ANOMALIYA ANIQLASH (ISOLATION FOREST)

Alijonov Diyorbek Do'stbek o'g'li
diyorbekalijanov023@gmail.com
+998502255220

Annotatsiya

Ushbu akademik maqola anomaliyalarni aniqlashning muhim usuli bo'lmish Isolation Forest (Izolyatsiya O'rmoni) algoritmini chuqur tahlil qiladi. Maqolada anomaliya aniqlash muammosining global ahamiyati, mavjud cheklovlar va Isolation Forest algoritmining nazariy asoslari, ishlash mexanizmlari hamda boshqa an'anaviy usullar bilan solishtirma tahlili ko'rib chiqiladi. Algoritmning samaradorligi, hisoblash tezligi, xotira sarfi, yuqori o'lchovli ma'lumotlar bilan ishlash qobiliyati kabi afzalliklari batafsil o'rganiladi. Shuningdek, uning cheklovlari, ayniqsa yuqori o'lchovli ma'lumotlardagi matematik asoslarining murakkabligi va giperparametrlarga sezgirligi muhokama etiladi. Maqola, firibgarlikni aniqlash va tarmoq xavfsizligini ta'minlash kabi real dunyo ilovalaridagi rolini yoritib, kelajakdagi tadqiqot yo'nalishlari bo'yicha takliflar beradi.

Kalit so'zlar: Anomaliya aniqlash, Isolation Forest, noxos kuzatuv, mashinaviy o'rganish, kuzatuvsiz o'rganish, ma'lumotlar tahlili

Abstract

This academic article provides an in-depth analysis of the Isolation Forest algorithm, a crucial method for anomaly detection. It examines the global importance of anomaly detection, existing limitations, and the theoretical foundations, operational mechanisms, and comparative analysis of Isolation Forest with traditional methods. The article elaborates on the algorithm's advantages, such as its efficiency, computational speed, memory usage, and ability to handle high-dimensional data. It also discusses its limitations, particularly the complexity of its mathematical underpinnings in high-dimensional data and its sensitivity to hyperparameters. The paper highlights its role in real-world applications like fraud detection and network security, offering suggestions for future research directions.

Keywords: Anomaly detection, Isolation Forest, outlier detection, machine learning, unsupervised learning, data analysis

Аннотация

Данная академическая статья глубоко анализирует алгоритм Isolation Forest, важный метод обнаружения аномалий. В статье рассматривается глобальное значение проблемы обнаружения аномалий, существующие ограничения, а также теоретические основы, механизмы работы и сравнительный анализ алгоритма Isolation Forest с другими традиционными

методами. Подробно изучаются преимущества алгоритма, такие как его эффективность, скорость вычислений, потребление памяти и способность работать с высокоразмерными данными. Также обсуждаются его ограничения, в частности, сложность математических основ в высокоразерных данных и чувствительность к гиперпараметрам. Статья освещает роль алгоритма в реальных приложениях, таких как обнаружение мошенничества и обеспечение сетевой безопасности, а также предлагает направления для будущих исследований.

Ключевые слова: Обнаружение аномалий, Isolation Forest, выбросы, машинное обучение, обучение без учителя, анализ данных

Kirish

Anomaliyalarni aniqlash, mashinaviy o'rganish sohasidagi eng muhim muammolardan biri bo'lib, firibgarlikni aniqlash, tarmoq himoyasi, uskunalar ishdan chiqishining oldini olish va tibbiy diagnostika kabi keng doiradagi amaliy masalalarni hal qilishda hal qiluvchi ahamiyatga ega. Global miqyosda raqamli ma'lumotlarning hajmi eksponensial ravishda o'sib borayotganligi sababli, ushbu ma'lumotlar oqimida yashiringan g'ayritabiiy yoki kutilmagan naqshlarni tez va samarali aniqlashga bo'lgan ehtiyoj har qachongidan ham dolzarbdir. Anomaliyalar, odatda, noyob, kam uchraydigan va yashirin ma'lumot nuqtalari bo'lib, ular odatdagi ma'lumot namunalaridan sezilarli darajada farq qiladi. Ushbu g'ayritabiiy kuzatuvlar ko'pincha ma'lumotlar to'plamining umumiy xususiyatlariga zid keladi va ba'zan jiddiy muammolar, tizimdagi nuqsonlar yoki kutilmagan hodisalarning dastlabki belgilari bo'lishi mumkin. An'anaviy bashoratli modellar, o'tmishdagi belgilangan ma'lumotlarga tayanib ishlashi sababli, yangi va kutilmagan anomaliyalarni aniqlashda jiddiy qiyinchiliklarga duch keladi. Bu holat, ayniqsa, anomaliyalar uchun oldindan belgilangan yorliqlar kam bo'lgan yoki umuman mavjud bo'lmagan stsenariylarda yaqqol namoyon bo'ladi, chunki anomaliyalar ta'rifi bo'yicha nodir bo'ladi va ularni o'qitish uchun etarli yorliqli ma'lumotlarga ega bo'lish qiyin. Shu sababli, ma'lumotlarga oldindan yorliq qo'yishni talab qilmaydigan va odatiy bo'lmagan xususiyat qiymatlarini aniqlashga qodir bo'lgan Isolation Forest (Izolyatsiya O'rmoni) kabi kuzatuvsiz usullar ustunlikka ega bo'ladi. Isolation Forest algoritmi noxos kuzatuvlarni aniqlash uchun samarali va tezkor yechim taklif etadi. Uning asosiy printsipi shundan iboratki, anomaliyalar ma'lumotlar to'plamida kamroq va o'ziga xos bo'lganligi sababli, ularni normal ma'lumot nuqtalariga qaraganda kamroq tasodifiy bo'linishlar orqali ajratib olish ancha oson bo'ladi. Bu xususiyat, uni boshqa anomaliya aniqlash usullaridan ajratib turadi va uni tez o'sib borayotgan ma'lumotlar hajmi sharoitida, ayniqsa yuqori o'lchovli ma'lumotlar to'plamlari uchun juda qulay vositaga aylantiradi. Ushbu maqola Isolation Forest algoritmining nazariy asoslarini, ishlash

mexanizmlarini, afzalliklari va cheklovlarini chuqur tahlil qilib, uning akademik va amaliy ahamiyatini atroflicha yoritadi.

Mavzuga oid adabiyotlar tahlili

Anomaliyalar — bu odatdagi namunalardan sezilarli darajada chetga chiquvchi ma'lumotlar nuqtalari. Ular uch asosiy turga bo'linadi: nuqtali anomaliyalar, kontekstli anomaliyalar va kollektiv anomaliyalar. Nuqtali anomaliyalar, masalan, biror bir bank hisobidan kutilmaganda juda katta miqdordagi pul o'tkazmasi, ma'lum bir kuzatuvning qolgan barcha kuzatuvlardan butunlay farq qilishini anglatadi. Kontekstli anomaliyalar ma'lum bir kontekstda g'ayritabiiy bo'lishi mumkin, ammo boshqa kontekstda normal hisoblanishi mumkin. Kollektiv anomaliyalar esa alohida nuqtalar emas, balki ma'lumotlar to'plamining bir qismi sifatida g'ayritabiiy xatti-harakatni ko'rsatadigan kuzatuvlar guruhini bildiradi. Anomaliyalar ko'pincha noyob, yashirin va ularga tegishli yorliqlar kamligi sababli, an'anaviy tasniflash modellarining ularni aniqlashda samaradorligi past bo'ladi. Shu sababli, Isolation Forest kabi kuzatuvsiz usullar belgilangan ma'lumotlarga ehtiyoj sezmasdan noxos kuzatuvlarni samarali aniqlashga qodir bo'lgan muhim yechim hisoblanadi.

Isolation Forest algoritmi Fei Tony Liu, Kai Ming Ting va Zhi-Hua Zhou tomonidan 2008 yilda joriy etilgan bo'lib, u kuzatuvsiz mashinaviy o'rganish usuli sifatida anomaliyalarni aniqlashga mo'ljallangan. Algoritmning asosiy g'oyasi shundan iboratki, anomaliyalar ma'lumotlar to'plamida kamdan-kam uchraydigan va boshqa nuqtalardan farqli bo'lganligi sababli, ularni normal ma'lumot nuqtalariga qaraganda kamroq tasodifiy bo'linishlar orqali ajratish ancha osonroq ekanligiga asoslanadi. Bu nazariya ikkilik qidirish daraxtlarining ishlash printsipiga o'xshash bo'lib, bunda noyob elementlar daraxtning yuqori qismida tezroq ajratib olinadi. Algoritm Isolation Trees (Izolyatsiya Daraxtlari) deb ataluvchi ko'plab daraxtlarni qurish orqali ishlaydi. Har bir izolyatsiya daraxti tasodifiy tanlangan ma'lumotlar quyi namunasidan foydalanib, tasodifiy xususiyat va tasodifiy bo'linish qiymati asosida ma'lumotlarni rekursiv ravishda bo'ladi. Bu jarayon har bir barg tugunida bitta misol qolguncha yoki maksimal chuqurlikka erishilguncha davom etadi. Aniqlangan anomaliya birgina bo'linish orqali boshqa ma'lumotlardan ajratilishi mumkin bo'lsa, normal ma'lumot nuqtalarini ajratish uchun ko'plab kichik bo'linishlar kerak bo'ladi, chunki ular bir-biriga yaqin joylashgan bo'ladi.

Ma'lumot nuqtasi uchun anomaliya bali (anomaly score) daraxtlardagi uning o'rtacha yo'l uzunligi (path length) asosida aniqlanadi. Yo'l uzunligi, daraxt ildizidan ma'lumot nuqtasini izolyatsiya qilish uchun zarur bo'lgan bo'linishlar sonini bildiradi. Anomaliyalar uchun bu o'rtacha yo'l uzunligi qisqaroq bo'ladi. Algoritm har bir izolyatsiya daraxtida h(x) deb nomlanuvchi yo'l uzunligini hisoblaydi, keyin esa barcha daraxtlar bo'yicha bu uzunliklarning o'rtachasini oladi. Ushbu o'rtacha yo'l uzunligi, ikkilik qidirish daraxti nazariyasidan kelib chiqqan, namuna hajmi n ga asoslangan

kutilgan o'rtacha yo'l uzunligi c(n) bilan normallashtiriladi. Natijada olingan anomaliya bali 0 dan 1 gacha diapazonda bo'ladi, bunda 1 ga yaqin qiymatlar aniq anomaliyani, 0 ga yaqin qiymatlar esa normal ma'lumotni bildiradi. Yuqori ballar, nuqtaning qisqaroq izolyatsiya yo'liga ega bo'lganligi sababli, uning anomaliya bo'lish ehtimoli yuqori ekanligini ko'rsatadi.

Isolation Forest ning muhim afzalliklaridan biri uning hisoblash samaradorligidir. Algoritm o'qitish uchun chiziqli vaqt murakkabligiga (O(n log n)) va bashorat qilish uchun O(log n) murakkablikka ega, bu uni katta hajmli ma'lumotlar to'plamlari uchun miqyosli qiladi. Uning xotira talablari ham juda past bo'lib, bu cheklangan resurslarga ega tizimlarda katta afzallik beradi. Isolation Forest, yuqori o'lchovli ma'lumotlar bilan ishlashda ayniqsa samarali bo'lib, "o'lchamli la'nat" muammosini chetlab o'tishga yordam beradi, chunki u barcha xususiyatlar orasidagi masofalarni hisoblamasdan, balki ularning tasodifiy kichik to'plamlarida ishlashga tayanadi. U ma'lumotlarning taqsimoti haqida hech qanday taxminlarni talab qilmaydi va belgilangan o'qitish ma'lumotlariga muhtoj emas, bu esa uni kuzatuvsiz anomaliya aniqlash vazifalari uchun ideal qiladi. Algoritm ko'p modallikdagi taqsimotlarga va anomaliyalar bilan ifloslangan o'qitish ma'lumotlariga nisbatan barqarorlikni namoyish etadi, bu esa boshqa aniqlash usullarini qiyinlashtiradigan muhim muammolardir.

Biroq, Isolation Forest ning ba'zi cheklovlari ham mavjud bo'lib, ularni algoritm qo'llanilganda hisobga olish lozim. Uning ishlashi giperparametr sozlamalariga, masalan, o'rmondagi daraxtlar soni (nestimators) va har bir daraxtni qurish uchun foydalaniladigan ma'lumotlar quyi namunasining hajmi (maxsamples) ga sezgir bo'lishi mumkin. Natijalarning barqarorligini ta'minlash uchun ushbu giperparametrlarni to'g'ri sozlash, odatda, tajriba va sinovlarni talab qiladi. Algoritm anomaliyalar haqiqatan ham noyob (ideal holda, ma'lumotlarning 10% dan kamrog'i) va normal ma'lumotlardan yaxshi ajratilgan deb taxmin qiladi. Agar anomaliyalar klasterlangan bo'lsa yoki anomaliya darajasi yuqori bo'lsa, algoritmning ishlashi sezilarli darajada yomonlashishi mumkin. Shuningdek, uning tasodifiy tabiati, ayniqsa kichik ma'lumotlar to'plamlarida, turli ijrolar orasida natijalarning nomuvofiqligiga olib kelishi mumkin.

An'anaviy anomaliya aniqlash usullari ko'pincha ma'lumot nuqtalari orasidagi masofaga (masalan, k-eng yaqin qo'shni (k-NN) asosidagi usullar) yoki mahalliy zichlikka (masalan, LOF - Local Outlier Factor) asoslanadi. Bunday usullar ma'lumotlarning zich klasterlarini aniqlash va ulardan statistik jihatdan uzoqda joylashgan nuqtalarni anomaliya deb belgilashga harakat qiladi. Biroq, ular yuqori o'lchovli ma'lumotlar bilan ishlashda "o'lchamli la'nat" muammosiga jiddiy ravishda duch kelishi mumkin, bunda nuqtalar orasidagi masofa tushunchasi o'z ahamiyatini yo'qotadi va barcha nuqtalar deyarli bir xil darajada uzoqlashadi. Natijada, hisoblash xarajatlari sezilarli darajada oshadi va algoritmning samaradorligi kamayadi. Isolation

Forest esa bu usullardan tubdan farq qiladi, chunki u normal ma'lumot profilini yaratish o'rniga, anomaliyalarni to'g'ridan-to'g'ri ajratishga e'tibor qaratadi. Bu uni yuqori o'lchovli ma'lumotlar uchun ayniqsa samarali qiladi va "o'lchamli la'nat" muammosidan mohirlik bilan qochishga yordam beradi, chunki u faqat bitta tasodifiy tanlangan xususiyat bo'yicha bo'linishlarni amalga oshiradi, bu esa hisoblash murakkabligini kamaytiradi.

Isolation Forest algoritmining amaliy qo'llanilishi juda keng va xilma-xildir, ko'plab sohalarda muhim amaliy qiymatga ega. U firibgarlikni aniqlashda, masalan, moliyaviy sohada kredit karta firibgarliklari, sug'urta firibgarliklari yoki bank hisoblaridagi g'ayritabiiy harakatlarni aniqlashda juda samarali. Tarmoq xavfsizligi sohasida, u tarmoqqa kirish xurujlari, zararli dasturlar faoliyati, DoS hujumlari yoki g'ayritabiiy trafik shakllari kabi noxos hodisalarni aniqlash uchun ishlatilishi mumkin. Sanoatda uskunalar diagnostikasi va texnik xizmat ko'rsatishda, u mashina qismlarining normal ishlashidan og'ishlarni aniqlash orqali potentsial nosozliklarni oldindan bashorat qilishga yordam beradi, shu bilan ishlab chiqarishda to'xtalishlarni minimallashtiradi va xizmat muddatini uzaytiradi. Tibbiyotda, u tibbiy tasvirlarda noodatiy patologiyalarni, bemor ma'lumotlarida g'ayritabiiy o'zgarishlarni aniqlashda yordam berishi mumkin.

Tadqiqot metodologiyasi

Isolation Forest algoritmining ishlash tamoyillarini chuqurroq tushunish va uning cheklovlarini aniqlash uchun uning matematik asoslarini ko'rib chiqish maqsadga muvofiqdir. Isolation Random Forest (IRF) algoritmi, ayniqsa, Katta Ma'lumotlar (Big Data) sharoitida masofa hisob-kitoblaridan yoki ehtimollik taqsimotini parametrlashdan voz kechganligi sababli keng qo'llaniladi. Bir o'lchovli ma'lumotlar uchun o'tkazilgan jiddiy matematik tahlillar IRF usulining konvergentsiyasini isbotladi va uning nisbiy masofalar bilan aniqlangan anomaliyalarni aniqlashdagi yakuniyligini tasdiqladi. Tadqiqotlar shuni ko'rsatadiki, izolyatsiya daraxtidagi kutilgan daraxt balandligi ma'lumot nuqtasining nisbiy masofasi bilan asimptotik tarzda bog'liq bo'lishi mumkin. Bu, bir o'lchovli holatlarda algoritmning nazariy jihatdan mustahkam va ishonchli ekanligini anglatadi.

Biroq, ushbu tahlil IRF usulining yuqori o'lchovli ma'lumotlarda anomaliyalarni aniqlashda har doim ham yakuniy emasligini ko'rsatadigan muhim cheklovni ham ochib berdi. Qarshi misollar orqali tadqiqotlar, ikki o'lchovli tekislikdagi bir xil masofadagi nuqtalar uchun bu yakuniylikning yo'qligini ko'rsatdi, bu esa usulning bunday stsenariylarda anomaliya aniqlash uchun matematik sifat sertifikatiga ega emasligini bildiradi. Ya'ni, yuqori o'lchovli fazoda barcha nuqtalar bir-biridan deyarli bir xil darajada uzoq bo'lishi mumkin bo'lgan "o'lchamli la'nat" ta'siri ostida, Isolation Forest ning "anomaliyalar kam bo'linishda ajratiladi" degan asosiy printsipi har doim ham to'g'ri kelmasligi mumkin. Bu nazariy cheklash algoritmning yuqori o'lchovli, zich

klasterlangan yoki bir-biriga yaqin anomaliyalar mavjud bo'lgan murakkab ma'lumotlar to'plamlarida ishlash qobiliyatiga ta'sir qilishi mumkin. Natijada, ushbu cheklovlarni bartaraf etish uchun to'liq asoslantirilgan matematik asosga ega IRF usulining muqobil versiyasi taklif qilingan.

Amaliy jihatdan, Isolation Forest ni qo'llashda giperparametrlarni, xususan, nestimators (o'rmondagi daraxtlar soni) va maxsamples (har bir daraxtni qurish uchun foydalaniladigan ma'lumotlar quyi namunasining hajmi) ni ehtiyotkorlik bilan sozlash zarur. Odatda, ko'proq daraxtlar yanada barqaror natijalarga olib keladi va tasodifiylikni kamaytiradi, ammo hisoblash xarajatlarini oshiradi. maxsamples parametri esa ma'lumotlar to'plamidagi anomaliyalar zichligi va hajmini hisobga olgan holda tanlanishi kerak. Yuqorida qayd etilgan nazariy va amaliy cheklovlar, ayniqsa, anomaliyalar juda noyob bo'lmaganida yoki ular klasterlarni hosil qilganida, algoritmning samaradorligini pasaytirishi mumkin. Shuning uchun, algoritmning ishlashi ma'lumotlar to'plamining o'ziga xos xususiyatlariga juda bog'liq bo'ladi.

Xulosa

Isolation Forest algoritmi anomaliyalarni aniqlash sohasida kuchli, samarali va miqyosli yechimni taqdim etadi, ayniqsa yuqori o'lchovli va yorliqlanmagan ma'lumotlar to'plamlari bilan ishlashda o'zining ustunligini ko'rsatadi. Uning asosiy kuchi anomaliyalarni normal ma'lumot nuqtalariga qaraganda kamroq bo'linishlar orqali izolyatsiya qilish qobiliyatiga asoslanadi, bu esa unga hisoblash samaradorligi va xotira tejamkorligini beradi. Algoritm firibgarlikni aniqlash, tarmoq xavfsizligi va tizim monitoringi kabi turli xil real dunyo ilovalarida muvaffaqiyatli qo'llanilgan. U ma'lumotlar taqsimoti haqida aniq taxminlarni talab qilmasligi, ya'ni parametrik bo'lmagan yondashuv ekanligi va "o'lchamli la'nat" ga nisbatan mustahkamligi bilan ajralib turadi. Ushbu xususiyatlar uni ko'plab an'anaviy usullarga nisbatan afzalroq qiladi, ayniqsa murakkab va katta ma'lumotlar to'plamlari bilan ishlashda.

Biroq, algoritmining ba'zi cheklovlari mavjud bo'lib, ular kelajakdagi tadqiqotlar uchun muhim yo'nalishlarni belgilaydi. Jumladan, giperparametrlarni sozlashga sezgirlik, anomaliyalar noyob va yaxshi ajratilgan deb faraz qilish, shuningdek, yuqori o'lchovli ma'lumotlardagi anomaliyalar uchun matematik yakuniylikning har doim ham ta'minlanmasligi uning asosiy cheklovlaridandir. Kelajakdagi tadqiqotlar ushbu cheklovlarni bartaraf etishga qaratilishi mumkin. Masalan, o'lchovlarning yuqori sonidagi anomaliyalarni yanada ishonchli aniqlash usullarini ishlab chiqish, ayniqsa, "o'lchamli la'nat" ta'sirini kamaytiradigan yangi o'lchovlilikni kamaytirish texnikalari bilan integratsiya qilish mumkin. Shuningdek, klasterlangan anomaliyalar yoki yuqori anomaliya darajasi bo'lgan stsenariylar uchun algoritmning moslashuvchanligini oshirish, masalan, adaptiv bo'linish strategiyalari yoki mahalliy kontekstni hisobga oluvchi yondashuvlarni joriy etish orqali erishish mumkin. Isolation Forest, shubhasiz, ma'lumotlar tahlilchilar va tadqiqotchilar uchun anomaliyani aniqlashda qimmatli

vosita bo'lib qoladi, ammo uning to'liq salohiyatini ochish va universal qo'llanilishini kengaytirish uchun doimiy takomillashtirish va chuqur tahlillar talab etiladi.

Foydalanilgan adabiyotlar

[1] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Izolyatsiyaga asoslangan anomaliyalarni aniqlash." ACM Ma'lumotlardan Bilim Keshfiyoti bo'yicha Tranzaktsiyalar (TKDD), vol. 6, no. 1, 2012, pp. 1-39.

[2] Chandola, Varun, Vipin Kumar, and Anuj Karpatne. "Anomaliyani aniqlash: Sharh." ACM Hisoblash Sharhlari (CSUR), vol. 41, no. 3, 2009, pp. 1-58.

[3] Chandola, Varun, Chih-Chung Chee, and Vipin Kumar. Anomaliya aniqlash tamoyillari va usullari. New York: Springer, 2010.

[4] Aggarwal, Charu C. Chetki qiymatlar tahlili. New York: Springer, 2017.

[5] Zimek, Arthur, Erich Schubert, and Hans-Peter Kriegel. "Yuqori o'lchamli raqamli ma'lumotlarda nazoratsiz anomaliyalarni aniqlash bo'yicha tadqiqot." Statistik Tahlil va Ma'lumotlarni Qazib Olish: ASA Ma'lumotlar Fanlari Jurnali, vol. 5, no. 5, 2012, pp. 363-397.