

SMART MA'LUMOTLARI ASOSIDA DISK NOSOZLIKLARINI MASHINA O'RGANISH MODELLARI YORDAMIDA BASHORATLASH

Xolmurov Shamurod Abdihamidovich

Termiz iqtisodiyot va servis universiteti magistranti

E-mail: moviyosmon77@gmail.com

Annotatsiya. Maqolada SMART monitoring ma'lumotlari asosida disk nosozligini bashoratlashda turli mashina o'rganish modellarining samaradorligi baholandi. Muvozanatsiz sinflarga ega 40 000 disk-kunlik namunaviy ma'lumotlar to'plamida Random Forest, XGBoost, SVM, Isolation Forest, Autoencoder va LSTM modellari qiyoslandi. Natijalar statik belgilar uchun Random Forest eng muvozanatli F1 ko'rsatkichini, vaqt qatorlarida esa LSTM yuqori salohiyatni namoyon etganini ko'rsatdi. Model tanlash shartlari hamda uch qatlamli konseptual gibriddan arxitektura asoslandi.

Kalit so'zlar: SMART monitoring, disk nosozligi, prediktiv texnik xizmat, mashina o'rganish, Random Forest, XGBoost, anomaliyani aniqlash, LSTM, sinflar muvozanatsizligi, F1-o'lcham.

Аннотация. В статье оценена эффективность моделей машинного обучения при прогнозировании отказов дисков по данным SMART-мониторинга. На модельном наборе из 40 000 наблюдений «диск-день» с несбалансированными классами сопоставлены Random Forest, XGBoost, SVM, Isolation Forest, Autoencoder и LSTM. Результаты показывают наиболее сбалансированное значение F1 у Random Forest для статических признаков и высокий потенциал LSTM при наличии временных рядов. Обоснованы условия выбора моделей и предложена концептуальная трехуровневая гибридная архитектура.

Ключевые слова: SMART-мониторинг, отказ диска, предиктивное обслуживание, машинное обучение, Random Forest, XGBoost, обнаружение аномалий, LSTM, дисбаланс классов, F1-мера.

Abstract. This article evaluates machine-learning models for predicting disk failures from SMART monitoring data. Random Forest, XGBoost, SVM, Isolation Forest, Autoencoder, and LSTM were compared using a model dataset of 40,000 disk-day observations with imbalanced classes. The results indicate that Random Forest provides the most balanced F1 performance for static features, whereas LSTM has strong potential when temporal sequences are available. Model-selection conditions are identified, and a conceptual three-layer hybrid architecture is proposed for practical predictive maintenance.

Keywords: SMART monitoring, disk failure, predictive maintenance, machine learning, Random Forest, XGBoost, anomaly detection, LSTM, class imbalance, F1-score.

KIRISH

Ma'lumotlarni saqlash qurilmalari raqamli infratuzilmaning asosiy tarkibiy qismlaridan biridir. Serverlar, ma'lumotlar markazlari va bulutli xizmatlarning uzluksiz ishlashi diskning texnik holatiga bevosita bog'liq. HDD yoki SSD qurilmasining kutilmagan ishdan chiqishi nafaqat uskunani almashtirish xarajatini, balki xizmatning to'xtashi, ma'lumotlarning buzilishi yoki yo'qolishi xavfini ham yuzaga keltiradi. Shu sababli reaktiv ta'mirlashdan nosozlik xavfini oldindan baholashga asoslangan prediktiv texnik xizmatga o'tish amaliy ahamiyat kasb etmoqda.

SMART (Self-Monitoring, Analysis and Reporting Technology) disk ichidagi diagnostik ko'rsatkichlarni muntazam qayd etadi. Biroq an'anaviy SMART nazorati ko'pincha har bir atributni ishlab chiqaruvchi belgilagan statik chegara bilan alohida taqqoslaydi. Bir nechta atributning bir vaqtdagi kichik o'zgarishi nosozlik xavfini bildirishi mumkin bo'lsa-da, ularning hech biri yakka holda ogohlantirish chegarasiga yetmasligi mumkin. Murray, Hughes va Kreutz-Delgado disk nosozligi kam uchraydigan hodisa ekanini va uni bashoratlash uchun maxsus mashina o'rganish yondashuvlari zarurligini ko'rsatgan [1, 783–816].

Mavjud tadqiqotlarda ko'pincha bitta model yoki bir turdagi ma'lumotga urg'u beriladi. Nazorat ostidagi klassifikatorlar yorliqlangan ma'lumotlarda yuqori natija ko'rsatsa, anomaliyani aniqlash usullari nosozlik yorliqlari kam bo'lgan muhitda foydalidir. Vaqt qatori modellari esa SMART atributlarining asta-

sekin o'zgarishini hisobga oladi. Ushbu farqlar birgina "eng yaxshi model"ni tanlashdan ko'ra, modelni ma'lumotlar tabiati va amaliy xatolar narxiga mos tanlash zarurligini ko'rsatadi.

Maqolaning maqsadi SMART atributlari asosida disk nosozligini bashoratlashda Random Forest, XGBoost, SVM, Isolation Forest, Autoencoder va LSTM modellarining qiyosiy samaradorligini baholashdan iborat. Asosiy tadqiqot savoli quyidagicha: qaysi model musbat bashorat aniqligi, to'liqlik va yolg'on signallar o'rtasidagi eng maqbul muvozanatni ta'minlaydi? Gipotezaga ko'ra, mashina o'rganish modellari atributlar o'rtasidagi chiziqli bo'lmagan bog'liqliklarni hisobga olib, oddiy chegaraviy nazoratga qaraganda mazmunliroq bashorat beradi; model samaradorligi esa yorliqlar va vaqt qatori mavjudligiga bog'liq.

TADQIQOT METODOLOGIYASI

Tadqiqot empirik, qiyosiy va hisoblash eksperimenti shaklida tashkil etildi. Disk nosozligi $y \in \{0,1\}$ chiqishli binar klassifikatsiya hamda yorliq mavjud bo'lmaganda anomaliyani aniqlash masalasi sifatida formalashtirildi. Asosiy tajriba real disk statistikasi xususiyatlariga mos ravishda generatsiya qilingan 40 000 disk-kun kuzatuvdan iborat namunaviy to'plamda o'tkazildi. Unda 39 000 sog'lom va 1 000 nosoz kuzatuv mavjud. Shu bois natijalarni to'liq real ishlab chiqarish muhiti uchun tashqi validatsiyasiz umumlashtirish mumkin emas.

Har bir kuzatuv sakkizta belgi bilan ifodalandi: asosiy SMART 5, 187, 188, 197 va 198 atributlariga SMART 1, SMART 9 va SMART 194 qo'shildi. Yetishmayotgan yoki mantiqan noto'g'ri qiymatlarni aniqlash, belgilarning diapazonlarini tekshirish va standartlashtirish dastlabki ishlov berishning asosiy bosqichlarini tashkil etdi. Standartlashtirish parametrlari faqat o'qitish qismida hisoblanib, sinov qismiga tatbiq etildi; bu ma'lumot sizib chiqishini cheklaydi.

Ma'lumotlar `random_state=42` bilan stratifikatsiyalangan holda 70 foiz o'qitish (28 000 kuzatuv) va 30 foiz sinov (12 000 kuzatuv) to'plamiga ajratildi. Sinflar muvozanatsizligi Random Forest va SVMda muvozanatlangan sinf vaznlari, XGBoostda `scale_pos_weight` orqali hisobga olindi. Isolation Forest va Autoencoder faqat sog'lom o'qitish kuzatuvlarida normal holatni o'rgandi. Random Forestda 200 ta daraxt, XGBoostda 200 ta daraxt, maksimal chuqurlik 5 va o'rganish tezligi 0,1 qo'llandi. SVM RBF yadrosi bilan qurildi. Autoencoder qayta tiklash xatosi, LSTM esa 14 kunlik ketma-ketlik asosida baholandi.

Dasturiy tajriba Python, scikit-learn, XGBoost va TensorFlow/Keras vositalari yordamida bajarildi. Baholashda umumiy to'g'rilik (accuracy), musbat bashorat aniqligi (precision), to'liqlik (recall), F1-o'lcham, ROC-AUC hamda yolg'on signal darajasi (FAR) qo'llandi. Muvozanatsiz to'plamda accuracy asosiy tanlov mezoni sifatida olinmadi; F1, recall va FAR birgalikda talqin qilindi. Hold-out sinovi qo'llangani, kross-validatsiya bajarilmagani, kutubxona versiyalari va apparat konfiguratsiyasi ko'rsatilmagani tadqiqotning takrorlanuvchanligi va umumlashtirilishini cheklaydi.

MAVZUGA OID ADABIYOTLAR TAHLILI

Murray, Hughes va Kreutz-Delgado qattiq disk nosozligini ko'p misolli o'rganish masalasi sifatida ko'rib, kam uchraydigan nosoz sinf sabab oddiy umumiy to'g'rilik yetarli emasligini asoslagan [1, 783–816]. Tomic va Pejic Bach anomaliyani aniqlash yondashuvining disk holatini yorliqlar cheklangan vaziyatda baholashdagi imkoniyatlarini ko'rsatgan [2, 1153–1172]. Ushbu ishlar disk nosozligini klassifikatsiya bilan cheklamasdan, anomaliya nuqtayi nazaridan ham o'rganish zarurligini ko'rsatadi.

Nazorat ostidagi modellar orasida Breiman taklif qilgan Random Forest bootstrap namunalari va tasodifiy belgilar tanloviga asoslangan ko'p daraxtli ansambl bo'lib, barqarorlik va belgilar muhimligini baholash imkonini beradi [3, 5–32]. Chen va Guestrin ishlab chiqqan XGBoost ketma-ket daraxtlar orqali oldingi xatolarni tuzatadi va regularizatsiyalangan maqsad funksiyasidan foydalanadi [4, 785–794]. Cortes va Vapnikning SVM modeli sinflar orasidagi maksimal chegarani izlaydi va RBF yadrosi yordamida chiziqli bo'lmagan bog'liqliklarni ifodalashi mumkin [7, 273–297]. Bu modellar yorliqlangan ma'lumot mavjud bo'lganda kuchli, ammo ularning precision va recall nisbatlari turlicha.

Liu, Ting va Zhou tomonidan taklif etilgan Isolation Forest anomal nuqtalarning tasodifiy daraxtlarda kamroq bo'linish orqali tez izolyatsiya qilinishiga tayangan [5, 413–422]. Autoencoder normal

kuzatuvni qayta tiklashni o'rganib, katta qayta tiklash xatosini anomaliya belgisi deb qabul qiladi; uning umumiy nazariy asoslari chuqur o'rganish adabiyotida yoritilgan [8, 493–516]. Bunday usullar yorliqsiz ishlashi bilan foydali, biroq anomaliya doimo nosozlikni anglatmagani sabab yolg'on signal ko'payishi mumkin.

Hochreiter va Schmidhuberning LSTM arxitekturasini uzoq vaqt bog'liqliklarini xotira yacheykasi va boshqaruv darvozalari orqali saqlaydi [6, 1735–1780]. Disklarda SMART atributlari nosozlik oldidan bosqichma-bosqich o'zgarsa, ketma-ketlik modeli statik kuzatuvda ko'rinmaydigan degradatsiya tendensiyasini aniqlashi mumkin. Biroq statik va vaqt qatori modellarini turli kirish tuzilmasi hamda turli eksperimental sharoitlarda olgan natijalari asosida mutlaq qiyoslash metodologik ehtiyotkorlikni talab qiladi.

TAHLIL VA NATIJALAR

1-jadval. Asosiy SMART atributlarining diagnostik mazmuni

Atribut	Inglizcha nomi	Diagnostik mazmuni
SMART 5	Reallocated Sectors Count	Nuqsonli deb topilib, zaxira sohaga ko'chirilgan sektorlar soni.
SMART 187	Reported Uncorrectable Errors	Xatoni tuzatish mexanizmi bartaraf eta olmagan xatolar soni.
SMART 188	Command Timeout	Belgilangan vaqtda bajarilmay, bekor qilingan buyruqlar soni.
SMART 197	Current Pending Sector Count	Qayta taqsimlanishni kutayotgan beqaror sektorlar soni.
SMART 198	Offline Uncorrectable	Oflayn tekshiruvda tuzatib bo'lmaydigan sektorlar soni.

Manba: Dissertatsiya ma'lumotlari asosida muallif ishlanmasi.

Ushbu atributlar disk yuzasi, o'qish-yozish operatsiyalari va tuzatib bo'lmaydigan sektorlar bilan bog'liq turli nosozlik alomatlarini ifodalaydi. Ularni birgalikda tahlil qilish alohida chegaradan oshmagan, biroq kombinatsiyasi xavf tug'diradigan holatlarni aniqlash imkonini beradi.

2-jadval. Mashina o'rganish modellarining qiyosiy samaradorligi

Model	Accuracy	Precision	Recall	F1	ROC-AUC	FAR
Random Forest	0,997	0,989	0,877	0,929	0,972	0,000*
XGBoost	0,994	0,841	0,917	0,877	0,983	0,004
SVM (RBF)	0,989	0,720	0,933	0,813	0,970	0,009
Isolation Forest	0,971	0,459	0,873	0,602	0,979	0,026
Autoencoder	0,973	0,477	0,890	0,621	0,979	0,025
LSTM	0,998	0,995	0,990	0,992	0,999	0,002

Manba: Dissertatsiyadagi eksperimental natijalar. *FAR uch xona aniqligida yaxlitlangan.

Barcha modellarda accuracy 0,97 dan yuqori bo'lsa-da, sog'lom kuzatuvlar 97,5 foizni tashkil qilgani sabab bu ko'rsatkich o'z-o'zicha model foydaliligini isbotlamaydi. Statik klassifikatorlar orasida Random Forest eng katta F1 (0,929) va precision (0,989)ni ta'minladi. Uning FAR qiymati jadvalda 0,000 gacha yaxlitlangan; chalkashlik matritsasida uchta yolg'on musbat holat mavjudligi sabab "yolg'on signal yo'q" degan xulosa chiqarilmaydi.

XGBoost recall bo'yicha Random Forestdan yuqori natija ko'rsatdi: haqiqiy nosozliklarning 91,7 foizi aniqlandi, biroq precision 0,841 gacha pasaydi. SVM recall bo'yicha statik modellar orasida eng yuqori qiymatga — 0,933 ga erishdi, lekin precision 0,720 va FAR 0,009 bo'ldi. Demak, nosozlikni o'tkazib yuborish juda qimmat bo'lgan muhitda SVM yoki XGBoost, keraksiz almashtirish va yolg'on ogohlantirish qimmat bo'lganda Random Forest ma'qulroq.

Isolation Forest va Autoencoderning F1 qiymatlari pastroq bo'lsa-da, ularning recall ko'rsatkichlari 0,873 va 0,890 ni tashkil etdi. Bu modellar yorliqlangan nosoz misollarini talab qilmasligi bilan ahamiyatli. Ular yangi muhitda dastlabki anomaliya filtri vazifasini bajarishi mumkin, ammo anomaliyani bevosita nosozlik deb qabul qilish yolg'on signallar ulushini oshiradi.

LSTM barcha ko'rsatkichlarda eng yuqori natijaga ega. Biroq u 14 kunlik vaqt ketma-ketligidan, boshqa modellar esa asosan statik belgilar vektoridan foydalangan. Shuning uchun LSTM natijasi uning mutlaq ustunligi emas, balki aniq degradatsiya tendensiyasi mavjud ketma-ket ma'lumotlarda vaqt tuzilmasidan foydalanishning yuqori salohiyati sifatida talqin qilinadi. Bu natijani real va shovqinli vaqt qatorlarida disk bo'yicha ajratilgan tashqi sinov bilan tekshirish zarur.

3-jadval. Random Forestda qaror chegarasining precision va recallga ta'siri

Chegara (τ)	Precision	Recall	F1
0,3	0,912	0,930	0,921
0,5	0,989	0,877	0,929
0,7	0,996	0,810	0,893
0,9	0,999	0,720	0,837

Manba: Dissertatsiyadagi chegaraviy qiymat eksperimenti.

Chegarani 0,3 gacha pasaytirish recallni 0,930 ga oshiradi, ammo precision 0,912 gacha kamayadi. Chegarani 0,9 gacha oshirish precisionni 0,999 ga olib chiqadi, lekin haqiqiy nosozliklarning 28 foizi o'tkazib yuboriladi. F1 bo'yicha 0,5 chegara ushbu namunaviy tajribada eng yaxshi muvozanatni ta'minlagan bo'lsa-da, amaliy chegara yolg'on signal va o'tkazib yuborilgan nosozlik xarajatlari asosida tanlanishi kerak.

NATIJALAR MUHOKAMASI

Natijalar barcha sharoitlar uchun yagona eng yaxshi model mavjud emasligini ko'rsatadi. Model tanlovi ma'lumotning yorliqlanganligi, vaqt ketma-ketligi, hisoblash resurslari va xatolar narxiga bog'liq. Random Forest statik, yorliqlangan ma'lumotda yuqori precision va F1, juda past FAR hamda belgilar muhimligini baholash imkoniyati bilan amaliy boshlang'ich model bo'la oladi. XGBoost yoki SVMning yuqoriroq recalli kritik nosozlikni o'tkazib yuborish keraksiz tekshiruvdan qimmatroq bo'lgan muhitda ustunlik beradi.

Yorliqlar cheklangan yangi muhitda Isolation Forest yoki Autoencoder dastlab normal ish rejimini o'rganib, shubhali kuzatuvlarni ajratishi mumkin. Bunda ularning ogohlantirishi yakuniy nosozlik hukmi emas, ikkinchi bosqich tekshiruviga kirish sifatida qaralishi kerak. Disk bo'yicha yetarli tarix mavjud bo'lsa, LSTM atributlarning tezligi va yo'nalishini hisobga oladi. Biroq vaqt qatori modeli uchun kelajakdagi ma'lumotning o'qitishga sizib kirishini oldini oladigan xronologik hamda disk bo'yicha guruhlangan validatsiya talab qilinadi.

Olingan natijalar asosida uch qatlamli konseptual arxitektura taklif etiladi. Birinchi qatlamda Autoencoder normal holatdan og'ishgan disklarni ajratadi. Ikkinchi qatlamda Random Forest yoki XGBoost shubhali kuzatuvni yorliqlangan qonuniyatlar asosida tasniflab, yolg'on signallarni kamaytiradi. Uchinchi qatlamda LSTM SMART ko'rsatkichlarining vaqt bo'yicha o'zgarishini baholab, xavf dinamikasini aniqlaydi. Ushbu arxitektura dissertatsiyada to'liq birlashtirilgan tizim sifatida eksperimental validatsiyadan o'tkazilmagan; shu bois u keyingi sinov uchun ilmiy-amaliy konsepsiya sifatida baholanadi.

Tadqiqotning asosiy cheklovi — namunaviy to'plam real disk statistikasi xususiyatlariga mos generatsiya qilingan bo'lsa-da, to'liq real ishlab chiqarish ma'lumotlari emasligidir. Disk ishlab chiqaruvchisi va modeli bo'yicha SMART semantikasidagi farqlar, konsepsiya siljishi, takroriy kuzatuvlarning bir diskka tegishliligi, kross-validatsiya va tashqi validatsiyaning yo'qligi natijalarni umumlashtirishni cheklaydi. LSTMning juda yuqori ko'rsatkichlari ayniqsa real ma'lumotlarda qayta tekshirilishi lozim.

XULOSA VA TAKLIFLAR

Tadqiqot savoliga javoban, statik SMART atributlari va yorliqlangan namunaviy ma'lumotlar sharoitida Random Forest precision, F1 va yolg'on signal darajasi o'rtasidagi eng maqbul umumiy

muvozanatni ta'minladi. Agar asosiy vazifa nosozliklarni imkon qadar kam o'tkazib yuborish bo'lsa, SVM yoki XGBoostning yuqoriroq recall ko'rsatkichlari afzal bo'lishi mumkin. Shu sababli tanlov bitta reyting emas, amaliy xarajat funksiyasiga asoslanishi kerak.

Eksperiment mashina o'rganish modellari bir nechta SMART atributlari o'rtasidagi murakkab bog'liqliklardan foydalanishi mumkinligi haqidagi gipotezani namunaviy to'plam doirasida qo'llab-quvvatladi. Biroq an'anaviy SMART chegaraviy nazoratining miqdoriy bazaviy natijasi alohida keltirilmagani sabab undan "sezilarli ustunlik" qat'iy statistik qiyos bilan to'liq isbotlangan deb bo'lmaydi.

Maqolaning ilmiy hissasi turli model sinflarini bir xil mezonlar asosida qiyoslash, ularning qo'llanish shartlarini ajratish va anomaliya filtri, klassifikator hamda vaqt qatori tahlilini birlashtiruvchi uch qatlamli konseptual arxitekturani asoslashdan iborat. Amaliyotda yorliqlangan tarixiy ma'lumotlar uchun Random Forest yoki XGBoost, yorliqlar kam bo'lganda Autoencoder yoki Isolation Forest, yetarli ketma-ket kuzatuv mavjud bo'lganda LSTMdan foydalanish tavsiya etiladi.

Tizimni joriy etishda qaror chegarasi tashkilotning yolg'on signal va o'tkazib yuborilgan nosozlik xarajatlariga moslashtirilishi, model yangi ma'lumotlarda davriy qayta o'qitilishi va ma'lumot taqsimotidagi siljish nazorat qilinishi kerak. Prediktiv modulni Zabbix yoki Prometheus kabi monitoring platformalariga ulash mumkin, biroq real vaqt integratsiyasi mustaqil sinov va xavfsizlik talablarini taqozo etadi.

Kelgusi tadqiqotlarda modellarni Backblaze va Alibaba kabi to'liq real ma'lumotlarda, disk bo'yicha guruhlangan va vaqtga asoslangan validatsiya bilan tekshirish, ishonch oralig'ini hisoblash, gibrid arxitekturani yagona tizim sifatida sinash hamda SHAP yoki LIME orqali qarorlarning izohlanishini oshirish zarur. Turli disk modellari bo'yicha tashqi validatsiya natijalarning amaliy ishonchligini belgilovchi asosiy bosqich bo'ladi.

FOYDALANILGAN ADABIYOTLAR RO'YXATI

1. Murray J.F., Hughes G.F., Kreutz-Delgado K. Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application // Journal of Machine Learning Research. – 2005. – Vol. 6. – P. 783–816.
2. Tomic A., Pejic Bach M. Anomaly Detection Model for Predicting Hard Disk Drive Failures // Applied Artificial Intelligence. – 2021. – Vol. 35, No. 15. – P. 1153–1172.
3. Breiman L. Random Forests // Machine Learning. – 2001. – Vol. 45, No. 1. – P. 5–32.
4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD Conference. – 2016. – P. 785–794.
5. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest // Proceedings of the 8th IEEE International Conference on Data Mining. – 2008. – P. 413–422.
6. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780.
7. Cortes C., Vapnik V. Support-Vector Networks // Machine Learning. – 1995. – Vol. 20, No. 3. – P. 273–297.
8. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge: MIT Press, 2016. – 800 p.