tasodifiy o'rmon algoritmining afzalliklari:

Tasodifiy o‘rmonlar hatto katta va murakkab ma’lumotlar to‘plamida ham yuqori bashorat qilish aniqligi bilan mashhur. Ular bitta qaror daraxti bilan solishtirganda haddan tashqari moslashishga kamroq moyil. Tasodifiy o‘rmonlar bir nechta daraxtlarning yig‘ilishi tufayli shovqinli ma’lumotlar va o‘zgarib turadigan ma’lumotlarga chidamli. Ular etishmayotgan qiymatlarni boshqarishi va yaxshi ishlashni saqlab qolishlari mumkin. Tasodifiy o‘rmonlar funksiya ahamiyati haqida tushuncha beradi, bu esa foydalanuvchilarga model bashoratiga ta’sir etuvchi o‘zgaruvchilarni tushunish imkonini beradi [4].

Tasodifiy o'rmon algoritmini qo'llash:

Tasodifiy o‘rmonlar spam elektron pochta xabarlarini aniqlash, his-tuyg‘ularni tahlil qilish va mijozlarning ishdan chiqishini bashorat qilish kabi tasniflash vazifalarida keng qo‘llaniladi.

Tasniflashdan tashqari, Tasodifiy o'rmonlar uy-joy narxlarini bashorat qilish, aksiyalar narxlari va sotishni prognozlash kabi regressiya muammolariga qo'llanilishi mumkin.

Tasodifiy o‘rmonlardan anomaliyalarni aniqlash vazifalari uchun ham foydalanish mumkin, bunda kamdan-kam uchraydigan hodisalarni yoki o‘zgacha ko‘rsatkichlarni aniqlash

muhim ahamiyatga ega. Tasodifiy o'rmon algoritmi mashinani o'rganish asboblar qutisida ko'p qirrali, mustahkam va yuqori samarali vosita sifatida ajralib turadi. Uning murakkab ma'lumotlar to'plamini boshqarish, aniqlikni saqlash va xususiyatlarning ahamiyati haqida qimmatli tushunchalarni taqdim etish qobiliyati uni turli xil bashoratli modellash vazifalari uchun asosiy tanlovga aylantiradi. Yangi boshlovchi yoki tajribali ma’lumot olimi bo‘lasizmi, “Tasodifiy o‘rmon” algoritmini o‘zlashtirish sizning mashinani o‘rganish imkoniyatlarini sezilarli darajada oshirishi mumkin [5].

Xulosa. Xulosa qilib aytganda, hujjatlarni tasniflash algoritmini tanlashda ma'lumotlar to'plamining o'ziga xos xususiyatlari va kerakli ishlash ko'rsatkichlari hisobga olinishi kerak.

Naive Bayes kichik ma'lumotlar to'plamlari bilan oddiy vazifalar uchun javob beradi, SVM esa yuqori o'lchamli ma'lumotlar bilan ishlashda samarali. Tasodifiy o'rmon mustahkam va kengaytirilishi mumkin, ammo muvozanatsiz ma'lumotlar to'plamlari bilan kurashishi mumkin.

CNN eng zamonaviy ishlashni taklif qiladi, lekin katta hisoblash resurslarini talab qiladi.

Har bir algoritmning kuchli tomonlari va cheklovlarini o'lchab, amaliyotchilar o'zlarining hujjatlarni tasniflash ehtiyojlariga samarali javob berish uchun oqilona tanlov qilishlari mumkin.

REFERENCES

1. B. Xu, J. Z. Wang, and D. Y. Peng, “Practical Protocol Steganography: Hiding Data in IP Header,” Proceedings - 1st Asia International Conference on Modelling and Simulation: Asia Modelling Symposium 2007, AMS 2007, pp. 584–588, 2007, doi: 10.1109/AMS.2007.80.[1]
2. K. Szczypiorski, “A performance analysis of HICCUPS - A steganographic system for WLAN,” 1st International Conference on Multimedia Information Networking and Security, MINES 2009, vol. 1, pp. 569–572, 2009, doi: 10.1109/MINES.2009.248.[2]
3. A. S. Nair, A. Sur, and S. Nandi, “Detection of packet length-based network steganography,” Proceedings - 2010 2nd International Conference on Multimedia Information Networking and Security, MINES 2010, pp. 574–578, 2010, doi: 10.1109/MINES.2010.126.[3]
4. C. H. Rowland, “Covert Channels in the TCP/IP Protocol Suite,” undefined, vol. 2, no. 5, 1997, doi: 10.5210/FM. V2I5.528.[4]
5. S. Cabuk, C. E. Brodley, and C. Shields, “IP Covert Channel Detection,” undefined, vol. 12, no. 4, Apr. 2009, doi: 10.1145/1513601.1513604[5]