

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 2981

Usporedba algoritama grupiranja u postupcima otkrivanja anomalija

Jelena Nemčić

Zagreb, svibanj 2022.

Zagreb, 11. ožujka 2022.

DIPLOMSKI ZADATAK br. 2981

Pristupnica: **Jelena Nemčić (0036497921)**

Studij: Računarstvo

Profil: Računarska znanost

Mentor: izv. prof. dr. sc. Goran Delač

Zadatak: **Usporedba algoritama grupiranja u postupcima otkrivanja anomalija**

Opis zadatka:

Odabrati, proučiti i opisati algoritme za grupiranje primjerene obradi velikih skupova podataka. Opisati obilježja algoritama i objasniti njihov princip rada nad pokaznim primjerima. Proučiti postojeće metrike za vrednovanje uspješnosti algoritama grupiranja. Odabrati primjeren skup podataka za postupak otkrivanja anomalija. Programski ostvariti i provesti vrednovanje odabranog podskupa algoritama nad odabranim skupom podataka. Opisati programsko ostvarenje sustava, rezultate vrednovanja algoritama te navesti korištenu literaturu i primljenu pomoć.

Rok za predaju rada: 27. lipnja 2022.

SADRŽAJ

1. Uvod	1
2. Anomalije	2
2.1. Pojava anomalija i njeni uzroci	2
2.2. Klasifikacija anomalija	3
2.3. Problem otkrivanja anomalija	4
2.4. Metode otkrivanja anomalija	6
3. Algoritmi grupiranja za otkrivanje anomalija	7
3.1. O algoritmima grupiranja	7
3.2. K-Means	8
3.3. DBSCAN	8
3.4. Gaussova mješavina	8
4. Korišteni skupovi podataka	9
4.1. prvi	9
4.2. drugi	9
4.3. treći	9
5. Implementacija	10
6. Rezultati	11
7. Zaključak	12
Literatura	13

1. Uvod

Svaki dan generira se velika količina podataka koja se zatim obrađuje kako bi se iz nje saznale nove informacije. Jedan od načina korištenja podataka jest otkrivanje neobičnog ponašanja i pronalaženje anomalija.

Anomalijom se smatra svaki događaj ili opažanje koje značajno odstupa od većine podataka i ne ponaša se na očekivan način. Takvi primjeri mogu izazvati sumnju da ih generira drugačiji mehanizam ili se činiti nedosljednima s ostatkom tog skupa podataka.

Otkrivanje anomalija pronalazi primjenu u mnogim domenama uključujući kibernetičku sigurnost, medicinu, računalni vid, statistiku, neuroznanost i oružane snage. Koristi se također i za otkrivanje financijskih prijevара, industrijskih oštećenja i poremećaja u ekosustavu. Anomalije mogu predstavljati problem te su tada tražene radi namjernog izostavljanja iz skupa podataka kako bi se dobila točnija statistička analiza ili bolje predviđanje nekog modela strojnog učenja. Međutim, u mnogim su primjenama anomalije najzanimljiviji dio skupa podataka i predstavljaju novu pojavu koju je potrebno identificirati i dalje istražiti.

Jedna od tehnika otkrivanja anomalija jest korištenje algoritama grupiranja s ciljem pronalaženja elemenata koji ne pripadaju niti jednoj grupi. U ovom radu dano je objašnjenje problema pronalaska anomalija, opis različitih algoritama grupiranja i korištenih skupova podataka te usporedba izvedbe tih algoritama u postupcima otkrivanja anomalija. Algoritmi odabrani za usporedbu su K-Means, DBSCAN i Gaussova mješavina, a testirani su na problemima otkrivanja ...

2. Anomalije

2.1. Pojava anomalija i njeni uzroci

Postoji više pokušaja definiranja anomalija, a većina njih opisuje anomaliju kao opažanje čiji se obrazac ponašanja razlikuje od očekivanog, najčešće se pojavljuje vrlo rijetko u skupu podataka i njegova su obilježja značajno drugačije od onih većine preostalih opažanja. Također, anomalijom se može smatrati podatak koji se čini nedosljedan i relativno udaljen od drugih podataka iz skupa ili izaziva sumnju da ga generira drugačiji mehanizam.

Anomalije se mogu pojaviti u bilo kojem skupu podataka i ponekad njihovo otkrivanje može biti od izuzetne važnosti. Često se otkrivanje anomalija provodi u predobradi kako bi se mogle ukloniti iz skupa podataka. Time se dobiva točnija statistika podataka, bolje predviđanje modela strojnog učenja i bolja vizualizacija podataka. S druge strane, anomalije mogu biti najvažnija i najzanimljivija opažanja i tada se otkrivanje anomalija provodi radi njih samih. Primjeri takve primjene su otkrivanje upada u području kibernetičke sigurnosti, otkrivanje financijskih prijevара i lažnih informacija, otkrivanje kvarova i pogrešaka, praćenje stanja sustava i vremenskih serija, detekciju događaja u senzorskim mrežama, otkrivanje poremećaja u ekosustavu, otkrivanje nedostataka na slikama pomoću računalnog vida te za postavljanje medicinske dijagnoze i provođenje zakona.

Mogući uzroci pojave anomalija su:

1. Podaci pripadaju različitim razredima.
 - Anomalije se razlikuju od ostalih podataka jer pripadaju drugom razredu, koji ima drugačija obilježja.
 - Primjer takvih anomalija su financijske prijevare, strani upad u sustav i pojava bolesti.
2. Prirodna varijacija.

- Neki skupovi podataka mogu se modelirati normalnom distribucijom, gdje su anomalije oni događaji koji imaju vrlo malu vjerojatnost pojavljivanja.

3. Pogreške u mjerenju ili prikupljanju podataka.

- Do pojave anomalija može doći ako podaci sadrže šum, ako postoji kvar u mjernim instrumentima ili zbog ljudske pogreške.
- Krajnji je cilj eliminirati ovakve anomalije jer smanjuju kvalitetu podataka.

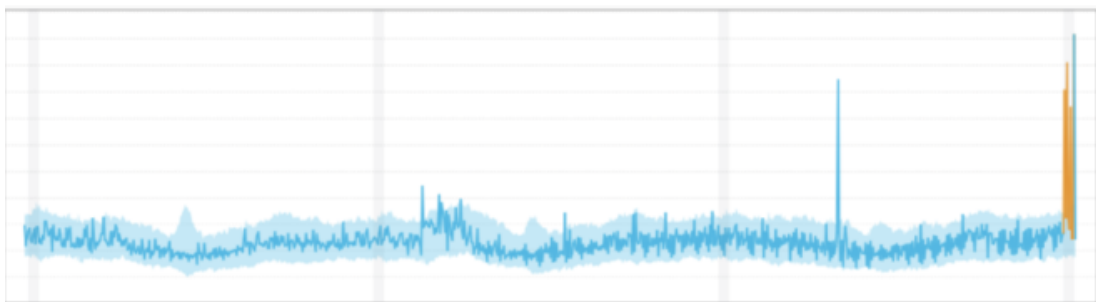
U ovom radu razmatrat će se samo anomalije koje se javljaju kao posljedica toga što podaci prirodno pripadaju različitim razredima.

2.2. Klasifikacija anomalija

Kako bi sustav za otkrivanje anomalija mogao točno identificirati potencijalna odstupanja, nužno je znati koja vrsta anomalije se očekuje. Anomalije se mogu podijeliti u tri glavne kategorije:

1. Globalne anomalije

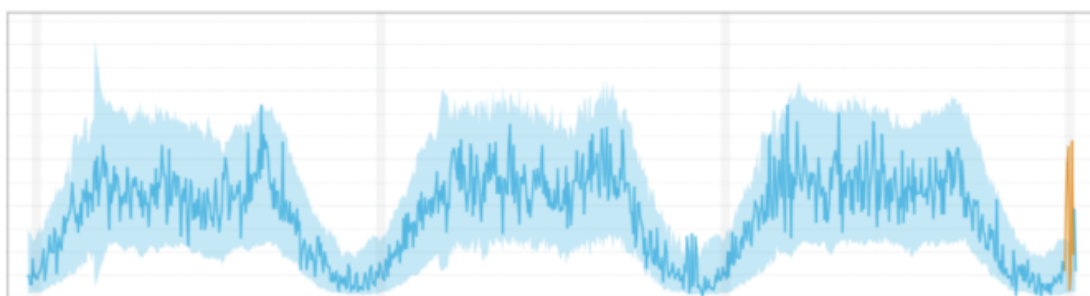
Opazanje se smatra globalnim odstupanjem ili globalnom anomalijom ako se njegova vrijednost ili vrijednost nekih njegovih obilježja značajno razlikuje od vrijednosti cjelokupnog skupa podataka. Gledano u n-dimenzionalnom prostoru, taj se podatak nalazi daleko od svih ostalih podataka iz skupa. Primjer globalne anomalije dan je na slici 2.1.



Slika 2.1: Globalna anomalija. Preuzeto s <https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6>

2. Kontekstualne anomalije

Kontekstualne ili uvjetne anomalije su opažanja čije se vrijednosti znatno razlikuju od ostalih opažanja koja postoje u istom kontekstu. Takve vrijednosti ne moraju biti izvan globalnih očekivanja, ali odudaraju od konteksta u kojem se nalaze. Također, jedan podatak koji je anomalija u kontekstu jednog skupa podataka ne mora biti anomalija u drugom. Ovakva odstupanja najčešća su u podacima vremenskih serija jer takvi skupovi podataka sadrže zapise ovisne o vremenskom razdoblju. Slika 2.2 prikazuje primjer takve anomalije.



Slika 2.2: Kontekstualna anomalija

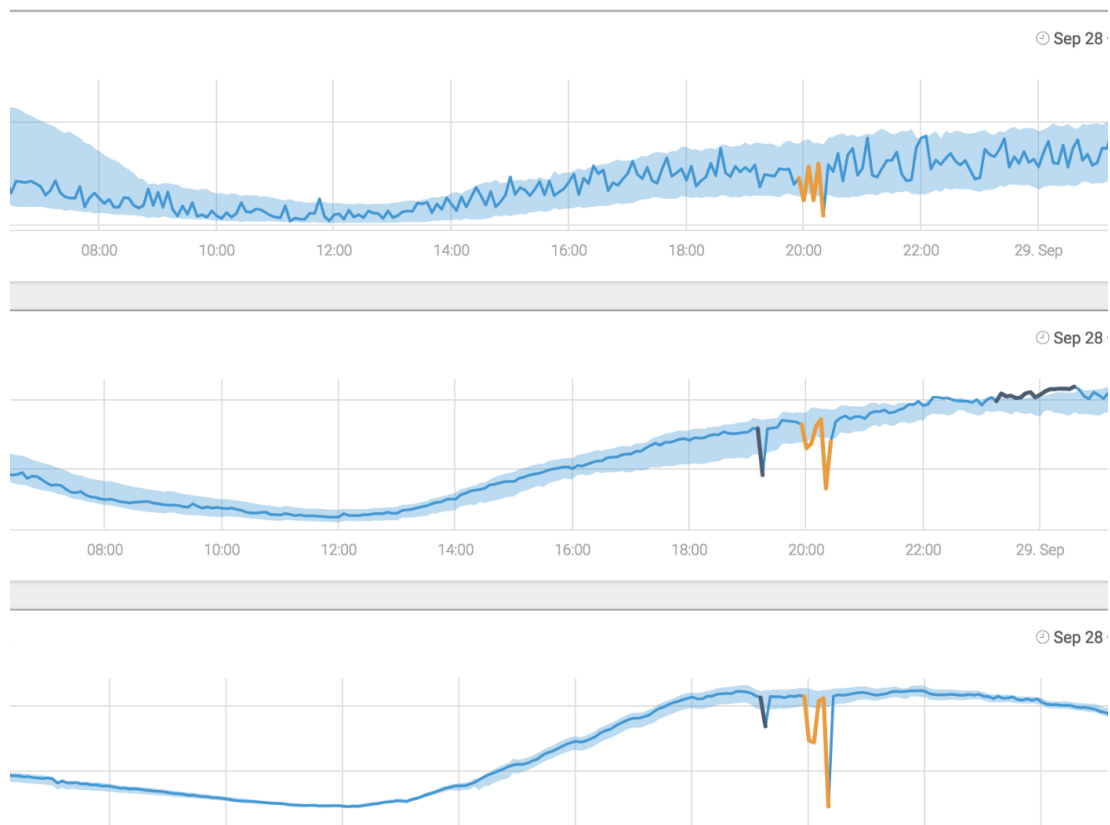
3. Kolektivne anomalije

Podskup podataka smatra se kolektivnom anomalijom ako njihove vrijednosti kao grupa značajno odstupaju od cijelog skupa podataka, ali vrijednosti pojedinačnih podataka nisu same po sebi anomalne ni u globalnom niti u kontekstualnom smislu. U podacima vremenskih serija, kolektivne anomalije mogu se manifestirati kao vrhovi i doline koje se javljaju izvan vremenskog okvira kada je takvo ponašanje normalno, kao što se vidi na slici 2.3.

Ovisno o vrsti anomalije primjenjuju se različite metode i načini detekcije. Ovaj rad fokusira se na globalne anomalije i njihovo pronalaženje.

2.3. Problem otkrivanja anomalija

Otkivanje anomalija svodi se na problem definiranja očekivanog ponašanja podataka ili granica unutar kojih se podaci smatraju normalnima te identificiranja točaka koje se ne nalaze unutar njih. Postoji nekoliko faktora koji čine ovaj problem vrlo teškim.



Slika 2.3: Kolektivna anomalija

- Učinkovito modeliranje normalnih vrijednosti i ponašanja može biti vrlo izazovan problem. Često je teško nabrojati sva moguća normalna ponašanja nekog objekta i klasificirati neki podatak kao anomaliju. Također, granica između normalnih podataka i anomalija može biti vrlo nejasna.
- Svaki problem zahtjeva specifičan način detekcije anomalija jer su odabir mjere sličnosti i modeliranje odnosa ovisni o svojstvima tog problema. Zbog toga nije moguć razvoj univerzalno primjenjive metode otkivanja anomalija.
- Prikupljeni podaci često sadrže šum koji može imati vrijednosti koje znatno odstupaju od normalnih ili čak nedostaju. Šum smanjuje kvalitetu podataka i otežava definiranje granica između normalnih podataka i anomalija te se često šum može pogrešno identificirati kao anomalija i obrnuto.
- U nekim primjenama, korisnik ne želi samo identificirati anomalije već i razumjeti zašto su ti podaci detektirani kao abnormalni. Zbog toga metoda otkrivanja anomalija mora biti razumljiva, smisljena i pružiti opravdanje detekciji.

2.4. Metode otkrivanja anomalija

Postoji puno različitih tehnika otkrivanja anomalija i one se mogu podijeliti u četiri glavne kategorije.

1. Statističke metode

Statistički pristup naziva se još i pristup temeljen na modelu jer sadrži model koji opisuje obilježja skupa podataka. Model najčešće sadrži distribuciju vjerojatnosti podataka i za svaki podatak računa se vjerojatnost njegova pojavljivanja u tom modelu. Ako je ta vjerojatnost vrlo mala, podatak se proglašava anomalijom.

2. Metode temeljene na blizini

(a) Metode temeljene na udaljenosti

Metode temeljene na udaljenosti pretpostavljaju da je podatak anomalija ako mu se najbliži susjedi nalaze daleko u prostoru značajki odnosno ako blizina objekta njegovim susjedima značajno odstupa od blizine većine drugih objekata njihovim susjedima u istom skupu podataka.

(b) Metode temeljene na gustoći

Metode temeljene na gustoći koriste broj podataka koji se nalaze unutar definiranog prostora ispitivanog podatka za definiranje lokalne gustoće. Što je lokalna gustoća objekta manja, veća je vjerojatnost da je on anomalija.

3. Metode temeljene na grupiranju

Metode koje se temelje na grupiranju pretpostavljaju da normalni podaci pripadaju velikim i gustim grupama, dok anomalije pripadaju malim i rijetkim grupama ili ne pripadaju niti jednoj. Razlika između grupiranja i metoda temeljenih na gustoći je u tome što grupiranje dijeli podatke u grupe dok metode temeljene na gustoći dijele podatkovni prostor.

U ovom radu za detekciju anomalija koristit će se algoritmi temeljeni na grupiranju. Za usporedbu su izabrani algoritmi K-Means, DBSCAN i Gaussova mješavina.

3. Algoritmi grupiranja za otkrivanje anomalija

3.1. O algoritmima grupiranja

Grupiranje je podjela skupa podataka u grupe, tako da su podaci u istoj grupi sličniji jedni drugima nego podacima iz ostalih grupa. Cilj jest pronalaženje intrinzičnih grupa u skupu podataka. Algoritmi grupiranja pripadaju u skupinu nenadziranih metoda strojnog učenja jer su ulazni podaci dani bez ciljnih vrijednosti odnosno nisu označeni.

Grupiranje se može podijeliti u dvije kategorije:

1. Tvrdo grupiranje - podatak ili pripada grupi ili ne pripada
2. Meko grupiranje - podatak pripada svakoj grupi s određenom vjerojatnošću

Osim po tipu grupiranja koje provode, algoritmi grupiranja razlikuju se i po tome kako definiraju pojam grupe i sličnost podataka. Svaki algoritam pretpostavlja specifičan model grupe, a najčešći modeli su:

1. Modeli povezanosti - na temelju udaljenosti podataka stvara se hijerarhijsko stablo grupa
2. Centroidni modeli - podaci se organiziraju u nehijerarhijske grupe ovisno o udaljenosti od centroida te grupe
3. Modeli distribucije - grupe se modeliraju pomoću vjerojatnosti da podaci pripadaju istoj statističkoj distribuciji
4. Modeli gustoće - područja iste gustoće povezuju se u grupe

Ne postoji objektivno najbolji algoritam grupiranja, već odabir algoritma ovisi o problemu koji se rješava. Algoritam se može odabrati na temelju modela grupe ili eksperimentalno. Također, algoritam dizajniran za jednu vrstu modela grupe općenito neće raditi na skupu podataka koji sadrži drugačiji tip grupa.

3.2. K-Means

3.3. DBSCAN

3.4. Gaussova mješavina

4. Korišteni skupovi podataka

4.1. prvi

4.2. drugi

4.3. treći

5. Implementacija

6. Rezultati

7. Zaključak

Zaključak.

LITERATURA

- [1] G Sandhya Madhuri i Dr. M. Usha Rani. Anomaly detection techniques causes and issues. *International Journal of Engineering & Technology*, 2018. URL <https://www.sciencepubco.com/index.php/ijet/article/view/22791>.
- [2] Animesh Patcha i Jung-Min Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *ScienceDirect*, 2007. URL <https://www.sciencedirect.com/science/article/abs/pii/S138912860700062X>.
- [3] Ramiz Aliguliyev Rasim Alguliyev i Lyudmila Sukhostat. Anomaly detection in big data based on clustering. *Statistics Optimization & Information Computing*, 2017. URL https://www.researchgate.net/publication/321448608_Anomaly_Detection_in_Big_Data_based_on_Clustering.
- [4] Ajay Sreenivasulu. Evaluation of cluster based anomaly detection, 2019. URL <https://www.diva-portal.org/smash/get/diva2:1382324/FULLTEXT01.pdf>.
- [5] Jiong Jin Srikanth Thudumu, Philip Branch i Jugdutt (Jack) Singh. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 2020. URL <https://doi.org/10.1186/s40537-020-00320-x>.

Usporedba algoritama grupiranja u postupcima otkrivanja anomalija

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.