Anomália-észlelési algoritmusok többszempontú összehasonlítása

A multi-criteria comparison of anomaly detection algorithms

Zoltán CZAKO[1], Gheorghe SEBESTYEN[2], Anca HANGAN[3]

Számítástechnikai Tanszék, Kolozsvári Műszaki Egyetem, Koloszvár, Románia

[1]zoltan.czako@cs.utcluj.ro, [2]gheorghe.sebestyen@cs.utcluj.ro, [3]anca.hangan@cs.utcluj.ro

Abstract

Anomaly detection is the technique of finding out-of-the-ordinary occurrences inside datasets. For this goal, several anomaly detection techniques were created in many areas (e.g., economics, industrial processes, health, environmental monitoring, etc.) employing general artificial intelligence and signal processing algorithms. The original dataset features and user-selected hyperparameter parameters substantially affect these algorithm's performance. Researchers must try several approaches with varied hyperparameter settings since there is no ideal solution for a domain or dataset. Researchers should evaluate the performance of these algorithms from several perspectives. Most work is done on particular problem contexts or areas, and methods are assessed from one or a few viewpoints using "default" parameters without hyperparameter optimisation. This article discusses the results of a comprehensive anomaly detection benchmarking that evaluated 34 algorithms on numerous relevant datasets from a broad range of fields. Results are presented from various angles. Hyperparameter optimisation is used to determine the optimal parameters for each method and dataset in each test scenario. This paper discusses the pros and cons of several methods and presents a taxonomy of anomaly detection algorithms depending on the problem context and input data.

Keywords: artificial intelligence, anomaly detection, benchmarking, taxonomy of anomaly detection algorithms

Kivonat

Az anomália-észlelés az adathalmazokon belüli, a szokásostól eltérő előfordulások megtalálásának technikája. E cél érdekében számos területen (pl. közgazdaságtan, ipari folyamatok, egészségügy, környezetmonitoring stb.) mesterséges intelligencia és jelfeldolgozó algoritmusok alkalmazásával számos anomália-detektáló technikát hoztak létre. A felhasználó által kiválasztott hiperparaméterek jelentősen befolyásolják ezen algoritmusok teljesítményét. A kutatóknak többféle megközelítést kell kipróbálniuk változatos hiperparaméter-beállításokkal, mivel nincs ideális megoldás egy tartományra vagy adatkészletre. Ezen algoritmusok teljesítményét a kutatók több szempontból kell értékeljék. A legtöbb munka bizonyos kérdéseken vagy területeken történik, és a módszereket egy vagy néhány nézőpontból értékelik "alapértelmezett" paraméterek használatával, hiperparaméter-optimalizálás nélkül. Ez a cikk egy átfogó anomália-észlelési benchmarking eredményeit tárgyalja, amely 34 algoritmust értékelt számos releváns adatkészleten, számos területről. Az eredményeket különböző szögekből mutatjuk be. A hiperparaméter-optimalizálást az egyes módszerek és adatkészletek optimális paramétereinek meghatározására használtuk minden tesztforgatókönyvben. Ez a cikk számos módszer előnyeit és hátrányait tárgyalja, és bemutatja az anomália-észlelő algoritmusok taxonómiáját a probléma és a bemeneti adatok függvényében.

Kulcsszavak: mesterséges intelligencia, anomália-észlelés, benchmarking, anomália-észlelő algoritmusok taxonómiája

1. Bevezetés

Az anomália-észlelés/anomália detektálás (AD), más néven outlier-azonosítás, kulcsfontosságú mesterséges intelligencia (AI) és gépi tanulási (ML) feladat több alkalmazással, mint például a pénzmosás elleni küzdelem, a behatolás-észlelés a kiberbiztonságban, a csalások felderítése a pénzügyekben, az egészségügyben, többek között az ipari hiba, az orvosi diagnózis és a betegség kitörésének felderítése.

Annak ellenére, hogy már számos benchmark és kiértékelő munka létezik az anomáliák felderítésére, ezeknek a munkáknak gyakran több hiányossága is van. A legfontosabb hiba az, hogy a cikkek túlnyomó többsége az algoritmusok teljesítményét az "alapértelmezett" beállítások használatával értékeli (lásd: [1], [2], [4], [5], [7], [8]). Az anomália-észlelő algoritmusokat és azok teljesítményét nagymértékben befolyásolja,

hogy a felhasználó hogyan állítja be az egyes algoritmusok hiperparamétereit, ezért azt állíthatjuk, hogy hiperparaméter-optimalizálás nélkül nem lehet értékelni, hogy az algoritmus milyen hatékonyan teljesít a használatbavételkor.

A meglévő kiértékelő munkák további hiányosságai, hogy ezeknek a cikkeknek a többsége csak a nem felügyelt anomáliák észlelésének problémáját elemzi, figyelmen kívül hagyva a félig felügyelt vagy felügyelt algoritmusokat (például [1], [3], [4], [7] esetén), nem veszik figyelembe a zaj hatásait, az adatok előfeldolgozásának vagy az adatok dimenziójának hatásait, nem veszik figyelembe a különböző típusú anomáliákat, és az algoritmusokat egy vagy korlátozott számú kiértékelési metrika segítségével értékelik (lásd [4], [5]).

E hiányosságok kiküszöbölésére tudásunk szerint az egyik legkiterjedtebb anomália-detektálási referenciaértéket/benchmarkot fejlesztettük ki. 19 felügyelt, 4 félig felügyelt és 11 felügyelt anomália-detektáló technikát hasonlítottunk össze és vizsgáltunk meg több szempontból is. Ezeket az algoritmusokat különféle körülmények között teszteltük, például zajos vagy sérült adatokkal az adatkészletben, hogy értékeljük robusztusságukat. Ezenkívül különböző típusú anomáliákat használtunk, hogy meghatározzuk a különböző környezetek hatását az elemzett algoritmusokra, és megváltoztattuk az adatkészletek dimenzióját, hogy értékeljük a redundáns vagy irreleváns jellemzők hatását. Ezeket a kísérleteket több adathalmazon végeztük, amelyekről a következő részekben lesz szó.

2. Kapcsolódó munkák

Az anomália-észlelés fontos gépi tanulási feladat, melyet számos helyen alkalmaznak, ez az oka annak, hogy sok különböző megoldás és algoritmus foglalkozik ezzel a problémával. Noha számos benchmark és összehasonlító tanulmány létezik (például [1], [2], [3], [4]), mindegyiknek vannak hiányosságai, amelyeket ebben a cikkben szeretnénk orvosolni.

A [6]-ban a szerzők inkább a hiperparaméter-hangolási hatásokra összpontosítottak az anomália-detektáló technikák teljesítménye helyett. Ebben a cikkben csak 6 nem felügyelt algoritmust vontak be az elemzésbe, ami egyértelműen azt mutatja, hogy ez az elemzés tovább fejleszthető. A cikk legjelentősebb tanulsága az, hogy a legtöbb algoritmus jobban teljesít, ha a hiperparamétereiket megfelelően hangoljuk, nem pedig az alapértelmezett hiperparamétereket használjuk. Ezen megállapítások miatt úgy döntöttünk, hogy benchmarkjainkat kisebb adatkészleteken futtatjuk, de hiperparaméter-optimalizálással hangoljuk őket.

Az AD-algoritmus-benchmarkingról szóló egyik legátfogóbb cikk a [7]. Ebben a cikkben a szerzők 30 anomália-észlelő algoritmust elemeztek 57 adathalmazon, több oldalról, például felügyelet, zajhatások vagy anomáliatípusok figyelembe vételével. Az elemzés egyetlen hátránya a hiperparaméter-hangolás hiánya, ezeket az algoritmusokat az "alapértelmezett" paraméterekkel futtatták, ami azt jelenti, hogy az eredmények szuboptimálisak voltak, és nagy az esély arra, hogy hiperparaméter-optimalizálással az algoritmusok rangsorolása módosulhat.

Ebben a munkában megoldottuk mindazokat a hátrányokat, amelyeket a korábban ismertetett benchmarkok tartalmaznak. Kutatásainkat több mint 30 algoritmuson végeztük hangolt hiperparaméterek felhasználásával.

3. Anomália-észlelési benchmarkok konfigurációja

3.1. Használt hiperparaméter-optimalizálási technika

A kiértékelt AD algoritmusok optimális hiperparaméter hangolása és megtalálása érdekében a PSO-SA (Particle Swarm Optimization - Simulated Annealing) [8] nevű algoritmusunkat használtuk. Ennek az algoritmusnak az egyik legjelentősebb előnye, hogy könnyen párhuzamosan futtatható, mert minden részecske más-más elosztott csomóponton, más-más gépen futtatható, és csak üzenettovábbítás szükséges közöttük.

3.2. Értékelési mutatók

Az AUCROC [9] érték a bináris osztályozó képességének mértéke az osztályok között. Az AUC az elválaszthatóság mértéke, míg a ROC egy valószínűségi görbe. Azt jelzi, hogy a modell képes-e megkülönböztetni az osztályokat. Minél magasabb az AUC, annál pontosabban jósolja a modell a nulla osztályokat nullának, az egy osztályt pedig egynek. Például minél nagyobb az AUC, annál nagyobb a modell azon képessége, hogy különbséget tudjon tenni a betegségben szenvedő és nem beteg betegek között.

Az AUCPR [10] modell metrikáját annak felmérésére használják, hogy egy bináris osztályozási modell mennyire képes megkülönböztetni a precíziós-visszahívás párokat vagy pontokat. Ezeket az értékeket

különböző küszöbértékek alkalmazásával kapjuk valószínűségi vagy más folyamatos kimenetű osztályozón. Az AUCPR a precízió-visszahívás átlaga, súlyozva egy adott küszöbérték valószínűségével.

Ez a két metrika akkor előnyös, ha a kiválasztott osztályozási küszöb (osztályozás-küszöb-invariáns) figyelembevétele nélkül akarjuk mérni a modell előrejelzéseinek minőségét, ami különösen a nem felügyelt vagy félig felügyelt anomália-detektáló algoritmusok esetében hasznos, ahol különböző besorolási küszöbök eltérő eredményeket hozhatnak.

3.3. Tesztelt algoritmusok

Három típusú anomália-észlelési beállítást különböztethetünk meg az adatkészletben található címkék alapján:

1. *Felügyelt anomália-észlelés* – az adatok teljes címkékkel ellátott betanítási és tesztadatkészletekből állnak. A következő felügyelt AD-algoritmusokat építettük be benchmarkainkba: KNearestNeighbors (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Extra Tree (ET), Extra Trees (ETs), Ridge Classifier (RC), Passive-Aggressive Classifier (PAC), Gradient Boosting Classifier (GB), SGD Classifier (SGD), XGB Classifier (XGB)
2. *Félig felügyelt anomália-észlelés* – a betanítási adatok teljes egészében normál, anomáliamentes adatokat tartalmaznak. Az alapelv az, hogy a normál osztály modelljét megtanuljuk, és ettől a modelltől eltérve anomáliákat találhatunk. Ezt a koncepciót gyakran "egyosztályos osztályozásnak" nevezik. A következő félig felügyelt AD-algoritmusokat építettük be benchmarkainkba: GANomaly, Deep Semi-Supervised Anomaly Detection (DeepSAD), Pairwise Relation prediction-based ordinal regression Network (PReNet), Feature Encoding with AutoEncoders for Weakly-supervised Anomaly Detection (FEAWD)
3. *Felügyelet nélküli anomália-észlelés* – a leginkább adaptálható konfiguráció, amely nem igényel címkézést. Egy nem felügyelt anomália-észlelő rendszer teljes mértékben az adatkészlet belső tulajdonságai alapján pontszámot rendel az adatokhoz. Általában a távolságokat vagy sűrűségeket használják annak becslésére, hogy mi jellemző és mi abnormális. A következő, nem felügyelt AD-algoritmusokat építettük be benchmarkainkba: kNN for Outlier Detection (KNNOD), Local Outlier Factor (LOF), Cluster-Based Local Outlier Factor (CBLOF), Histogram-based Outlier Score (HBOS), One-Class Support Vector Machine, Robust Principal Component Analysis (PCA), Angle-base Outlier Detection (ABOD) Subspace outlier detection (SOD), Feature Bagging Outlier Detector (FBOD), Isolation-based Anomaly Detection (IBAD), Isolation Forest (IF), Kernel Density Estimation for Anomaly Detection (KDE), Linear Method for Deviation-based Outlier Detection (LMDD), Lightweight Online Detector (LOD), Locally Selective Combination (LSC), LSTM for Anomaly Detection (LSTMAD), Minimum Covariance Determinant Anomaly Detection (MCDAD), Single Objective Generative Adversarial Active Learning (SOGAAL), Variational Auto Encoder for Anomaly Detection (VAE)

3.4. Tanulási és tesztelési adatkészletek

Kísérleteink futtatása előtt kulcsfontosságú, hogy tisztázzuk a bemeneti adatok természetét, és meghatározzuk az anomália definícióját az alábbi adatkészletek alapján:

1. *annthyroid* - a cél a pajzsmirigy alulműködés azonosítása, így a pajzsmirigy működési zavara jelenti az anomáliát
2. *breastw* - a probléma az emlőrák diagnózisa és prognózisa
3. *cardio* - a probléma a szív- és érrendszeri betegség jelenlétének vagy hiányának meghatározása, ez a betegség az anomália osztályt képviseli
4. *yeast* - ez az adatkészlet egy fehérje-fehérje interakciós hálózatból áll, ahol a hibás fehérjéket (abnormális fehérjéket) anomáliának tekintik
5. *glass* - ez az adatkészlet a bűnkutatásra vonatkozik, a bűncselekmény helyszínén hagyott üveg bizonyítékként használható fel, ha megtalálható és felismerhető. Ez az adatkészlet több különböző típusú (több osztályú) üvegről tartalmaz információkat. A 6. osztály a kisebbségi osztály, és ennek pontjai anomáliaként, míg az összes többi pont normálként vannak jelölve.
6. *ionosphere* - radaradatok, amelyek 16 nagyfrekvenciás vevőegységből származnak. Ezek körülbelül 6,4 kilowatt teljesítményt adnak ki, melyeket a légkör elektronjaira lőttek. A "jó" radar eredmények azt mutatják, hogy az ionoszférának van valamiféle szerkezete. A "rossz" hozam azok, amelyek nem, a jeleik áthaladnak az ionoszférán.

7. *lymphography* - a vizsgálatok eredményei alapján a betegek ebben az adathalmazban négy csoportra vannak osztva. Mind az 1., mind a 4. osztályból csak 6 példa van. Ezeket a csoportokat abnormálisnak soroljuk be.
8. *pageblocks* - az adatkészlet információkat tartalmaz a dokumentumoldalakon lévő különböző típusú blokkokról. Egy dokumentum elemzéséhez fontos, hogy különbséget tudjunk tenni a szöveg, a képek és a grafika között. Ha a blokk tartalma szöveg, azt "inlier"-nek nevezik. Ha nem szöveges, akkor "outlier"-nek hívják.
9. *wine* - ennek az adathalmaznak a fő célja a rossz minőségű borok felismerése és a bor minőségének általános javítása.
10. *stamps* – ez az adatkészlet a csaló (fénymásolt vagy szkennelt és nyomtatott) és hiteles (tinta) bélyegzőket írja le. A jellemzőket a bélyegek színe és nyomtatási minősége határozza meg. A hamisított bélyegzők anomáliának minősülnek.

4. Kísérletek

Összesen 34 x 3 x 10 x 10 x 10 tesztesetet futtattunk le, ami 102 000 tesztet jelent. Minden teszt során hiperparaméter-optimalizálást is végeztünk, hogy minden egyes algoritmushoz megtaláljuk a legjobb hiperparaméter-beállításokat.

Ezekkel a tesztekkel több kérdésre próbáltunk választ adni, például, hogy milyen típusú algoritmusok jobbak különböző szintű felügyelet esetén, mely algoritmusokra vannak leginkább hatással a különböző zajtípusok, hogyan befolyásolja az anomália típusa a különböző algoritmusok teljesítményét, és mi az irreleváns jellemzők hatása ezen algoritmusok teljesítményére.

4.1. A különböző anomáliatípusok hatása

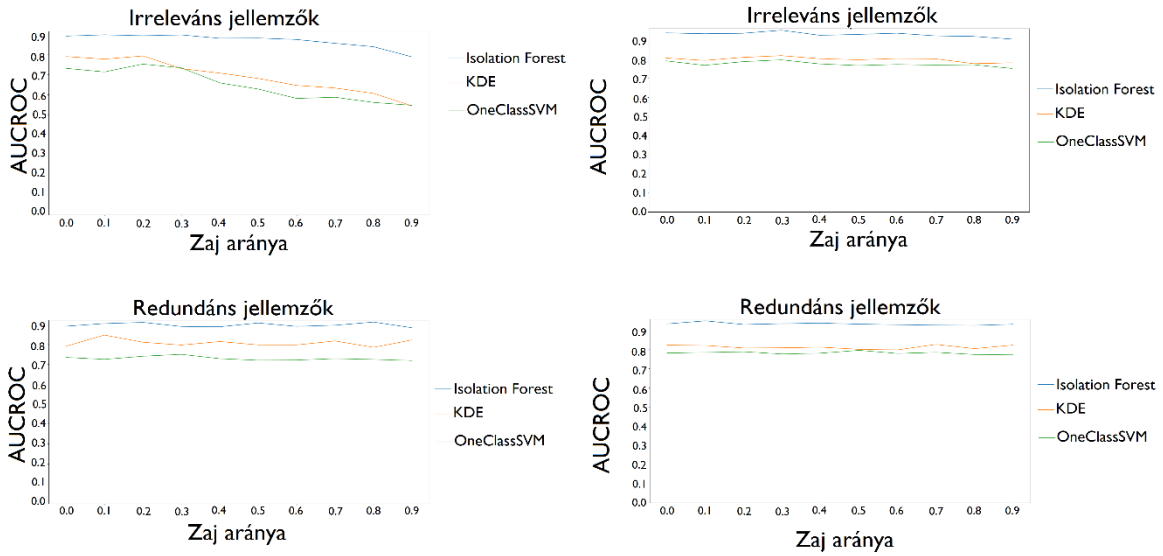
A különböző anomáliatípusok hatásának megértéséhez minden adathalmazon lefuttattuk a teszteket, zaj nélkül, így az eredményeket nem befolyásolják a sérült adatok. Meglepő módon azt az eredményt kaptuk, hogy a felügyelt algoritmusok többsége (a Random Forest kivételével) rosszabbul teljesít, mint a legjobb nem felügyelt módszerek a helyi és globális anomáliák esetében. A klaszteres anomáliák esetében a felügyelt algoritmusoknak sikerült legyőzniük a nem felügyelteket. A félig felügyelt algoritmusok kiváló választások globális vagy akár klaszter anomáliák esetén, de a lokális anomáliák esetén jóval gyengébb a teljesítményük.

Ez az anomáliatípusokon alapuló különbség rávilágít az adatkészleteinkben előforduló anomáliatípusok előzetes ismerete fontosságára.

4.2. A különböző zajtípusok hatása

Ahogy az 1-es ábrán látható, a felügyelt algoritmusok nagyon robusztusak ismétlődések vagy irreleváns jellemzők esetén, teljesítményük nem változott. Ezzel szemben a nem felügyelt algoritmusokra hatással van az irreleváns jellemzők száma, amint azt az 1-es ábra felső diagramján láthatjuk.

Az 1-es ábrán látható, hogy a tesztelt félig felügyelt algoritmusok nagyon robusztusak duplikált adatok esetén, de teljesítményüket nagymértékben befolyásolják az irreleváns jellemzők, ami nagymértékben csökkenti általános teljesítményüket.

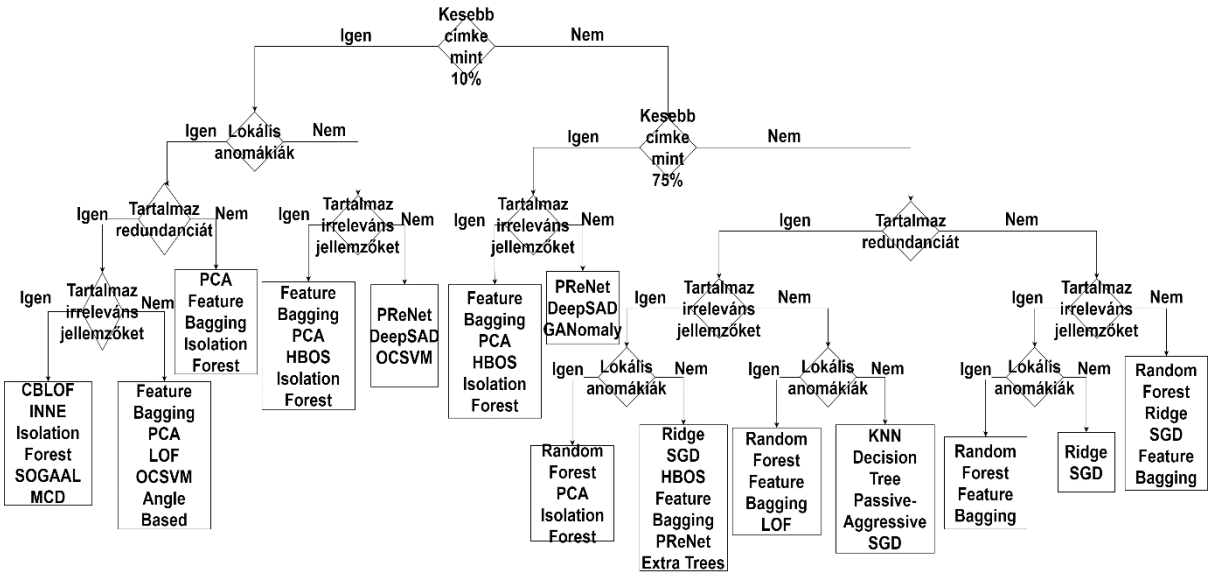


1. ábra. *Példák a zaj hatására. Balra fent – az irreleváns jellemzők hatása a nem felügyelt algoritmusokra; Jobbra fent – a irreleváns jellemzők hatása a felügyelt algoritmusokra; Balra lent – a redundáns jellemzők hatása a felügyelt algoritmusokra; Alul-jobb – a redundáns jellemzők hatása a felügyelt algoritmusokra*

4.3. Javaslatok az algoritmus kiválasztásához

Az előző megfigyelések és kísérleteink eredményei alapján meghatározhatunk néhány szabályt a keresési tér leszűkítésére, amikor egy adott kontextushoz anomália-észlelő algoritmust akarunk választani, ahogy az a 2-es ábrán látható. Ha például több mint 10%, de kevesebb, mint 75% címkénk van, és nem tudjuk, melyek a releváns jellemzők (tehát a bemeneti adatok tartalmazhatnak irreleváns jellemzőket), akkor a 2-es ábra alapján az anomáliák kimutatására FB, PCA, HBOS vagy IF algoritmus használható. Abban az esetben, ha biztosak vagyunk abban, hogy mely jellemzőket kell használni az anomália észleléséhez, választhatunk a PReNet, DeepSAD vagy OCSVM anomália-detektáló algoritmusok közül.

A 2-es ábrán bemutatott döntési fa segítségével a bemeneti adatok alapján optimalizálhatjuk a keresési teret az anomália-észlelő algoritmusok kiválasztásához, ezzel csökkentve a kísérletekhez szükséges időt.



2. ábra. *Javaslatok az AD algoritmus kiválasztásához*

5. Következtetések

Ebben a cikkben 34 anomália-észlelő algoritmust elemeztünk 10 benchmark adatkészlet felhasználásával. Az algoritmusok pontos összehasonlítása érdekében hiperparaméter-optimalizálást alkalmaztunk, így biztosítva, hogy optimális körülmények között korrekt összehasonlítást végezzünk.

A benchmarkok futtatása során kiemeltük a különböző zajtípusok hatását a különböző algoritmusok teljesítményére valamint elemeztük a különböző típusú anomáliák hatását. Tesztjeink eredményei alapján gyakorlati javaslatokat készítettünk a probléma kontextusa alapján legmegfelelőbb anomália-detektáló algoritmus kiválasztásához. Ezek a javaslatok segíthetnek csökkenteni az algoritmustípusok keresési terét, ezáltal csökkentve az összes lehetséges anomália-észlelő algoritmus teszteléséhez szükséges időt egy adott probléma esetén.

Irodalmi hivatkozások

[1] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data mining and knowledge discovery, 30(4):891–927, 2016.

[2] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui. A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 74:406–421, 2018.

[3] A. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. A meta-analysis of the anomaly detection problem. ArXiv, 1503.01158, 2015.

[4] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one, 11(4):e0152173, 2016.

[5] Goldstein M., Uchida S., A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLoS One. 2016 Apr 19;11(4):e0152173. doi: 10.1371/journal.pone.0152173. PMID: 27093601; PMCID: PMC4836738.

[6] Soenen, J., Leuven, K., Wolputte, E.V., Perini, L., Vercruyssen, V., Meert, W., Davis, J., Blockeel, H. (2021). The Effect of Hyperparameter Tuning on the Comparative Evaluation of Unsupervised Anomaly Detection Methods.

[7] Han S., Hu X., Huang H., Jiang M., Zhao Y., ADBench: Anomaly Detection Benchmark (November 3, 2022). Advances in Neural Information Processing Systems (NeurIPS), 2022, Available at SSRN: https://ssrn.com/abstract=4266498 or http://dx.doi.org/10.2139/ssrn.4266498

[8] Czako, Z., Sebestyen, G., Hangan, A. (2021). AutomaticAI - A hybrid approach for automatic artificial intelligence algorithm selection and hyperparameter tuning. Expert Syst. Appl., 182, 115225.

[9] Developers Google, 2022, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[10] Boyd K., Eng K.H., Page C.D., Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Machine Learning and Knowledge Discovery in Databases, vol 8190. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-40994-3_29