XORIJIY TALABALAR MA'LUMOTLARIDA SHUBHALI YOZUVLARNI ANIQLASHDA ISOLATION FOREST ALGORITMINI QO'LLASH

Abdullayev Javoxir Jaxongirovich
Buxoro davlat texnika universiteti o'qituvchisi, i.f.f.d. (PhD)

Sun'iy intellekt texnologiyalarini davlat boshqaruvi va ijtimoiy sohalarda qo'llash O'zbekistonda ustuvor vazifa sifatida belgilangan [1]. Oliy ta'lim muassasalarida xorijiy talabalar haqidagi ma'lumotlarni nazorat qilish ham ana shunday yo'nalishlardan biri hisoblanadi.

Xorijiy talabalar ma'lumotlaridagi xatolarning bir qismi aniq qoidalar bilan topiladi: bo'sh maydon, noto'g'ri format, mantiqiy ziddiyat. Biroq ayrim yozuvlar formal jihatdan to'g'ri bo'lib, faqat umumiy manzaradan ajralib turadi. Masalan, fuqarolik hisobot topshirilishidan bir necha kun oldin o'zgartirilgan yoki viza muddati tugaganiga qaramay talaba faol holatda qolgan. Bunday holatlarning barcha turlarini oldindan qoida sifatida yozib bo'lmaydi, shu sababli ularni aniqlash uchun ma'lumotlarning o'zidan o'rganadigan usullar kerak.

Anomaliyalarni aniqlash usullari V. Chandola va hammualliflar tomonidan statistik, masofaga asoslangan, zichlikka asoslangan va tasniflovchi guruhlarga ajratilgan [2]. Zichlikka asoslangan LOF usuli [3] va bir sinfli tayanch vektorlar mashinasi [4] keng tarqalgan, lekin katta hajmdagi ma'lumotlarda hisoblash xarajati yuqori. F. Liu, K. Ting va Z. Zhou taklif etgan Isolation Forest algoritmi boshqacha g'oyaga tayanadi: anomal yozuvlar kam va boshqalardan farqli bo'lgani uchun tasodifiy bo'linishlar orqali tezroq ajratiladi [5]. Algoritm belgilangan namunalarni talab qilmaydi, chiziqli murakkablikka ega va ko'p o'lchamli ma'lumotlarda barqaror ishlaydi [6]. Chuqur o'qitishga asoslangan usullar [7] murakkab naqshlarni yaxshiroq aniqlasa-da, ko'proq ma'lumot va hisoblash resursi talab qiladi. Shu sababli dastlabki bosqich uchun Isolation Forest maqbul tanlov hisoblanadi.

Algoritmning yaroqliligini dastlabki baholash uchun sintetik tajriba o'tkazildi. 2000 ta yozuvdan iborat to'plam to'rt belgi bilan tavsiflandi: fuqarolik o'zgarishlari soni, oxirgi o'zgarishdan hisobot sanasigacha o'tgan kunlar, qabul paytidagi yosh va viza tugash sanasi bilan holat o'rtasidagi farq. To'plamga 60 ta (3 %) shubhali yozuv kiritildi. Tajriba scikit-learn kutubxonasida [8] 200 ta daraxt bilan bajarildi, kutilayotgan anomaliyalar ulushi parametri uch qiymatda sinaldi (1-jadval).

1-jadval. Sintetik to'plamda olingan natijalar

Anomaliyalar ulushi parametri	Aniqlik (precision)	To'liqlik (recall)	F1
0,02	1,000	0,667	0,800
0,03	0,900	0,900	0,900
0,05	0,600	1,000	0,750

Parametr haqiqiy anomaliyalar ulushiga yaqin tanlanganda eng yaxshi muvozanat olindi (F1 = 0,90). Parametr kichik bo'lsa algoritm faqat eng yaqqol holatlarni topadi, katta bo'lsa ortiqcha signallar ko'payadi. Xodim vaqti cheklangan sharoitda aniqlikni yuqori ushlab turish muhimroq.

Xulosa qilib aytganda, Isolation Forest algoritmi xorijiy talabalar ma'lumotlarida oldindan belgilangan qoidalarga tushmaydigan shubhali yozuvlarni aniqlash uchun istiqbolli vosita hisoblanadi. Natijalar sintetik ma'lumotlarga asoslangani uchun dastlabki xarakterga ega. Keyingi bosqichda

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

algoritm real, anonimlashtirilgan ma'lumotlarda sinaladi va qoidaga asoslangan tekshiruv bilan birlashtirilgan gibrid sxema ko'rinishida qo'llanadi.

Foydalanilgan adabiyotlar

1. O'zbekiston Respublikasi Prezidentining 2024-yil 14-oktabrdagi PQ-358-son "2030-yilgacha sun'iy intellekt texnologiyalarini rivojlantirish strategiyasini tasdiqlash to'g'risida"gi qarori.
2. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey // ACM Computing Surveys. – 2009. – Vol. 41, No. 3. – Article 15. – DOI: 10.1145/1541880.1541882.
3. Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. LOF: Identifying density-based local outliers // Proceedings of the ACM SIGMOD International Conference on Management of Data. – 2000. – P. 93–104. – DOI: 10.1145/342009.335388.
4. Schölkopf B., Platt J.C., Shawe-Taylor J., Smola A.J., Williamson R.C. Estimating the support of a high-dimensional distribution // Neural Computation. – 2001. – Vol. 13, No. 7. – P. 1443–1471.
5. Liu F.T., Ting K.M., Zhou Z.-H. Isolation forest // Proceedings of the 8th IEEE International Conference on Data Mining (ICDM). – 2008. – P. 413–422. – DOI: 10.1109/ICDM.2008.17.
6. Liu F.T., Ting K.M., Zhou Z.-H. Isolation-based anomaly detection // ACM Transactions on Knowledge Discovery from Data. – 2012. – Vol. 6, No. 1. – Article 3. – DOI: 10.1145/2133360.2133363.
7. Pang G., Shen C., Cao L., van den Hengel A. Deep learning for anomaly detection: A review // ACM Computing Surveys. – 2021. – Vol. 54, No. 2. – Article 38. – DOI: 10.1145/3439950.
8. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.