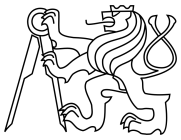
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



ZADÁNÍ DIPLOMOVÉ PRÁCE

Název:	Analýza leteckých dat a hledání anomálních pasažérů
Student:	Bc. Matúš Tóth
Vedoucí:	Ing. Pavel Kordík, Ph.D.
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra teoretické informatiky
Platnost zadání:	Do konce zimního semestru 2018/19

Pokyny pro vypracování

Prozkoumejte metody detekce anomalit z grafových dat a dat o leteckém provozu. Zpracujte data poskytnutá policií ČR do formy použitelné pro modelování a detekci anomalit. Použijte základní techniky předzpracování dat k přípravě kvalitních atributů. Ve spolupráci s policií formulujte analytické otázky, na které poté odpovíte výsledkem analýz. Analýzu dat proveďte v některém DM nástroji (např. Rapid Miner, nebo h2o.ai). Soustřeďte se zejména na detekci pasažérů podezřelých z pašování lidí, zbraní a chráněných zvířat. Výsledkem budou odpovědi na analytické otázky podpořené datovými reporty.

Seznam odborné literatury

Dodá vedoucí práce.

doc. Ing. Jan Janoušek, Ph.D.
vedoucí katedry

prof. Ing. Pavel Tvrdík, CSc.
děkan

V Praze dne 4. dubna 2017

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Diplomová práca

Analýza leteckých dát a hľadanie anomálnych pasažierov

Bc. Matúš Tóth

Vedúci práce: Ing. Pavel Kordík, Ph.D.

8. mája 2017

Poďakovanie

V prvom rade by som chcel poďakovať Ing. Pavlovi Kordíkovi, Ph.D. za cenné rady, pomoc a odborné vedenie tejto práce. Ďalej by som chcel poďakovať svojim najbližším a rodine za dôveru a neustálu podporu pri vypracovávaní tejto diplomovej práce.

Prehlásenie

Prehlasujem, že som predloženú prácu vypracoval(a) samostatne a že som uviedol(uviedla) všetky informačné zdroje v súlade s Metodickým pokynom o etickej príprave vysokoškolských záverečných prác.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona, v znení neskorších predpisov, a skutočnosť, že České vysoké učení technické v Praze má právo na uzavrenie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe 8. mája 2017 .....

Odkaz na túto prácu

Tóth, Matúš. *Analýza leteckých dát a hľadanie anomálnych pasažierov.* Diplomová práca. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2017.

Abstrakt

V tejto diplomovej práci sa venujem preskúmaniu možností detekcie anomálií v rôznych typoch dát, spracovaniu leteckých dát poskytnutých Políciou ČR do formy vhodnej pre modelovanie a detekciu spomínaných anomálií. Pre letecké dáta tiež v spolupráci s políciou definujeme analytické otázky, na ktoré následne odpoviem na základe vykonanej analýzy. Analýza bude vykonávaná pomocou pythonovských skriptov a dataminingového nástroja RapidMiner.

Klíčová slova Detekcia anomálií, Analýza, Anonymizácia dát, Predspracovanie dát, Letecké dáta, Strojové učenie

Abstract

In this diploma thesis I examine the possibilities of detecting anomalies in different types of data, pre-processing of flight data provided by the Czech Police to a form suitable for modeling and detection of the mentioned anomalies. For flight data, we also define analytical questions in co-operation with the police, which I will then answer according to the results of analysis. The analysis will be performed using Python scripts and the RapidMiner datamining tool.

Keywords Anomaly detection, Analysis, Data anonymization, Data pre-processing, Flight data, Machine learning

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Obranyschopnosť a bezpečnosť štátu a jeho obyvateľov patrí k základným funkciám štátu. Zaisťuje sa tým trvanie a suverenita štátu a je nevyhnutným predpokladom na to, aby občania mohli na území štátu užívať svoje práva a slobody. Je to komplexný pojem zahŕňajúci veľkú množinu rôznorodých činností. Dá sa však rozdeliť na dve hlavné odvetvia a to:

1. Medzinárodná bezpečnosť
2. Vnútroštátna bezpečnosť.

Existujú však oblasti, v ktorých je zaistenie bezpečnosti nad sily jednotlivca. Nie je možné individuálne sa brániť proti ozbrojenej agresii, zabrániť šíreniu zbraní hromadného ničenia alebo ovplyvniť ekonomické a environmentálne problémy sveta. Jednotlivec preto dobrovoľne ochranu svojich záujmov posúva na vyššiu organizačnú štruktúru – štát. Ten disponuje vnútornou a vonkajšou suverenitou a prostredníctvom svojich bezpečnostných zložiek (polícia, ozbrojené sily) zaisťuje bezpečnosť svojich občanov. Vnútorná suverenita mu umožňuje vykonávať jurisdikciu v rámci štátneho celku, zatiaľ čo vonkajšia suverenita mu garantuje rovnocenné postavenie v systéme medzinárodných vzťahov a to, že žiadny iný štát nemá právo zasahovať do jeho interných záležitostí.

Medzi oblasti, ktoré musí zastrešovať štát patrí aj oblasť ochrany hraníc. Už samotný pojem ochrany hraníc je nadmieru komplexný, keďže je potrebné identifikovať mnohé druhy hrozieb, od pašovania ľudí, zvierat alebo drog až po nelegálnu imigráciu a terorizmus. Pri ochrane hraníc pojednávame rôzne možnosti dopravy. Pri automobilovej a vlakovej je potrebná fyzická kontrola hraníc a kontrola jednotlivých ľudí. Keďže však letecká doprava poskytuje bohaté informácie o pasažieroch dostupné ešte pred príletom, je možné skúsiť automatizovať identifikáciu podozrivých pasažierov na základe týchto údajov.

Kapitola 1

Teória

V tejto kapitole sa budem venovať teoretickému základu pre identifikáciu spomínaných podozrivých pasažierov.

1.1 Monitorovanie leteckej dopravy

Letecký dopravce je povinen za účelem zdokonalení hraničních kontrol a boje proti nedovolenému přistěhovalectví předávat útvaru Policie České republiky údaje o cestujících, kteří překročí vnější hranici (pouze přílet) na vybraných pravidelných linkách, u charterových letů na vyžádání.

- směrnice č. 2004/82/ES
- zákon č. 49/1997 Sb., o civilním letectví
- Interní akty řízení PP a ŘSCP

Ako vidíme, letecký dopravca je zo zákona povinný poskytovať dáta o cestujúcich. Tieto poskytované údaje sú (§ 69 z.č. 4/1997):

1. číslo a typ použitého cestovného dokladu
2. štátna príslušnosť
3. meno (mená) a priezvisko
4. dátum narodenia
5. hraničný prechod vstupu na územie členských štátov
6. kódové číslo letu
7. čas odletu a príletu
8. celkový počet osôb prepravovaných uvedeným letom
9. počiatočné miesto nástupu na palubu

1.1.1 Informačný systém OBZOR

Tieto dáta tiež musia byť istým spôsobom organizované. Preto Polícia Českej republiky prišla s informačným systémom OBZOR, ktorý plne prepojuje leteckých dopravcov s Políciou pomocou siete leteckej dopravy SITA. Bol uvedený do prevádzky 1.7.2012 a posiela doň svoje dáta 30 leteckých spoločností na 73 leteckých spojoch.

Medzi jeho funkcie patria:

- vyhodnotenie formátu API správy
- vyhodnotenie správnosti jej obsahu
- základné analýzy rizík
- vykonanie previerky osôb
- prehľadové zobrazenie výsledkov
- štatistické a analytické funkcie (vytváranie profilov cestujúcich)

Ako vidíme, tento informačný systém ponúka rôzne možnosti prehľadu. Ďalšou možnosťou je zobraziť profil cestujúceho. V tomto profile sú zahrnuté všetky lety tohoto pasažiera a aj prípadné problémy.

Keďže isté profily sú rizikové už na základe národnosti cestujúceho, dátumu narodenia, miestom odletu alebo nejakou kombináciou týchto vlastností. Preto OBZOR tiež ponúka vytvorenie istých profilov a cestujúci, ktorí vyhovujú tomuto profilu sú označení na ďalšie preskúmanie.

Jednou z najdôležitejších funkcií pre nás je však export dát, pre automatické spracovanie pomocou externých nástrojov.

1.2 Gnumeric

Gnumeric je tabuľkový procesor[2], počítačový program vytvorený GNOME projectom, ktorý slúži na manipuláciu a analýzu číselných dát. Gnumeric pomáha sledovať informácie v zoznamoch, organizovať číselné hodnoty do stĺpcov a riadkov, vykonávať a aktualizovať zložité výpočty tým, že definujeme jednotlivé kroky výpočtu a následne ich modifikujeme. Umožňuje tiež vytvárať a zobraziť alebo vytlačiť rôzne typy grafov a vykonávať zložité optimalizačné modelovanie alebo vykonávať mnoho ďalších úloh, zahŕňajúcich čísla, dátumy, časy, mená alebo iné dáta.

1.2.1 ssconvert

Ssconvert je nástroj príkazového riadka pre konverziu tabuľkových súborov na rôzne formáty. Jeho syntax:

ssconvert [*OPTIONS*] *infile outfile*

1.3 Data mining

Data mining je proces objavovania vzorov vo veľkých dátových súboroch zahrňujúcich metódy založené na umelej inteligencie, strojovom učení, štatistike v spolupráci s rôznymi databázovými systémami [3][4]. Celkovým cieľom data miningového procesu je získanie informácií z dátového súboru a premeniť ho na zrozumiteľné štruktúry pre ďalšie použitie. Okrem analýzy zahŕňa aj správu dát, ich predspracovanie, modelovanie a následne odvodzovanie záverov, postprocessing získaných štruktúr, vizualizácie a iné.

1.3.1 Data miningové nástroje

Keďže by bolo nesmierne obtiažne vykonávať tieto operácie manuálne, existujú rôzne data miningové sofvéry a frameworky pre uľahčenie práce a názornejšie zobrazovanie výsledkov.

1.3.1.1 RapidMiner

RapidMiner je data miningový nástroj napísaný v programovacom jazyku Java [5]. Ponúka rôzne možnosti analýzy dát a vďaka užívateľskej prívetivosti je široko využívaný.

Okrem získavania dát, RapidMiner tiež poskytuje funkcie na predspracovanie a vizualizáciu dát, prediktívne analýzy a štatistické modelovanie, vyhodnotenie a nasadenie.

RapidMiner je šírený pod AGPL open source licenciou a možno ho stiahnuť zo SourceForge, kde je hodnotený ako najlepší analytický softvér.

1.4 Nahrádzanie chýbajúceho atribútu

V bežných dátach sa často môže stať, že niektoré záznamy neobsahujú všetky z atribútov. Toto môže nastať z rôznych príčin (porucha jedného zo senzorov na sonde, chybujúci ľudský faktor, atp.). Tieto defekty však musia byť odhalené a vrámci predspracovania dát by mala byť zvolená jedna z možností ako sa s nekonzistenciami vysporiadať. Tieto techniky zohľadňujú dôležitosť informácie, že atribút chýba.

1.4.1 Nespraviť nič

Prvou možnosťou je ponechať atribút chýbajúci. Zachováme tak informáciu, že niečo pri tomto zázname nebolo v poriadku. Nevýhodou tohoto prístupu je, že mnohé techniky učenia sa nevedia vysporiadať s chýbajúcim atribútom.

1.4.2 Vynechať záznam

Druhou možnosťou je celý záznam zmazať. Takto prídeme nielen o informáciu, že atribút chýbal, ale aj o ostatné (nechýbajúce atribúty). Tento prístup je vhodný, ak máme veľké množstvo záznamov a len nebatateľné percento z nich má chýbajúci nejaký z atribútov. Nevýhodou je, že môžeme prichádzať o cenné informácie.

1.4.3 Nahradenie priemerom

Ďalšou možnosťou je chýbajúci atribút nahradiť priemerom hodnôt (ak daný atribút poskytuje možnosť priemerovania - numerické atribúty), alebo hodnotou, ktorú atribút najčastejšie nadobúda pri záznamoch, kde nechýba. Takto sa síce zbavíme nekonzistencie, ale zase prídeme o informáciu, že atribút chýbal a navyše zo záznamov, ktoré boli do veľkej miery odlišné od ostatných sa môžu stať záznamy, ktoré nie sú odlišné batateľným spôsobom. Táto technika je vhodná ak môžeme o dátach predpokladať, že sa vyskytujú v zhlukoch a takéto vyhladenie nespôsobí žiadny problém.

1.4.4 Nahradenie význačnou hodnotou

Táto metóda spočíva v nahradení atribútu istou hodnotou, ktorú tento atribút nenadobúda v žiadnom inom prípade (napríklad pre počty je vhodné zvoliť -1, keďže počet nadobúda hodnoty prirodzených čísel). Takto nestratíme ani záznam, ani informáciu o tom, že atribút chýbal a ani nemôže nastať vyhladenie v dátach. Potrebujeme však isté znalosti o dátach, ktoré majú aby sme zvolili význačnú hodnotu správne.

1.5 Normalizácia dát

Normalizácia dát je proces predspracovania dát. Pomocou tejto normalizácie upravujeme (štandardizujeme) rozsah premenných alebo vlastností dát.

Keďže rozsah hodnôt nespracovaných údajov sa môže značne líšiť, v niektorých algoritmoch strojového učenia funkcie nemusia fungovať správne bez normalizácie. Napríklad väčšina klasifikátorov vypočíta vzdialenosť medzi dvoma bodmi podľa istej miery vzdialenosti (mnohokrát euklidovská). Ak niektorý z atribútov má veľký rozptyl hodnôt, vzdialenosť bude značne ovplyvnená práve týmto atribútom. Rozptyl všetkých atribútov by sa mal normalizovať tak, aby každý z nich prispel ku konečnej vzdialenosti rovnako.

Techniky normalizácie dát:

- Min-Max normalizácia. Tento druh normalizácie spočíva v naškálovaní atribútu do istého intervalu (min - max). Štandardným intervalom je interval $[0, 1]$, kde normalizovanú hodnotu atribútu získame ako $x' = \frac{x-min(x)}{max(x)-min(x)}$. Tento spôsob je jednoducho rozšíriteľný na akýkoľvek interval $[a, b]$ a to spôsobom: $x' = \frac{x-min(x)}{max(x)-min(x)} * (b - a) + a$
- Desatinné škálovanie (decimal scaling). Jedná sa o normalizáciu takým spôsobom, že každá hodnota daného atribútu sa vynásobí rovnakou celočíselnou mocninou 10.
- Štandardizácia každú hodnotu atribútu upraví spôsobom: $x' = \frac{x-\bar{x}}{\sigma}$, kde $\bar{x}$ je stredná hodnota atribútu a σ jeho štandardná odchylka.
- Eliminácia odľahlých hodnôt. Táto technika spočíva v nájdení odľahlých hodnôt a následnom vymazaní alebo nahradení týchto hodnôt.

1.6 Anonymizácia dát

Anonymizácia dát je úprava dát za účelom ochrany súkromia. Je to proces, pri ktorom sú z data setov zašifrované alebo odstránené informácie, ktoré vedú k jednoznačnej identifikácii človeka.

1.7 Detekcia anomálií

Detekcia anomálií predstavuje problém nájdenia vzorov v dátach, ktoré nedosahujú očakávané správanie. Tieto nevyhovujúce vzory sú často označované ako anomálie alebo odľahlé hodnoty. Detekcia anomálií nachádza rozsiahle uplatnenie v širokej škále aplikácií, ako je detekcia chýb v bezpečnostných systémoch, vojenský dohľad nad nepriateľskými aktivitami alebo tiež detekcia anomálií medzi leteckými pasažiermi.

1.7.1 Čo sú to anomálie?

Anomálie sú vzory v dátach, ktoré nezodpovedajú normálnemu chovaniu. Možno ich spôsobiť v dátach rôznymi spôsobmi, ako je škodlivá činnosť, napríklad podvody s kreditnými kartami, teroristická činnosť alebo porucha systému. Všetky tieto neštandardné vzory majú istú hodnotu a to „zaujímavosť“ alebo význam v reálnom živote, čo je hlavným rysom detekcie anomálií.

1.7.2 Problematickosť domény

Na abstraktnej úrovni, anomália je definovaná ako vzor, ktorý nie je v súlade s normálnym chovaním. Jednoduchým prístupom pre detekciu anomálií

je preto vymedziť rozsah reprezentujúci normálne správanie a každé pozorovanie/záznam, ktoré nepatrí do tohto rozsahu označiť ako anomáliu. Avšak, niekoľko faktorov spôsobuje, že tento zdanlivo jednoduchý prístup sa stáva náročným:

- Definovanie tejto oblasti, ktorá zahàňa všetko možné normálne správanie je veľmi ťažké. Taktiež hranica medzi normálnym a abnormálnym chovaním často nie je presná.

- Keď sú anomálie výsledkom škodlivých akcií, útočníci sa snažia javiť ako bežní užívatelia, preto aj ich akcie sú často veľmi podobné akciám bežných užívateľov, čím sa zase sťažuje detekcia týchto útokov.

- V mnohých doménach sa toto normálne správanie zase časom vyvíja a čo bolo normálnym správaním v minulosti, už v budúcnosti normálnym správaním byť nemusí.

- Presný pojem anomálie sa líši v rôznych aplikačných oblastiach. Napríklad, v medicínskej oblasti už malá odchýlka od normálu (napríklad kolísanie telesnej teploty) môže byť anomália, zatiaľ čo podobná odchylka na burze cenných papierov (napríklad výkyvy v hodnote akcie) by mohla byť považovaná za normálnu. Vyvinutie jednej stratégie pre detekciu anomálií teda nemusí byť aplikovateľná na inú doménu.

- Dostupnosť označených dát pre učenie a validáciu modelov je tiež často problémom.

- Dáta často obsahujú šum, ktorý má tendenciu byť podobný reálnym anomáliám a preto je ťažké ich rozlíšiť a odstrániť.

Vzhľadom k vyššie uvedeným problémom, je problém detekcie anomálií vo svojej najvšeobecnejšej forma obtiažne vyriešiť. V skutočnosti väčšina súčasných techník detekcie anomálií rieši jednu konkrétnu formuláciu problému. Formulácia je vyvolaná rôznymi faktormi, ako je povaha dát, dostupnosť označených dát, typu anomálie, ktorú sa snažíme detekovať, atď. Tieto faktory sú určené doménou v ktorej anomálie hľadáme. Pri riešení tohto problému sa využívajú poznatky z rozmanitých odborov, ako je štatistika, machine learning a data mining.

1.8 Rôzne aspekty problému detekcie anomálií

Ako som už spomenul, konkrétna formulácia problému je daná niekoľkými faktormi, ako je povaha vstupných dát, dostupnosť (či nedostupnosť) značených dát.

1.8.1 Povaha vstupných dát

Kľúčovým aspektom akejkoľvek techniky detekcie anomálií je povaha vstupných dát. Vstup je obvykle kolekcia inštancií dát. Každá inštancia dát je označená sadou atribútov/dimenzií. Atribúty môžu byť rôznych druhov(numerické, binárne, atď). Povaha atribútov určuje použiteľnost techník na detekciu anomálií. Napríklad pre techniky založené na metóde najbližšieho suseda (NN) potrebujeme atribúty pre ktoré vieme určiť vzdialenosť medzi dvoma inštanciami/záznamami.

Vstupné dáta môžu tiež byť klasifikované na základe vzťahu medzi nimi. Väčšina existujúcich techník na detekciu anomálií funguje na základe obdržaných alebo nameraných dát (alebo bodových údajov), v ktorých sa nepredpokladá žiadny vzťah medzi inštanciami dát. Všeobecne však platí, že inštancie dát môžu byť vo vzájomnom vzťahu. Niektoré príklady sú dáta sekvencií, priestorové údaje a grafové dáta. V sekvenčných dátach sú jednotlivé inštancie zoradené, napríklad na základe času (časové postupnosti), sekvencie genómov a iné. V priestorových dátach, každá inštancia dát sa vzťahuje k jeho susedným inštanciách. Keď priestorové dáta majú aj časovú (sekvenčnú) zložku sú označované ako časopriestorové dáta, napríklad dáta o klíme, alebo letecké dáta. V grafových dátach, inštancie sú reprezentované ako vrcholy v grafe a sú prepojené s ďalšími vrcholmi hranami.

1.8.2 Druhy anomálií

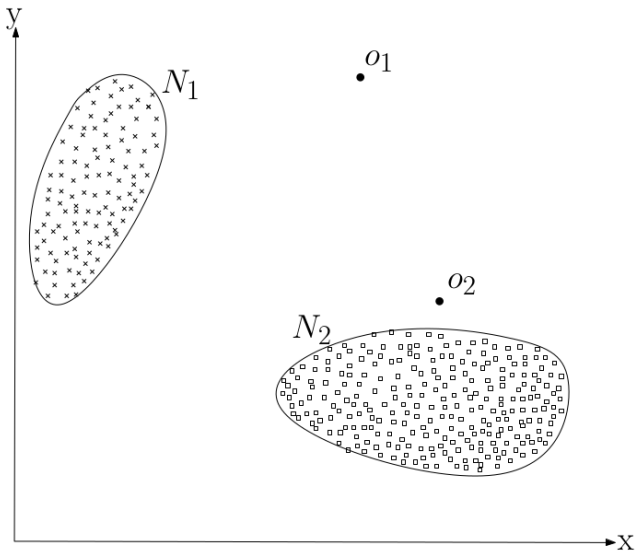
Dôležitým aspektom techniky detekcie anomálií je povaha požadovanej anomálie. Anomálie možno zaradiť do troch kategórií:

1.8.2.1 Bodové anomálie

Ak jednotlivé inštancie dát môžu byť považované za anomálne vzhľadom ku zvyšku dát, potom je táto inštancia bodovou anomáliou. Jedná sa o najjednoduchší typ anomálie. Ako príklad z reálneho života zoberme detekciu podvodov s kreditnými kartami. Súbor dát obsahuje transakcie kreditnou kartou. Predpokladajme, že dáta sú definované použitím iba jedného atribútu: zaplatená suma. Transakcie, pre ktoré je táto suma veľmi vysoká v porovnaní s ostatnými výdavkami bude klasifikovaná ako bodová anomália.

1.8.2.2 Kontextové anomálie

Ak je inštancia dát anomálnou v špecifickom kontexte (inak nie), potom sa nazýva kontextuálna anomália (tiež označovaný ako podmienené anomálie). Kontext je tvorený štruktúrou v súbore dát, a musí byť zadaný ako súčasť formulácie problému. Každá inštancia dát je definovaná dvoma typmi atribútmi:



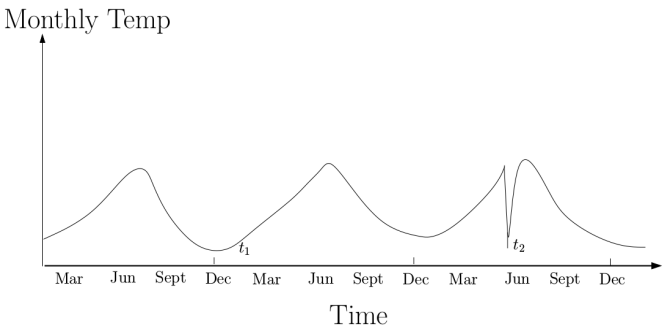
Obr. 1.1: Bodové anomálie[1]

1. Kontextové atribúty určujú kontext (alebo susednosť) pre túto inštanciu. Napríklad v priestorových dátach, zemepisná dážka a šírka sú kontextové atribúty. V dátach časových postupností, čas je kontextový atribút, ktorý určuje pozíciu jednej inštancie vrámci sekvencie.

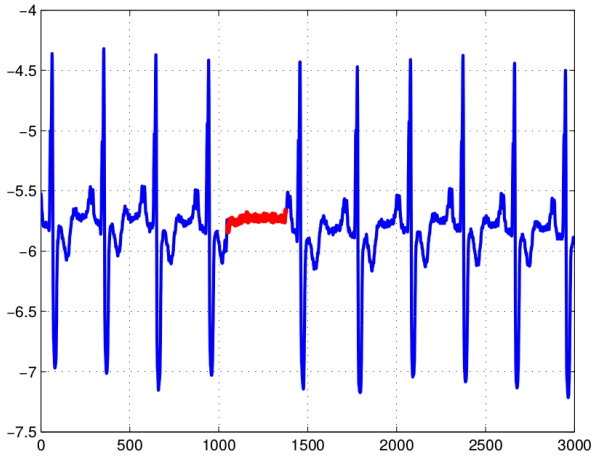
2. Behaviorálne atribúty definujú nekontextuálne charakteristiky inštancie. Napríklad v priestorových dátach priemerných zrážok z celého sveta, je množstvo zrážok v akomkoľvek mieste behaviorálny atribút.

Anomálne správanie je stanovená s použitím hodnôt pre behaviorálne atribúty v určitom kontexte. Inštancia dát môže byť kontextuálnou anomáliou v danom kontexte, ale identická inštancia dát (pokiaľ ide o behaviorálne atribúty) by mohla byť považovaná za normálnu v inom kontexte. Táto vlastnosť slúži k identifikácii kontextových a behaviorálnych atribútov pre techniku detekcie kontextových anomálií.

Kontextové anomálie sú najčastejšie skúmané v časových a priestorových dátach. Voľba použitia techniky detekcie kontextových anomálií závisí od zmysluplnosti kontextuálnych anomálií v doméne cieľovej aplikácie. ďalším kľúčovým faktorom je dostupnosť kontextových atribútov. V niekoľkých prípadoch je definícia kontextu jednoduchá a teda aj použitie metód pre detekciu kontextových anomálií dáva zmysel. V iných prípadoch môže byť definovanie kontextu obtiažne, čo znemožňuje použitie mnohých metód.



Obr. 1.2: Kontextová anomália[1]



Obr. 1.3: Kolektívna anomália[1]

1.8.2.3 Kolektívne anomálie

Ak kolekcia súvisiacich inštancií je anomálna vzhľadom na celý súbor dát, nazýva sa kolektívna anomália. Individuálne inštancie dát v kolektívnej anomálii nemusia byť anomálie samy o sebe, ale ich výskyt spolu ako kolekcia je abnormálny.

Je potrebné poznamenať, že zatiaľ čo bodová anomália sa môže objaviť v každom súbore dát, ku kolektívnym anomáliám môže dôjsť iba v dátach, ktorých inštancie spolu súvisia. Na druhú stranu, výskyt kontextových anomálií závisí od dostupnosti kontextových atribútov v dátach. Bodová alebo kolektívna anomália môže byť tiež kontextovou ak je analyzovaná vzhľadom na kontext. Problém detekcie bodových alebo kolektívnych anomálií môže byť transformovaný na detekciu kontextuových anomálií začlenením kontextových atribútov.

1.8.3 Označenie dát

Označenie dát hovorí o tom, či inštancia je normálna alebo anomália. Avšak získanie značených dát, ktoré sú presnou reprezentáciou všetkých typov chovania je často nemožné. Značenie sa často vykonáva ručne a preto sa vyžaduje značná snaha na získanie značených dát pre trénovanie. Zvyčajne je obtiažnejšie získať značené dáta pre normálne chovanie ako pre všetky typy anomálií.

Navyše anomálne správanie má často dynamický charakter, môžu vzniknúť nové typy anomálií, pre ktoré nie sú k dispozícii žiadne značené trénovacie dáta. V niektorých prípadoch, ako je napríklad bezpečnosť letovej prevádzky, anomálne prípady by mohli ústiť do katastrofických udalostí, a preto budú veľmi vzácne. Na základe rozsahu, v akom sú k dispozícii značené dáta, detekcia anomálií môže detekcia anomálií prebiehať nasledovnými spôsobmi:

1.8.3.1 Supervised detekcia anomálií

Techniky natrénované supervised predpokladajú dostupnosť trénovacích dát, ktoré obsahujú inštancie pre bežné ako aj anomálne triedy. Typický prístup v takýchto prípadoch je vybudovanie prediktívneho modelu pre klasifikáciu normálnych vs. anomálnych inštancií. Akúkoľvek inštanciu dát je potom možné pomocou tohto modelu klasifikovať. Existujú dva hlavné problémy ktoré vznikajú v supevised detekcii anomálií. Po prvé, anomálne prípady sú ďaleko menej frekventované v porovnaní s bežnými prípadmi v dátach pre trénovanie modelu [6].

Po druhé, získanie presných a reprezentatívnych označení, najmä pre triedu anomálií je zvyčajne náročné. Niekoľko techník bolo navrhnutých tak, aby vkladali umelé anomálie medzi normálne dáta pre získanie obsiahlejšieho trénovacieho setu dát [7][8].

1.8.3.2 Semi-supervised detekcia anomálií

Semi-supervised detekcia anomálií znamená detekovať anomálie, ak máme označené len normálne inštancie. Keďže tieto techniky nepotrebujú označenie anomálnej triedy, sú všeobecne viac uplatniteľné. Typickým prístupom týchto metód je vytvoriť model reprezentujúci normálne dáta a tento model následne použiť na identifikáciu anomálií (ktoré tomuto modelu neodpovedajú). Naopak modely natrénované len na anomálnych inštanciách sú neobvyklé, keďže je obtiažne zachytiť každý možný druh anomálie.

1.8.3.3 Unsupervised detekcia anomálií

Tieto techniky nevyužívajú trénovacie dáta pre žiadnu z tried a teda sú použiteľné najviac. Metódy v tejto kategórií predpokladajú, že normálne inštancie

sú ďaleko viac frekventované ako anomálne v testovacích dátach (inak by anomálne inštancie mohli byť považované za druh normálneho chovania a teda detekcia by neprebiehala správne).

1.8.4 Výstup detekcie anomálií

Dôležitým aspektom detekcie je tiež požadovaný výstup, ktorým sú anomálie identifikované. Typicky sa jedná o dva typy výstupov:

1.8.4.1 Skóre

Tieto techniky priraďujú každej inštancii z testovacieho data setu isté skóre, ktoré určuje mieru anomálnosti. Výstupom je teda ohodnotený zoznam anomálií. Za anomálie môžeme označiť zvolené množstvo ištancií s najvyšším anomálnym skóre, alebo zvoliť istú hranicu skóre a označíme za anomálne všetky inštancie, ktoré dosiahli vyššie skóre.

1.8.4.2 Označovanie

Techniky využívajúce označovanie (za normálnu alebo anomálnu inštanciu) priradzujú každej inštancii „nálepku“.

Techniky využívajúce skóre umožňujú analytikovi priamo ovplyvňovať citlivosť detekcie anomálií. Na druhú stranu označovacie metódy neposkytujú možnosť túto citlivosť ovplyvniť priamo, ale cez nastavovanie jednotlivých parametrov vrámci týchto metód.

1.9 Detekcia anomálií založená na klasifikácii

Používa vo dvoch krokoch a to naučenie modelu na označených dátach (trénovanie) a následnej klasifikácii inštancií, o ktorých chceme zistiť či sú anomáliou alebo nie (testovanie)[7]. Pri tomto prístupe predpokladáme, že model dokážeme naučiť na základe zadaného priestoru.

Na základe počtu „nálepiek“ rozdeľujeme techniky na one-class a multi-class detekcie anomálií. Ako pri one-class detekcii máme len jednu triedu pre normálne dáta, tak v multi-class máme viac druhov normálneho správania a preto vieme rozoznávať medzi nimi. V tomto prípade je inštancia anomálna, ak ju ani jeden z klasifikátorov pre normálne triedy neklasifikuje ako normálnu. Niektoré techniky tiež využívajú mieru istoty klasifikátora so svojím rozhodnutím. Ak žiadny z klasifikátorov nemá túto mieru vysokú pri tom, ako inštanciu klasifikuje ako normálnu, rozhodneme, že táto inštancia je anomálna.

1.9.1 Neurónové siete

Neurónové siete sa využívajú ako pri multi-class, tak aj pri one-class detekcii anomálií. Základnou myšlienkou je natrénovať neurónovú sieť na normálnych

dátach (naučiť ju rozpoznávať rôzne normálne triedy) a následne v testovacej fázi použiť inštanciu, ktorú chceme klasifikovať ako vstup do neurónovej siete. Ak ju prijme, jedná sa o normálnu inštanciu, ak nie o anomáliu[9].

1.9.2 Bayesovské siete

Bayesovské siete sa využívajú pri multi-class detekcii anomálií. Tento spôsob je založený na určení posteriornej pravdepodobnosti, že inštancia patrí do danej triedy. Keďže máme viac tried, zvolíme ako výslednú triedu tejto inštancie tú s najväčšou pravdepodobnosťou [10]. Závislosti medzi jednotlivými atribútmi a výslednou triedou sú získané z trénovacej množiny. Táto technika predpokladá nezávislosť medzi atribútmi. Niektoré techniky tiež zachytávajú závislosti medzi rôznymi atribútmi využívajúc komplexné Bayesovské siete [11].

1.9.3 Support vector machine

Support vector machine (SVM) sa využíva pri one-class detekcii anomálií [12]. Táto technika pracuje tak, že sa snaží zachytiť normálne správanie oblasťou, ktorá zachytáva trénovacie dáta. Pre komplexné normálne oblasti sa využívajú rôzne jadrové funkcie (napríklad radial basis function - RBF[13]). Klasifikácia následne prebieha pozorovaním, či testovaná inštancia spadá do naučeného regiónu a je normálnou alebo nespadá a je anomáliou. Niektoré techniky sú schopné trénovať SVM v lineárnom čase [14].

1.9.4 Techniky založené na pravidlách

Ako všetky klasifikátory, aj tento spôsob sa snaží zachytiť normálne chovanie dát. Ak inštancia, ktorú testujeme nie je zachytená žiadnym pravidlom, predpokladáme, že sa jedná o anomáliu. Tieto metódy sa používajú ako na multi-class tak aj one-class detekciu [8].

Prvou fázou je tiež trénovanie na základe trénovacej množiny, kde sa objavujú pravidlá v dátach. Typickými reprezentantmi týchto metód sú napríklad rozhodovacie stromy. Každé získané pravidlo má priradenú takzvanú confidence hodnotu, ktorá je podielom počtu inštancií ktoré spĺňajú toto pravidlo a všetkých inštancií, ktoré sú zahrnuté týmto pravidlom. Druhým krokom je samotná detekcia anomálií. Keďže by sme už mali mať zachytené normálne chovanie pravidlami, čo sme vytvorili, pre testovanú inštanciu zvolíme to najviac vyhovujúce pravidlo. Anomálnym skóre budeme nazývať prevrátenú hodnotu confidence tohto najviac vyhovujúceho pravidla.

Dolovanie asociačných pravidiel je taktiež používané na one-class detekciu anomálií a to generovaním pravidiel bez učiteľa (unsupervised) [15]. Aby sme predišli uplatňovaniu pravidiel s veľmi nízkym supportom (podiel počtu trénovacích inštancií k celkovému počtu inštancií) môžeme zvoliť istú hranicu a pravidlá so supportom menším ako táto hranica nebrať v úvahu.

1.9.5 Výhody a nevýhody

Výhody:

1. Pri presnom a dostatočne obsiahlom trénovacom data sete vieme zachytiť rôzne triedy normálneho správania a tým veľmi presne detekovať anomálie.

2. Testovanie prebieha rýchlo, keďže len využívame už natrénovaný model.

Nevýhody:

1. Multi-class detekcia sa spolieha na dostupnosť presne označených dát, čo v mnohých prípadoch vôbec nie je reálne

2. Výstupom týchto metód je „nálepka“ a nie anomálne skóre, ktoré je častokrát viac odpovedajúce.

1.10 Detekcia anomálií založená na metóde najbližšieho suseda

Prístupy založené na metóde najbližšieho suseda predpokladajú, že normálne dáta sa sú v zhlukoch zatiaľ čo anomálie sa vyskytujú ďaleko od svojho najbližšieho suseda [16]. Všetky tieto techniky tiež vyžadujú mieru, podľa ktorej môžeme jednotlivé inštancie porovnávať a tým získať istú mieru podobnosti alebo vzdialenosť medzi nimi. Pre spojité atribúty je klasickou voľbou Euklidovská vzdialenosť, pre iné je často potrebné použiť nejakú komplexnejšiu mieru. Ak inštancia obsahuje rôzne druhy atribútov je táto vzdialenosť obyčajne spočítaná pre jednotlivé atribúty zvlášť a následne skombinovaná.

Techniky založené na metóde najbližšieho suseda sa všeobecne delia na dve kategórie

1. Techniky využívajúce vzdialenosť ku k-temu susedovi ako anomálne skóre

2. Techniky počítajúce relatívnu hustotu susedov pre každú inštanciu

1.10.1 Techniky využívajúce vzdialenosť ku k-temu susedovi

Pri tomto prístupe je anomálne skóre inštancií počítané ako vzdialenosť ku k-temu susedovi. Citlivosť detekcie môžeme ovplyvňovať parametrom k, ale aj zvolením istej hranice anomálneho skóre alebo namiesto tejto hranice zvoliť n inštancií s najvyšším anomálnym skóre a prehlásiť ich za anomálie [17].

Iným spôsobom, ako vypočítať anomálne skóre je spočítať susedov (n), ktorí nie sú ďalej ako d[18][19][20]. Jedná sa o určovanie globálnej hustoty, keďže počítame susedov v hyperguli o polomere d so stredom v danej inštancii. Avšak anomálne skóre by malo stúpať ak predpokladáme s vyššou pravdepodobnosťou že sa jedná o anomáliu. Preto sú dva rôzne prístupy:

1. Stanoviť fixné d a anomálne skóre zvoliť ako $1/n$

2. Stanoviť fixné n a anomálne skóre zvoliť ako $1/d$

Keďže výpočetná zložitosť pri týchto metódach je $O(n^2)$, kde n je počet inštancií (rátame vzájomné vzdialenosti medzi všetkými inštanciami), mnohé prístupy sa snažia vylúčiť inštancie, ktoré nemôžu byť anomálne. Medzi tieto prístupy patrí napríklad technika, kde sa najskôr dáta rozdelia do zhlukov (clustering), v ktorých sa vypočíta spodná a horná hranica pre vzdialenosť od k-teho najbližšieho suseda. Táto informácia je následne použitá na identifikáciu partícií, v ktorých sa nemôže nachádzať k inštancií s najvyšším anomálnym skóre a ďalej ich neberieme v úvahu (anomálie hľadáme vo zvyšných partíciách). Ďalším prístupom ako zefektívniť túto metódu je hľadať najbližšieho suseda vrámci malej vzorky z data setu, čím sa zníži zložitosť na $O(mn)$, kde m je počet inštancií vo zvolenej vzorke.

1.10.2 Techniky počítajúce relatívnu hustotu susedov

Tieto techniky počítajú relatívnu hustotu susedov pre každú inštanciu. Inštancie, ktoré ležia v hustom susedstve označujeme za normálne a naopak tie, ktoré v riedkom označujeme za anomálne. Pre zadanú inštanciu, vzdialenosť k jej k-temu susedovi odpovedá polomeru hypergule so stredom v tejto inštancii zahŕňajúcej k najbližších susedov našej inštancie. Z toho plynie, že táto vzdialenosť môže byť považovaná za inverziu k hustote a teda základná technika využívajúca vzdialenosť ku k-temu susedovi môže byť tiež technikou počítajúcou s relatívnou hustotou susedov.

Metódy rátajúce s touto hustotou nemusia pracovať správne nad dátami, kde sú oblasti s rôznymi hustotami výskytu inštancií. Aby sa tomuto predišlo, zaviedli sa metódy, ktoré zohľadňujú relatívnu hustotu svojich susedov. Jedným z riešení je napríklad Local Outlier Factor (LOF)[21]. Pre danú inštanciu dát, LOF skóre je pomer priemernej hustoty k najbližších susedov a lokálnej hustoty tejto inštancie. Pre vypočítanie tejto lokálnej hustoty najskôr nájdeme polomer najmenšej hypergule, ktorá obsahuje k najbližších susedov a následne vydeleniu k jej objemom. Pre normálne inštancie bude lokálna hustota podobná ako hustota ich susedov, pričom anomálne inštancie budú mať túto hustotu menšiu (jej LOF skóre bude vyššie).

Výpočetná zložitosť je pri LOF zase $O(n^2)$, kde n je počet inštancií a preto existujú rôzne modifikácie:

- Connectivity-based Outlier Factor (COF)[22]. Funguje inkrementálne, do okolia sa pridáva vždy inštancia, ktorá je najbližšie k súčasnému okoliu (najmenšia vzdialenosť od akejkoľvek inštancie v okolí) až kým nedosiahneme veľkosť okolia k. Následne sa anomálne skóre spočíta rovnako ako pri LOF.

- Outlier Detection using In-Degree Number (ODIN)[16]. Pre každú inštanciu spočítame počet k najbližších inštancií, pre ktoré sa zadaná inštancia nachádza v ich k najbližšom okolí. Prevrátená hodnota tohto počtu je anomálne skóre inštancie.
- Multi-granularity Deviation Factor (MDEF)[23]. Pre danú inštanciu spočítame štandardnú odchýlku lokálnych hustôt najbližších susedov (aj samotnej inštancie). Prevrátenou hodnotou tejto odchýlky je anomálne skóre inštancie.

1.10.3 Výhody a nevýhody

Výhody:

1. Jedná sa o unsupervised metódy detekcie anomálií a nepredpokladáme žiadne tvrdenia ohľadom distribúcie dát. Sú čisto založené na dátach.
2. Prispôsobovanie týchto metód na rôzne dáta je priamočiare, jediné čo je pri tom potrebné je mať mieru podobnosti pre inštancie.

Nevýhody:

1. Ak majú anomálie dostatok blízkych susedov a tiež naopak ak normálne inštancie majú málo blízkych susedov sa môže stať, že detekcia neprebehne správne.
2. Výpočetná zložitosť pri týchto metódach je vysoká (bežne $O(n^2)$), keďže musíme rátať vzdialenosti medzi všetkými inštanciami, alebo inštanciami patriacimi do nejakého okolia inštancie.
3. Spoľahlivosť detekcie anomálií sa spolieha na zvolenú mieru podobnosti inštancií. Zvoliť mieru môže byť nadmieru obtiažna úloha, ak sa jedná o komplexné dáta (nespojité atribúty, postupnosti a iné).

1.11 Detekcia anomálií založená na zhlukovaní

Zhlukovanie (clustering)[24] sa používa na organizovanie podobných dát do zhlukov. Zhlukovanie je zvyčajne bez učiteľa, ale existujú aj semi-supervised prípady. Aj keď sa môže zdať, že zhlukovanie a detekcia anomálií sú dve odlišné veci, existujú metódy detekcie anomálií založené na zhlukovaní.

Tieto metódy sa delia do troch kategórií podľa predpokladov o dátach:

1. Normálne inštancie patria do zhluku, pričom anomálie nepatria do žiadneho.

2. Normálne inštancie ležia blízko centroidu najbližieho zhluku, zatiaľ čo anomálie ležia ďaleko.

3. Normálne inštancie patria do veľkých a hustých zhlukov, pričom anomálie patria do malých alebo riedkych zhlukov.

Techniky založené na prvom tvrdení označujú všetky inštancie, ktoré sme nezaradili do zhluku za anomálne (príkladom je algoritmus ROCK [25]). Nevýhodou týchto techník však je, že nie sú optimalizované na nachádzanie anomálií, ale ich cieľom je nájsť zhluky.

Metódy založené na druhom tvrdení pozostávajú z dvoch krokov. V prvom kroku sa dáta zhlukujú pomocou nejakého zhlukovacieho algoritmu. V druhom pre každú inštanciu vyrátame vzdialenosť od centroidu najbližšieho zhluku, čo následne berieme ako anomálne skóre. Bežne používanými algoritmami pre zhlukovanie sú napríklad zhlukovanie K-means, Self-Organizing Maps (SOM)[26]. Tieto techniky však neodhalia anomálie, ak budú tvoriť vlastný zhluk [27].

Tretia kategória metód označuje za anomálne také inštancie, ktoré patria do zhlukov, ktorých veľkosť alebo hustota je pod zvolenou hranicou[28]. Jednou z techník ako takéto anomálne skóre zvoliť je Cluster-Based Local Outlier Factor (CBOLF), ktorý je prakticky zhlukovou variantou Local Outlier Factor. Zahŕňa ako veľkosť zhluku, tak aj vzdialenosť od centroidu zhluku do ktorého patrí.

Výpočetná zložitosť týchto metód závisí od zvoleného algoritmu. Ak je potrebné vypočítať vzdialenosti medzi dvojicami inštancií, je zložitosť obyčajne kvadratická, ale na druhú stranu ak sú použité algoritmy založené na heuristike (napríklad K-means), môže byť zložitosť lineárna. Testovacia fáza je obyčajne rýchla, keďže porovnávame inštancie s obmedzeným množstvom zhlukov.

1.11.1 Rozdiely medzi zhlukovou metódou a metódou najbližšieho suseda

Niektoré metódy založené na zhlukovej analýze vyžadujú mieru podobnosti dvoch inštancií, v čom sú veľmi podobné metódam najbližšieho suseda. Voľba tejto miery do veľkej miery ovplyvňuje presnosť detekcie. Kľúčovým rozdielom medzi týmito technikami je fakt, že metódy založené na zhlukovej analýze vyhodnocujú inštanciu vzhľadom na zhluk, do ktorého patrí, pričom metódy najbližšieho suseda berú v úvahu lokálne okolie.

1.11.2 Výhody a nevýhody

Výhody:

1. Metódy zhlukovania dokážu fungovať bez učiteľa.

2. Tieto techniky sa dajú jednoducho prispôsobiť na komplexné dátové typy zvolením zhlukovacieho algoritmu, ktorý si s nimi vie poradiť.

3. Fáza testovania inštancie prebieha rýchlo, pretože porovnávame túto inštanciu s obmedzeným počtom zhlukov.

Nevýhody:

1. Presnosť detekcie je závislá na schopnosti zhlukovacieho algoritmu zachytiť zhluky normálnych inštancií.

2. Tieto algoritmy nie sú optimalizované na hľadanie anomálií, ich detekcia je len vedľajším produktom.

3. Niektoré algoritmy zaradia všetky dáta do zhluku a preto aj anomálie budú priradené do zhluku a tým pádom metódy predpokladajúce, že anomália nepatrí do žiadneho nebudú fungovať.

4. Niektoré algoritmy zase pracujú správne iba ak anomálie netvoria žiadne zhluky.

5. Výpočetná zložitosť zhlukovania môže byť vysoká.

1.12 Štatistická detekcia anomálií

Tieto metódy sa zakladajú na myšlienke, že anomáliou je taká inštancia, ktorá neodpovedá predpokladanému stochastickému modelu. Spolieha sa pritom na tvrdenie, že normálne dáta sa vyskytujú vo vysoko pravdepodobných oblastiach stochastického modelu, pričom anomálie naopak v oblastiach s nízkou pravdepodobnosťou [29].

Štatistické techniky detekcie fitujú štatistický model na dané dáta a následne sledujú či ďalšie inštancie patria do tohto modelu alebo nie. Inštancie, čo majú nízku pravdepodobnosť, že sú generované týmto modelom (na základe aplikovanej testovacej štatistiky) prehlásime za anomálie. Ako parametrické, tak aj neparametrické techniky sú využívané.

Zatiaľ čo parametrické techniky predpokladajú znalosť distribúcie a určujú parametre tejto distribúcie na základe daných dát [30], neparametrické techniky nepredpokladajú znalosť distribúcie.

1.12.1 Parametrické techniky

Predpokladáme, že dáta sú generované parametrickou distribúciou s parametrami θ a s hustotou pravdepodobnosti $f(x, \theta)$, kde x je pozorovanie. Anomálne skóre testovanej inštancie vypočítame ako prevrátenú hodnotu $f(x, \theta)$. Parametre θ určujeme na základe daných dát.

Alternatívnou možnosťou detekcie anomálií v tomto modeli je tiež testovanie hypotéz. Zvolíme nulovú hypotézu H_0 tak, že inštancia x bola generovaná predpokladanou distribúciou (s parametrami θ). Ak štatistický test zamietne hypotézu H_0, prehlásime x za anomáliu. Testovanie hypotéz je spojené s testovacou štatistikou, ktorá môže byť použitá na získanie pravdepodobnostného anomálneho skóre pre inštanciu x.

Na základe predpokladaného rozdelenia môžu byť tieto techniky ďalej delené.

1.12.1.1 Gaussovský model

Tieto techniky predpokladajú, že dáta boli generované Gaussovským rozdelením. Parametre sú určené pomocou metódy Maximum Likelihood Estimates (MLE). Vzdialenosť inštancie od priemeru je potom braná ako anomálne skóre. Pre označenie anomálií sa volí hranica a inštancie nad túto hranicu sú označené za anomálie. Rôzne techniky rátajú túto vzdialenosť od priemeru rôznym spôsobom.

Jednou z najjednoduchších detekcií odľahlých inštancií je označiť všetky inštancie, ktoré sú od priemeru μ vzdialené viac ako 3σ,kde σ je smerodatná odchylka rozdelenia. Oblasť $\mu \pm 3\sigma$ zahŕňa 99.7% inštancií.

Ďalšou jednoduchou metódou je využitie box plot rule. Box-plot graficky znázorňuje najmenšie neanomálne pozorovanie, dolný kvartil (Q_1), medián, horný kvartil (Q_3) a najväčšie neanomálne pozorovanie. $Q_3 - Q_1$ sa nazýva Inter Quartile Range (IQR). Box plot tiež indikuje, kedy pozorovanie pokladať za anomáliu. Inštancia dát, ktorá leží viac ako $1.5 * IQR$ pod Q_1, alebo $1.5 * IQR$ nad Q_3, je označovaná za anomáliu. Oblasť $Q_1 - 1.5 * IQR$ až $Q_3 + 1.5 * IQR$ obsahuje 99.3% pozorovaní a teda voľba $1.5 * IQR$ ako hranice anomálnosti je takmer ekvivalentná 3σ technike.

Grubbov test zase využíva výpočet z skóre pre každú inštanciu (predpokladáme jednorozmerné dáta) x: $z = \frac{|x-\bar{x}|}{s}$, kde $\bar{x}$ je priemer a s je štandardná odchylka vzorky dát. Inštancia je potom anomálna ak $z > \frac{(N-1)}{\sqrt{N}} * \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$, kde N je počet inštancií, $t^2_{\alpha/(2N),N-2}$ je hranica určujúca, či je inštancia anomálna (hodnota t-rozdelenia na hladine významnosti $\alpha/2N$) [31].

Varianta Grubbovho testu pre viacrozmerné dáta počíta s Mahalanobisovou vzdialenosťou inštancie od priemeru na redukovanie viacrozmerného priestoru do jednorozmerného skaláru.

$y^2 = (x - \bar{x})' S^{-1} (x - \bar{x})$

Následne je na y uplatnený Grubbov test podobne ako pri jednorozmerných dátach.

Jednou z ďalších variant detekcie je použitie χ^2 štatistiky. Predpokladáme, že máme viacrozmerné dáta s normálnym rozdelením. Potom je hodnota χ^2 štatistiky definovaná ako:

$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i}$

kde X_i je hodnota i-teho atribútu, E_i je priemerná hodnota i-teho atribútu (získaná z trénovacieho data setu) a n je počet atribútov. Veľká hodnota χ^2 značí, že sa v pozorovanej vzorke nachádzajú anomálie.

1.12.1.2 Regresný model

Detekcia anomálií využitím regresného modelu sa využíva na časové rady [32]. Základná myšlienka tohto prístupu spočíva v nafitovaní regresného modelu na dáta a v ďalšom kroku sa pre každú inštanciu vypočíta anomálne skóre ako rozdiel modelu a inštancie.

1.12.1.3 Kombinácia parametrických rozdelení

Táto kategória je rozdelená na dva smery:

1. Modelovanie normálnych a anomálnych inštancií odlišnými parametrickými rozdeleniami. Testovanie prebieha sledovaním, do ktorého rozdelenia patrí daná inštancia.

2. Modelovanie normálnych inštancií ako kombináciu parametrických rozdelení. Testovanie prebieha skúmaním, či daná inštancia patrí do nejakého naučeného rozdelenia. Ak nie, je prehlásená za anomáliu.

1.12.2 Neparametrické techniky

Techniky v tejto kategórii využívajú neparametrické štatistické modely, štruktúra dát nie je definovaná predom, ale je určená z dát.

1.12.2.1 Histogramy

Pre jednorozmerné dáta je základnou myšlienkou vytvoriť histogram nad týmito dátami a následne sledovať či testovaná inštancia spadá do niektorého z binov. Ak áno, je inštancia prehlásená za normálnu, ak nie, za anomálnu. Veľkosť binov je predmetom optimalizácie. Ak zvolíme príliš malé biny, môže sa stať, že aj normálne inštancie budú spadať do prázdnych oblastí a tým pádom

budú nesprávne detekované. Naopak pri príliš veľkých binoch môžu zase byť anomálie klasifikované ako normálne inštancie [30].

Pre viacrozmerné dáta histogramová metóda pracuje s atribútmi oddelene a vždy sledujeme veľkosť binu, do ktorého hodnota atribútu spadá a následne tieto veľkosti sčítame. Anomálne skóre získavame ako prevrátenú hodnotu týchto veľkostí [33].

1.12.3 Výhody a nevýhody

Výhody:

1. Ak sú splnené predpoklady, štatistické metódy poskytujú štatisticky dokázateľné riešenie pre detekciu anomálií.

2. Anomálne skóre je spojené s konfidenčným intervalom, čo môže byť použité pri voľbe hranice.

Nevýhody:

1. Štatistické metódy sa spoliehajú na predpoklady o dátach. Tieto predpoklady častokrát nie sú splnené (zvlášť pri multidimenzionálnych dátach [31]) a teda štatistické metódy sú nepoužiteľné.

2. Aj keď sú predpoklady splnené, testovanie hypotéz je obtiažnou úlohou (napríklad už zostaviť testovaciu hypotézu pre dáta s vysokou dimenziou je netriviálne).

3. Histogramové metódy sú síce jednoduché na implementáciu, ale nie sú schopné zachytiť závislosti medzi jednotlivými atribútmi (anomália môže mať hodnoty atribútov normálne, ale ich kombinácia môže byť nezvyčajná).

1.13 Teória informácie

Techniky založené na teórii informácie analyzujú informačný obsah data setu použitím rôznych mier ako Kolomogorova zložitosť, entropia a iné. Predpokladáme, že anomálie spôsobujú nepravideľnosti v informačnom obsahu data setu [34].

Nech $C(D)$ značí zložitosť daného data setu, D. Základná technika založená na teórii informácie hľadá minimálnu podmnožinu D, I takú, že $C(D) - C(D-I)$ je maximálne. Všetky inštancie v tejto podmnožine sú následne označené za anomálne. Hľadáme teda paretooptimálne riešenie, kaďže sa optimalizuú dve zložky. Spomínaná zložitosť C môže byť zvolená rôznymi spôsobmi [13][35][34].

1.13.1 Výhody a nevýhody

Výhody:

1. Tieto metódy sú schopné pracovať bez učiteľa.
2. Nekladú žiadne predpoklady o štatistickom rozdelení dát.

Nevýhody:

1. Spoľahlivosť týchto metód je do vysokej miery ovplyvnená výberom miery. Často tieto miery dokážu detekovať anomálie iba ak sa v data sete nachádzajú vo väčších počtoch.
2. Tieto techniky neposkytujú anomálne skóre.

1.14 Spektrálne techniky

Spektrálne techniky sa snažia o aproximáciu dát použitím kombinácie atribútov zachytávajúcich rozptyl v dátach. Predpokladáme, že dáta môžu byť transformované do priestoru s nižšou dimenziou, kde sa normálne a anomálne inštancie javia značne odlišné.

Niektoré z týchto techník využívajú Principal Component Analysis (PCA) pre projekciu dát do nového priestoru [36]. Jednou z nich je napríklad analýza projekcie každej inštancie do hlavných komponent s nízkym rozptylom. Normálna inštancia, ktorá odpovedá korelačnej štruktúre má nízku hodnotu projekcie zatiaľ čo anomália vysokú.

Spektrálnou technikou na hľadanie anomálií v časových radách grafov je napríklad reprezentovať graf ako maticu susednosti pre daný časový okamih. Pre každú časovú inštanciu bude zvolený vektor aktivity (zmeny) ako hlavná komponenta. Časová rada týchto vektorov je braná ako matica a z nej získavame hlavný ľavý singulárny vektor (principal left singular vector) pre zachytenie normálnych závislostí v dátach vzhľadom na čas. Pre nový záznam (graf) získavame jeho anomálne skóre ako uhol medzi týmto vektorom a vektorom aktivity nového záznamu [37].

1.14.1 Výhody a nevýhody

Výhody:

1. Tieto techniky sú vhodné na analýzu vysokodimenzionálneho priestoru, keďže ho redukujú. Tiež môžu byť použité ako predspracovanie pre iné techniky.
2. Vieme ich aplikovať v prostredí bez učiteľa.

Nevýhody:

1. Sú použiteľné len ak sú anomálie a normálne inštancie separabilné v priestore s nižšou dimenziou.
2. Vysoká výpočetná zložitosť.

1.15 Kontextové anomálie

Predchádzajúce techniky boli primárne zamerané na identifikáciu bodových anomálií. Detekcia kontextových anomálií vyžaduje, aby dáta mali kontextuálne a behaviorálne atribúty. Kontextuálne atribúty môžu byť:

- Priestorové - máme polohu a tým pádom aj priestorové okolie [38].
- Grafové - máme hrany, ktoré spájajú jednotlivé uzly (inštancie), čím sa zase určuje okolie.
- Sekvenčné - atribúty, ktoré určujú pozíciu v postupnosti. Jedná sa napríklad o časové rady [32][39].
- Profilové - sú to atribúty, ktoré zaraďujú inštancie do skupín (profilovanie), vrámci ktorých sa potom testuje anomálnosť.

Techniky zaoberajúce sa kontextovými anomáliami môžeme deliť na dve kategórie:

1. Redukcia problému na bodovú detekciu anomálií a následne použitie niektorého z opísaných prístupov.
2. Modelovanie štruktúry v dátach a následne použitie tohto modelu na detekciu anomálií.

1.15.1 Redukcia problému na bodovú detekciu anomálií

Keďže kontextové anomálie sú inštancie, ktoré sú anomálne len vzhľadom na kontext, jedným z prístupov je aplikovať bodovú detekciu anomálií v tomto kontexte.

Táto redukcia najskôr určí kontext pre každú z inštancií využívajúc kontextuálne atribúty a následne vypočíta anomálne skóre pomocou niektorej z techník bodovej detekcie anomálií.

1.15.2 Využitie štruktúry dát

V niektorých prípadoch nie je rozdelenie na kontexty priamočiare (typicky pre časové rady). Základnou myšlienkou tohto prístupu je naučenie modelu na trénovacích dátach, tak aby vedel určovať behaviorálne atribúty na základe kontextu. Ak je očakávané chovanie iné, predpokladáme anomáliu.

1.15.3 Výhody a nevýhody

Výhody: Sú schopné detekovať anomálie, ktoré by nemuseli byť odhalené bodovými detekciami. Nevýhody: Sú aplikovateľné len keď môže byť kontext jasne definovaný.

1.16 Kolektívne anomálie

Je to taká podmnožina inštancií, ktorých výskyt ako celku je neobvyklý.

Primárnym predpokladom pre detekciu kolektívnych anomálií sú závislosti medzi inštanciami dát.

1.16.1 Sekvenčné anomálie

Tieto anomálie môžu byť rozdelené do troch kategórií:

1.16.1.1 Detekcia anomálnej sekvencie v množine sekvencií

Tieto techniky pracujú semi-supervised, alebo unsupervised. Najväčšími problémami v tejto oblasti sú rozdielne dĺžky sekvencií a tiež rozdielne zarovnanie.

1. Prvým prístupom ako tieto anomálie detekovať je zase redukcia na bodovú detekciu anomálií. Snažíme sa teda jednotlivé sekvencie previesť do konečného priestoru a v ňom aplikujeme jednu z metód bodovej detekcie.

2. Druhým je modelovanie sekvencií. Najčastejšou metódou na toto modelovanie je pomocou Markovských modelov.

1.16.1.2 Detekcia anomálnej subsekvencie v sekvencii

Jedná sa o detekciu anomálneho vzoru vrámci sekvencie udalostí alebo časovej rady [40]. Táto detekcia pracuje zvyčajne v unsupervised móde a teda predpokladá, že sa časová rada odpovedá definovanému vzoru. Táto detekcia naráža opäť na problémy. Jedným z najzávažnejších je fakt, že vo všeobecnosti nepoznáme dĺžku anomálnej sekvencie [32][41].

1.16.1.3 Detekcia, či frekvencia vzoru v sekvencii nie je anomálna

Detekovať tento typ anomálií znamená nájsť vzory, ktorých frekvencia výskytu v inštancii sa líši od frekvencie v normálnom data sete[42]. Bežne sa využíva metóda pohyblivého okna[43].

1.16.2 Priestorové anomálie

Kolektívna detekcia anomálií v priestorových dátach zahŕňa nachádzanie podgrafov alebo subkomponent v dátach, ktoré sú anomálne. Táto kategória je značne nepreskúmaná.

Kapitola 2

Vstupy

Našimi vstupmi sú dáta z OBZORu 1.1.1.

2.1 Dáta

Dáta z obzoru sú organizované do adresárovej štruktúry podľa roku a mesiaca príletu. Jednotlivé lety sú uložené v separátnych súboroch (.csv, .xlsx), kde každý záznam odpovedá jednému pasažierovi. Každý zo záznamov pozostáva z niekoľkých atribútov jednotlivca a to:

1. FlightNumber - Číslo letu. Keďže sa jedná o kombináciu písmen a čísel o obmedzenej dĺžke, nie je unikátnym identifikátorom letu, nieto ešte pasažiera.

2. ScheduledArrival - Plánovaný čas a dátum príletu. Tiež sa nemusí jednať o jednoznačnú identifikáciu letu, keďže v jeden čas môže pristávať aj viac letov.

3. Nationality - Národnosť pasažiera. Zakódovaná v trojpísmenových skratkách štátu (CZE, SVK atp.).

4. Surname - Priezvisko pasažiera.

5. Names - Všetky zvyšné mená pasažiera.

6. BirthDate - Dátum narodenia pasažiera.

7. Sex - Pohlavie pasažiera. Nadobúda hodnôt - M pre muža, F pre ženu a U.

8. DocumentType - Pri odlete sa udáva identifikačný dokument pasažiera. Malo by sa jednať o buď pas alebo občiansky preukaz (ak je pasažier občanom členského štátu európskej únie).

9. DocumentIssued - Štát, v ktorom bol daný dokument vydaný. Zakódovaný v trojpísmenových skratkách štátu (CZE, SVK atp.).
10. DocumentNumber - Číslo tohoto dokumentu.
11. FlightFrom - Kód letiska, z ktorého let odlieta.
12. FlightTo - Kód letiska, na ktoré let prilieta. Spolu s FlightNumber, ScheduledArrival, FlightFrom môže byť použitý ako jednoznačná identifikácia letu.
13. Reservation - Ak má záznam aj tento atribút, tak sa jedná o rezerváciu dopredu. Jedná sa o kód rezervácie. Ak ho dvaja pasažieri zdieľajú, letia títo pasažieri spolu (na jednu rezerváciu).
14. HitType - Atribút označujúci jednotlivé hrozby. Ak je hodnota tohto atribútu 1 tak sa jedná o normálneho pasažiera, inak nie. Jedná sa o označenie na základe porovnania s databázou už známych nebezpečných ľudí. Nie všetky záznamy obsahujú tento atribút.

2.2 Nekonzistencie

Po podrobnom preskúmaní som narazil na isté problémy v týchto dátových súboroch.

2.2.1 Formáty súborov

Prvým problémom boli rozdielne formáty súborov, v ktorých sú dáta pasažierov uložené. Jedným z nich je formát .xlsx, ktorý je štandardným formátom programu Microsoft Excel. Druhým formátom je .csv (comma-separated values). Vrámci .csv súborov však tiež dochádza k nekonzistenciám a to v spôsobe oddelenia jednotlivých záznamov. Prvým je oddelenie záznamov pomocou bodkočiarky, druhým pomocou čiarky. Taktiež v niektorých .csv súboroch, v ktorých sú záznamy oddelené pomocou bodkočiarky sa nachádzajú záznamy obsahujúce čiarky (nie ako oddeľovače, ale ako hodnoty atribútov), čo znemožňuje jednoduchú konverziu medzi týmito formátmi.

2.2.2 Dátumy

Ďalším problémom boli dátumy. Ako pre atribút ScheduledArrival, tak aj pre BirthDate sa dátumy vyskytovali v 7 rôznych formátoch.

- d.m.Y H:M
- Y-m-d H:M:S
- Y-m-d H:M
- Y-m-d
- Y/m/d H:M:S
- Y/m/d H:M
- Y/m/d

2.2.3 Atribúty HeadGUID a BodyUID

Niektoré z letov osahujú ešte pred atribútom FlightNumber atribúty HeadGUID a BodyUID.

- HeadGUID - jedná sa o alfanumerický atribút (jednou z hodnôt, ktoré nadobúda je napríklad *74cf9b88-dcc3-40f2-9960-44cc88c76a54*)
- BodyUID - jedná sa o numerický atribút.

Ani po konzultácii s poskytovateľmi dát nie je jasný význam týchto dvoch atribútov, preto ich považujeme za nekonzistenciu.

2.2.4 Identifikácia pasažiera

Pri identifikácii pasažiera dochádza k viacerým nekonzistenciám.

2.2.5 Atribút Nationality

Prvým problémom je, že pri niektorých záznamoch chýba atribút reprezentujúci národnosť pasažiera.

2.2.5.1 Atribút Names

Ďalším problémom pri jednoznačnej identifikácii pasažiera je pri jeho menách (okrem priezviska). Pri väčšine záznamov sú jednotlivé mená oddelené medzerou, čo však nie je pravdou pri všetkých záznamoch. Pri niektorých nie sú tieto mená oddelené vôbec, čo môže znemožniť identifikáciu.

2.2.5.2 Atribút Sex

Atribút sex hovorí o pohlaví pasažiera. Pri niektorých záznamoch však chýba.

2.2.5.3 Atribút DocumentType

Tento atribút má pojednávať o type dokumentu, ktorým sa pasažier preukazuje. Malo by sa jednať o občiansky preukaz, alebo pas. Avšak, nie je to tak, keďže tento atribút pri každom zázname nadobúda len jednej hodnoty a to hodnoty P (pas). Kvôli praktickým dôvodom sa teda nemôžeme spoliehať na informačnú hodnotu tohoto atribútu. Pri niektorých záznamoch sa tento atribút zase nenachádza.

2.2.5.4 Atribút DocumentIssued

Tento atribút hovorí o tom, v akom štáte je tento identifikačný dokument vydaný. Malo by sa teda jednať konkrétne o trojpísmenovú skratku tohoto štátu. Sú však aj záznamy, ktoré tento atribút nemajú, ale väčšinou majú určený typ dokumentu ako pas.

2.2.5.5 Atribút DocumentNumber

Taktiež číslo dokumentu je niekedy nekonzistentné. Malo by sa jednať o alfanumerickú hodnotu, ale pri niektorých záznamoch tento atribút nadobúda hodnôt desatinných čísel (vo formáte 1,10693E+11). Naviac, pri niektorých záznamoch tento atribút zase chýba. Toto teda tiež považujeme za nekonzistenciu v dátach.

2.2.6 Atribút Reservation

Atribút Reservation je atribút, ktorý chýba pri najväčšom množstve záznamov. Toto však nie je chybou. Jedná sa o informáciu, že daný pasažier nemal rezerváciu. Ďalším problémom s týmto atribútom je, že pri niektorých letoch namiesto toho aby pasažierom ponechali chýbajúci atribút, nadobúda tento Reservation kladných celých čísel (vždy rôzna hodnota). V niektorých prípadoch zase nadobúda hodnôt desatinných čísel (vo formáte 1,10693E+11).

2.2.7 Atribút HitType

Aj keď sa môže javiť, že sa jedná o smerodajný atribút pri identifikácii potenciálnych hrozieb, nie je to tak. Tento atribút sa nevyskytuje pri mnohých záznamoch a keď sa vyskytuje, nemôžeme sa spoliehať na jeho pravdivosť. Napríklad vieme, že ak tento atribút nadobúda hodnoty 1, malo by sa jednať o bežného a bezpečného pasažiera. Ak však nenadobúda 1, mal by nadobúdať hodnotu 2 alebo 3 (tak sú označené známe hrozby). V poskytnutých dátach sa však vyskytujú celé lety, čo majú nastavený HitType na hodnotu mimo tejto množiny známych označení.

Pre tieto dôvody neprítomnosť atribútu a nadobúdanie neznámych hodnôt považujem za nekonzistenciu v dátach.

Kapitola 3

Požadované výstupy

V tejto kapitole rozoberám ciele tejto práce.

3.1 Spracovanie dát

V predchádzajúcej kapitole som opisoval mimo iné aj nekonzistencie v dátach. Tieto nekonzistencie spôsobujú, že dáta nie sú vhodné na automatické spracovanie a tým pádom ani vhodné na strojové učenie a tiež na detekciu anomálií. Prvou úlohou je teda analyzovať spôsoby, akými je možné zbaviť sa opísaných nekonzistencií.

Ďalej je potrebné dáta dostať do najvhodnejšej formy na automatické spracovanie. Keďže pôvodne dáta boli rozdelené do štruktúry podľa dátumu príletu, už len prístup k jednotlivým letom a teda aj k pasažierom je problematický.

Ďalším problémom je formát súborov, v ktorých sa dáta nachádzajú. Je potrebné zvoliť jednotný formát.

3.2 Detekcia anomálií

Ďalším cieľom je preskúmať v týchto dátach možnosti detekcie anomálií, analyzovať vhodnosť jednotlivých techník a prípadne demonštrovať tieto techniky na dátach.

3.3 Analytické otázky

Posledným bodom je odpovedať na analytické otázky zadané políciou Českej republiky. Týmito otázkami sú:

1. Je možné na základe poskytnutých dát zachytiť atribúty nebezpečného pasažiera? (označiť všetkých nebezpečných pasažierov)
   - Ak áno, s akou presnosťou vieme určiť týchto pasažierov?
   - Dokážeme vymodelovať „bezpečného pasažiera“?
2. Dajú sa určiť na základe týchto dát celé lety (alebo letiská), ktoré majú oproti ostatným vyššiu pravdepodobnosť, že v nich budú nebezpeční pasažieri?
3. Existujú ľudia, čo stále cestujú spolu v lietadle, ale nikdy nie na jednu rezerváciu?

Na tieto otázky bude možné odpovedať po vykonaní zvyšných analýz. Odpoveď bude podložená analýzou a experimentami v jednom z data miningových nástrojov.

Kapitola 4

Analýza a návrh

V predchádzajúcej kapitole som vytýčil ciele tejto práce a v tejto kapitole analyzujem spôsoby akými tieto ciele dosiahnuť.

4.1 Spracovanie dát

Pre spracovanie dát musíme splniť viaceré požiadavky.

4.1.1 Formáty súborov

Prvým problémom, ktorý je potrebné vyriešiť, je zjednotiť formáty súborov, v ktorých sa dáta nachádzajú. Keďže je .xlsx príliš špecifickým formátom a tým pádom aj ťažko spracovateľným automaticky, je vhodné zvoliť iný formát. Na druhú stranu, .csv je štandardizovaný formát, s ktorým vieme dostatočne jednoducho pracovať (ako s textovým súborom) a preto som sa rozhodol použiť tento formát.

Avšak, keď už sme sa rozhodli použiť formát .csv, je potrebné tiež zvoliť oddeľovač. Aj keď niektoré z poskytnutých súborov obsahujú desatinné čísla, ktoré používajú ako desatinnú čiarku znak ',' namiesto bežného '.', záznamy, ktoré obsahujú desatinné číslo nie sú správne (žiadny atribút by nemal nadobúdať hodnoty desatinných čísel). S týmito nesprávnymi záznamami je potrebné sa vysporiadať skôr ako sa dáta budú automaticky spracovávať alebo zlučovať a preto nie je žiaden dôvod nepoužiť ako oddeľovač znak ','.

4.1.2 Dátumy

Dátumy, keďže sa vyskytujú v mnohých rôznych formátoch, je potrebné ich zjednotiť. RapidMiner (opísaný v 1.3.1.1) dokáže spracovať viaceré formáty dátumov:

- yyyy.MM.dd G 'at' HH:mm:ss z
- EEE, MMM d, "yy
- yyyy.MMMMM.dd GGG hh:mm aaa
- EEE, d MMM yyyy HH:mm:ss Z
- yyMMddHHmmssZ
- yyyy-MM-dd'T'HH:mm:ss.SSSZ
- yyyy-MM-dd HH:mm:ss
- M/d/yy h:mm a

Z týchto formátov je najprirodzenejší a zároveň najľahšie interpretovateľný formát Y-m-d H:M:S, ktorý som sa preto rozhodol použiť ako pre BirthDate, tak aj pre ScheduledArrival. Pri BirthDate budú hodiny, minúty a sekundy nastavené na 0 a pri spracovaní ignorované.

4.1.3 Atribúty HeadGUID a BodyUID

Keďže sa tieto atribúty vyskytujú len vo výnimočných prípadoch (samozrejme ich majú celé lety, ale len niektoré) a tiež ich význam nie je žiadnym spôsobom smerodajný, rozhodol som sa tieto atribúty vynechať.

4.1.4 Identifikácia pasažiera

Ako som opísal v predchádzajúcej kapitole, pri identifikácii pasažiera dochádza k viacerým problémom.

4.1.4.1 Atribút Nationality

Neprítomnosť atribútu pasažiera, ktorý určuje národnosť pasažiera je rozhodne zvláštnosťou. Keďže však o záznamy, ktoré tento atribút nemajú nechceme prísť, zvolíme nahradenie tohto atribútu nejakou špecifickou hodnotou. Jeho nahradenie priemerom by nebolo rozumné, kvôli tomu, že by sme stratili informáciu o tom, že tento pasažier daný atribút nemal (táto informácia by však mohla viesť k nejakej kľúčovej závislosti medzi podozrivými pasažiermi).

4.1.4.2 Atribút Names

Keďže nemáme ako určiť, kde by mala byť medzera medzi menami a spojenie všetkých mien do reťazca bez medzier by mohlo viesť k zjednoteniu rôznych pasažierov, preto necháme tento atribút v pôvodnej forme.

4.1.4.3 Atribút Sex

Keďže nechceme prísť o informáciu, že daný atribút pri zázname chýbal, volíme doplnenie špeciálnej hodnoty.

4.1.4.4 Atribút DocumentType

Prvou nekonzistenciou tohoto atribútu je jeho neprítomnosť pri niektorých záznamoch. Nechceme prísť o informáciu, že daný atribút pri zázname chýbal, takže zase volíme doplnenie špeciálnej hodnoty.

Druhou nekonzistenciou je, že pri takmer každom zázname je uvedený typ dokumentu ako pas (aj keď z formátu čísla dokumentu je jasne vidno, že sa jedná o občiansky preukaz). Týmto atribút stráca na svojej informačnej hodnote a zároveň nevieme tieto chybné hodnoty upraviť tak aby reálne odpovedali typu dokumentu, takže zvyšné hodnoty nechávame v pôvodnej podobe.

4.1.4.5 Atribút DocumentIssued

Neprítomnosť atribútu DocumentIssued je zase pozoruhodnou informáciou, takže tieto neprítomnosti nahrádzame špeciálnou hodnotou.

4.1.4.6 Atribút DocumentNumber

Informáciu o neprítomnosti tohoto atribútu nechceme stratiť, preto chýbajúce hodnoty nahrádzame špeciálnou hodnotou.

Číslo dokumentu, ktorým sa pasažier preukazuje by tiež nemalo nadobúdať desatinných hodnôt. Tieto hodnoty považujeme za nesprávne a teda ich tiež nahradíme za špeciálnu hodnotu (inú ako pre chýbajúce dáta, keďže tieto prípady chceme rozlišovať).

4.1.5 Atribút Reservation

Prvým problémom pri atribúte Reservation je neprítomnosť atribútu. Ako aj pri ostatných chýbajúcich atribútoch, aj tu použijeme nahradenie špeciálnou hodnotou, lebo informácia, že atribút chýbal je pre nás cenná.

Druhým problémom bolo, že atribút nadobúdal hodnôt kladných celých čísel aj keď sa nejednalo o rezerváciu. Tieto hodnoty však nemáme ako odlíšiť od reálnych rezervácií a preto ich ponecháme v pôvodnom stave.

Ďalším problémom bolo, že tento atribút nadobúdal hodnôt desatinných čísel. Keďže pasažieri, čo cestujú na jednu rezerváciu majú túto hodnotu rovnakú, tieto hodnoty ponecháme tiež v pôvodnom stave.

4.1.6 Atribút HitType

Sú dva typy nekonzistencií, čo sa týkajú atribútu HitType. Prvou je neprítomnosť atribútu. Keďže aj neexistencia atribútu je istou informáciou, o ktorú nechceme prísť, doplníme do záznamov, kde atribút chýba špeciálnu hodnotu.

Druhým druhom nekonzistencie je, že tento atribút nadobúda hodnôt, ktorých význam nie je známy. Kedže však HitType bezpečného pasažiera by mal nadobúdať vždy hodnotu 1, pasažieri, čo majú inú hodnotu tohoto atribútu sú istým spôsobom zaujímaví. Preto tieto hodnoty môžeme nechať v pôvodnom stave (prípadne pri spracovaní niektorým data miningovým nástrojom hodnoty označiť ako 1 - bezpeční a 0 - všetci ostatní, kde teda spadajú aj pasažieri, čo majú inú hodnotu HitType ako 1, aj tí, pri ktorých tento atribút chýbal).

4.1.7 Zjednotenie dát

Predpokladajme, že predchádzajúce kroky prebehli bez problémov a máme všetky dáta v jednotnej forme a to v súboroch .csv, kde každý záznam má presne 14 atribútov. Pre jednoduchšiu manipuláciu s týmito dátami by bolo lepšie ich mať v jednom súbore. Rozčlenenie do jednotlivých adresárov síce zlepšuje prehľadnosť pre užívateľa, ale predpokladáme, že užívateľ do týchto dát bude zasahovať (bude ich skúmať manuálne) v čo najmenšej miere.

Keďže jednoznačne identifikovať let vieme už z jednotlivých záznamov v týchto súboroch (kombinácia atribútov FlightNumber, ScheduledArrival, FlightFrom, FlightTo), nie je potrebné túto adresárovú štruktúru udžiavať. Preto som sa rozhodol spojiť všetky .csv súbory do jedného.

4.1.8 Anonymizácia dát

Pre rozdelenie práce je potrebná anonymizácia dát. Keďže chceme zachovať všetky informácie a zároveň chceme, aby daný pasažier nebol dohľadateľný na základe anonymizovaných záznamov, musíme niektoré z údajov vynechať alebo zakódovať. Na hľadanie závislostí v dátach nie je potrebné ani meno pasažiera (keďže hľadáme vyššie súvislosti a nie filtrovanie podľa mena), ani jeho presná identifikácia. Môžeme preto vynechať ako typ dokumentu, ktorým sa pasažier preukazuje (zvlášť keď je takmer pri všetkých záznamoch použitý typ dokumentu pas), tak aj číslo tohoto dokumentu. Atribútmi po anonymizácii teda budú:

1. FlightNumber
2. ScheduledArrival
3. Nationality
4. BirthDate
5. Sex
6. DocumentIssued
7. FlightFrom
8. FlightTo
9. Reservation
10. HitType

Pre určenie kontextu nechávame všetky atribúty identifikujúce let. Keďže zoznamy pasažierov nie sú verejné, nejedná sa o citlivú informáciu.

Z atribútov identifikujúcich pasažiera:

- Nationality - Národnosť nie je citlivým údajom pri pasažierovi, ale aj napriek tomu môže viesť k odhaleniu zaujímavých závislostí v dátach.
- BirthDate - Dátum narodenia, z ktorého si jednoducho vieme odvodiť vek pasažiera v období letu, čo napomôže pri sledovaní meniacich sa trendov vo vekovom zložení letov.
- Sex - Pohlavie. Taktiež sa nejedná o citlivý údaj. Môžeme však pomocou neho sledovať meniace sa trendy v zložení letov.
- DocumentIssued - Štát vydávajúci dokument, ktorým sa pasažier preukazuje. Ak sa nezhoduje s národnosťou, môže sa jednať o zaujímavý záznam. Nevedie k jednoznačnej identifikácii pasažiera.

Ani kombinácia týchto atribútov jednoznačne neidentifikuje človeka a pritom pomocou nej môžeme skúmať rôzne závislosti a trendy. Preto ju považujem za primeranú.

4.2 Detekcia anomálií

Na základe dát, ktoré máme sa jednoznačne jedná o kontextuálnu detekciu anomálií (máme časovo závislé dáta, lieta sa z rôznych letísk). Prvou úlohou je teda definovať kontext, v ktorom sa anomálie budú určovať.

4.2.1 Definícia kontextu

Definícia kontextu nie je jednoznačnou úlohou. Vyžaduje analýzu a preskúmanie rôznych možností, akými môže byť kontext definovaný. V tejto sekcii priblížim spôsoby definície od najviac všeobecného až po špecializované kontexty.

4.2.1.1 Analýza dát ako celku

Prvou možnosťou ako analyzovať tieto dáta je vnímať ich ako celok. Takto môžeme v dátach identifikovať také záznamy, ktoré sú vrámci celku anomálne. Môže sa jednať o neobvyklý vek pasažiera, neobvyklú národnosť alebo iné.

4.2.1.2 Delenie na základe informácií o pasažierovi

Kontext vieme určiť z informácií o pasažierovi.

Rozčlenenie na základe národnosti Prvou možnosťou je zoskupenie záznamov s rovnakou národnosťou. Keďže kontextov pri jednom kontexte pre každú krajinu by bolo veľké množstvo a zároveň, črty istých národností sú nadmieru podobné, môžeme zaviesť kontext ako istú skupinu národností. Kontext teda bude obsahovať záznamy, ktoré majú rovnakú národnosť a anomálie budeme skúmať vrámci neho. Takto môžeme nájsť záznamy, ktoré majú vzhľadom na danú národnosť podivuhodné miesto odletu, miesto vydania dokumentu, alebo iné.

Rozčlenenie na základe veku pasažiera Ďalšou možnosťou je rozdeliť dáta na základe veku pasažiera. Vek pasažiera získame jednoduchým odčítaním dátumu narodenia od plánovaného času príletu. Pre jednoduchosť nám bude stačiť vek v rokoch. Vek následne znormalizujeme do intervalu $[0, 1]$ pomocou min-max normalizácie. Takto definovaný kontext bude pozostávať zo záznamov, ktorých vek v čas príletu patrí do jedného intervalu. Počet intervalov, na ktoré bude vek pasažiera rozdelený bude parametrom, ktorý budeme sledovať.

4.2.1.3 Delenie na základe informácií o lete

Ako sme videli, je možné vytvoriť kontext na základe informácií o pasažierovi. Tiež je teda možné vytvoriť kontext na základe informácií o lete.

4.2.1.4 Rozčlenenie na základe miesta odletu

Jednou z možností je kontext určiť na základe miesta odletu. Jednotlivé miesta odletu môžu tvoriť kontext. Je však zase rozumné zvoliť všeobecnejší kontext.

Letiská vrámci jedného štátu, alebo aj letiská vrámci štátov, ktoré majú rovnaké črty (napríklad krajiny a letiská stredného východu) môžu tvoriť ďalšiu možnosť kontextu. Do jedného kontextu teda budú patriť také záznamy, ktoré majú rovnaké miesto odletu, alebo tieto letiská patria do jednej skupiny.

Rozčlenenie na základe čísla letu Na druhú stranu, vieme kontext zvoliť aj viac špecificky. Každé číslo letu by znamenalo nový kontext. Takto by sme vedeli skúmať anomálie v pravidelných charterových letoch na konkrétnej linke.

4.2.1.5 Rozčlenenie na základe času príletu

Taktiež vieme kontext určiť na základe času príletu. Takto zvolený kontext bude zjednocovať lety, ktoré majú plánovaný čas príletu v istom intervale. Veľkosť týchto časových intervalov je zase parametrom.

4.2.1.6 Profily pasažierov

Ďalšou možnou voľbou kontextu je vytvoriť každému pasažierovi profil, v ktorom budú zaznamenané všetky jeho lety. Vzniká tak akýsi multigraf, kde sú všetky hrany orientované smerom k letiskám v Českej republike (keďže máme dáta letov začínajúcich mimo Schengenský priestor a končiacich v Českej republike). Týmto spôsobom vieme identifikovať anomálie v letovom profile jednotlivca, skúmať nepravidelnosti v miestach odletu.

Detekcia anomálií založená na profile pasažiera prebieha spôsobom odlišným od klasických detekcií anomálií. Je pri nej vhodné sledovať spolu s miestom odletu aj čas odletu, aby sme boli schopní nové záznamy zohľadňovať viac ako záznamy, ktoré sú staršie.

Na tom, na aké české letisko let prilieta nám nezáleží, takže profil bude obsahovať vektor, kde každá zložka odpovedá jednému letisku. Keďže chceme viac zohľadňovať nové záznamy v profile, budeme uplatňovať na celý vektor faktor rozpadu (decay rate) $d \in (0, 1)$, ktorým za každý určený časový úsek vynásobíme tento vektor.

Pri novom zázname do profilu porovnáme posledný záznam (bez uplatneného faktoru rozpadu) s novým záznamom, kde najskôr uplatníme faktor rozpadu a tiež pripočítame 1 k zložke vektoru odpovedajúcej regiónu, do ktorého patrí letisko (FlightFrom) tohoto nového záznamu. Toto porovnávanie môže prebiehať rôznymi spôsobmi:

1. Rozdiel vektorov - Pri rozdiele vektorov (ani normalizovanom) však nevieme presne zachytiť anomálie ak sa jedná o letisko (alebo región), z ktorého letí pasažier prvýkrát, ale inak lieta veľa (rozdiel bude malý aj po uplatnení faktor rozpadu).

2. Porovnávanie uhlu dvoch vektorov - Takto určená vzdialenosť vektorov odhalí aj anomálie, ktoré rozdiel vektorov nie je schopný zachytiť. Ak sa jedná o letisko, z ktorého pasažier bežne nelieta, bude uhol vektorov veľký. Taktiež, ak pasažier dlhšiu dobu nelieta z letiska, z ktorého lietal bežne a často, faktor rozpadu spôsobí, že vysoké hodnoty klesajú rýchlejšie a tým pádom bude uhol medzi vektormi vyšší.

Porovnávaním uhlov vieme teda detekovať rôzne typy anomálií a teda ho zvolíme pre porovnanie vektorov.

4.2.2 Voľba techniky detekcie anomálií

Predpokladajme, že máme zvolený kontext (iný ako profil pasažiera). Musíme teda zvoliť techniku vhodnú na identifikáciu anomálií.

Prvou otázkou je, či budeme záznamy, ktorých atribút HitType nadobúda hodnotu inú ako 1 brať ako anomálne:

- HitType určuje anomálnosť Ak HitType určuje anomálnosť, môžeme využívať supervised techniky detekcie anomálií. Pre zjednodušenie rozdelíme záznamy na normálne - tie, ktoré majú HitType 1 a tie, ktoré majú inú hodnotu HitType - anomálne záznamy. Problém detekcie anomálií sme takto zredukovali na problém binárnej klasifikácie záznamov. Aplikovať teda budeme techniky opísané v 1.9. Problémom tohoto prístupu je chýbajúci atribút HitType. O záznamoch, ktoré tento atribút nemajú nevieme jednoznačne povedať, či sú anomálne alebo nie a ich zaradenie do jednej alebo druhej kategórie by mohlo spôsobiť chybnú klasifikáciu a teda aj nepresnú detekciu anomálií. Pri učení teda tieto záznamy nemôžeme použiť.

- HitType neurčuje anomálnosť V prípade, že HitType neberieme ako označenie anomálnosti záznamu, prichádzajú v úvahu unsupervised techniky detekcie anomálií ako techniky najbližšieho suseda 1.10, alebo techniky založené na zhlukovaní 1.11. Tieto metódy sú založené na istej miere vzdialenosti medzi jednotlivými záznamami.

Keďže nepoznáme štatistické rozdelenie dát a ani nemôžeme žiadne takéto rozdelenie predpokladať, štatistické techniky detekcie anomálií sú nepoužiteľné.

4.2.3 Definovanie vzdialenosti medzi dvoma záznamami

Definovanie vzdialenosti medzi dvoma záznamami má kľúčový význam pre unsupervised techniky detekcie anomálií, keďže sa pri týchto metódach skúmajú vzdialenosti medzi susedmi a tiež hustota.

Túto vzdialenosť som sa rozhodol definovať po zložkách, keďže nemáme dostatočné znalosti o podobnosti dvoch pasažierov.

1. FlightNumber - Ak je číslo letu rovnaké, je vzdialenosť medzi hodnotami 0, inak 1
2. ScheduledArrival - Dátum a čas príletu ponúka mieru vzdialenosti už triviálne.
3. Nationality - Národnosť pasažiera má viac možností ako bude určovať vzdialenosť medzi dvoma hodnotami. Prvou je určovať vzdialenosť geograficky. Druhou voľbou by mohlo byť preskúmať politické a kultúrne podobnosti jednotlivých štátov a regiónov v spolupráci s Políciou ČR (čo by mohlo byť náplňou ďalších prác) a na základe týchto informácií zostaviť maticu podobností.
4. Surname - Priezvisko pasažiera pre určenie vzdialenosti medzi dvoma záznamami ignorujeme.
5. Names - Všetky zvyšné mená pasažiera pre určenie vzdialenosti medzi dvoma záznamami tiež ignorujeme.
6. BirthDate - Dátum narodenia pasažiera vrámci detekcie anomálií konvertujeme na vek. Vek je numerický atribút a pri numerických atribútoch je vzdialenosť vypočítaná triviálne.
7. Sex - Ak je pohlavie dvoch záznamov rovnaké, je vzdialenosť 0, ak rôzne tak 1.
8. DocumentType - Typ dokumentu, ktorým sa preukazuje pasažier pre určenie vzdialenosti vynechávame, keďže nadobúda takmer vždy hodnoty P - pas.
9. DocumentIssued - Štát vydávajúci tento dokument. Vzdialenosť určíme rovnako ako pri národnosti.
10. DocumentNumber - Číslo tohoto dokumentu pre vzdialenosť ignorujeme.
11. FlightFrom - Kód letiska, z ktorého let odlieta. Prvou možnosťou, ako určiť vzdialenosť medzi dvoma letiskami je zvoliť geografickú vzdialenosť (podobne ako pri národnosti). Ďalšou voľbou by mohlo byť preskúmať politické a kultúrne podobnosti jednotlivých štátov a regiónov v spolupráci s Políciou ČR (čo by mohlo byť náplňou ďalších prác) a na základe týchto informácií zostaviť maticu podobností letísk.

12. FlightTo - Kód letiska, na ktoré let prilieta. Vzdialenosť určujeme rovnako ako pri FlightFrom.

13. Reservation - Ak majú 2 záznamy rovnaký FlightNumber, ScheduledArrival a Reservation, ich vzdialenosť v atribúte Reservation je 0, inak 1.

14. HitType - Atribút, či sa jedná o hrozbu pri určovaní vzdialenosti budeme ignorovať, keďže sa jedná o atribút, ktorý nadobúda hodnôt na základe porovnania proti databáze a teda pre vzdialenosť nemá význam.

Takto definované vzdialenosti medzi jednotlivými atribútmi je potrebné normalizovať, keďže vzdialenosť medzi dvoma záznamami v atribúte FlightNumber môže byť maximálne 1, pričom vo veku môže byť aj ďaleko vyššia. Využijeme min-max normalizáciu aby sme zredukovali vplyv atribútov, ktorých rozptyl je väčší ako 1 na celkovú vzdialenosť medzi dvoma záznamami.

4.3 Analytické otázky

Ďalším požadovaným výstupom sú odpovede na dohodnuté analytické otázky.

4.3.1 Klasifikácia nebezpečných pasažierov

Prvou otázkou je: „Je možné na základe poskytnutých dát zachytiť atribúty nebezpečného pasažiera? (označiť všetkých nebezpečných pasažierov)“

Táto otázka je obdobnou otázke, či dokážeme detekovať anomálie supervised technikou - trénovaním modelu a následnou klasifikaciou jednotlivých pasažierov. Na natrénovanie aj testovanie modelu však nebudeme používať všetky atribúty. Sústredíme sa na všeobecné črty ľudí, takže budeme brať v úvahu národnosť, vek, pohlavie, typ dokumentu, ktorým sa preukazuje, štát, kde bol tento dokument vydaný, či má rezerváciu a pre označenie, či je bezpečný využijeme HitType. Pre porovnanie vyberieme rôzne modely a zhodnotíme ich úspešnosť.

4.3.1.1 S akou presnosťou vieme určiť týchto pasažierov?

Po natrénovaní modelu pomocou dát pasažierov budeme skúmať úspešnosť modelu pri klasifikácii. Ako som spomenul, budeme trénovať viaceré modely a následne porovnávať ich úspešnosť. Pre spoľahlivé výsledky použijeme krížovú validáciu.

4.3.1.2 Dokážeme vymodelovať „bezpečného pasažiera“?

Vymodelovanie bezpečného pasažiera zase súvisí s trénovaním modelu. Ak je model schopný sa natrénovať z poskytnutých dát a tiež bude mať pri nebezpečných pasažieroch vysokú hodnotu „recall“ (podiel počtu správne klasifi-

kovaných nebezpečných pasažierov a počtu všetkých nebezpečných pasažierov), čo zaručí, že ak už tento model klasifikuje pasažiera ako bezpečného, môžeme s vysokou pravdepodobnosťou povedať, že tento pasažier je naozaj bezpečný. Takto vymodelujeme bezpečného pasažiera. Pre názorné (ľudsky pochopiteľné) podmienky na to, aby bol pasažier bezpečný môžeme sledovať podmienky rozodovacieho stromu, alebo skúsiť vytvoriť asociačné pravidlá a sledovať tie, ktoré vedú k HitType 1.

4.3.2 Nebezpečné lety a letiská

Redukovaním záznamov na atribúty

1. FlightFrom
2. HitType

získavame základ pre identifikáciu nebezpečných letísk. HitType v takomto prípade berieme ako označenie a teda ho využívame na skúmanie, aké veľké percento pasažierov z daných letísk má túto hodnotu vyššiu ako 1 (záznamy, kde HitType chýbal vynechávame).

Druhou úlohou je analyzovať lety. Podobným spôsobom ako pre letiská zredukujeme atribúty

1. FlightNumber
2. HitType

a následne sledujeme, ktoré lety majú najvyššiu pravdepodobnosť, že obsahujú nebezpečného pasažiera.

4.3.3 Neznámi spolucestujúci

Poslednou analytickou otázkou bolo, či existujú pasažieri, čo stále cestujú spolu, ale nikdy nie na jednu rezerváciu.

Jedná sa o jednoduché vyhľadávanie v dátach.

Kapitola 5

Realizácia

Podľa vykonanej analýzy je potrebné uskutočniť aj realizáciu.

5.1 Spracovanie dát

Prvým bodom je spracovanie dát. Keďže sa v dátach nachádzajú rôzne nekonzistencie a iné problémy, je potrebné sa ich zbaviť aby sme dáta pripravili do formy vhodnej na ďalšie (automatické) spracovanie, detekciu anomálií a iné.

5.1.1 Formáty súborov

Ako som opísal v sekcii 4.1.1, musíme dátové súbory dostať do jednotného formátu.

Keďže niektoré csv. súbory obsahujú desatinné čísla, kde je desatinná čiarka znak ',', musíme tieto čiarky nahradiť znakom '.'. Tento typ nekonzistencie sa vyskytuje v .csv súboroch, kde je ako oddeľovač použitý znak ';', a preto túto úpravu uskutočníme jednoducho, príkazom *sed s/,/./g*.

Ďalším problémom bolo .csv súbory využívajúce ako oddeľovač ';' prekonvertovať na súbory využívajúce ako oddeľovač ','. Keďže sme si už teraz istí, že sa v týchto súboroch nenachádza žiadny znak ',' (príkaz *sed s/,/./g* nahradil všetky znakom '.'), vieme že nahradením znaku ';' znakom ',' neporušíme formát súborov. Nahradenie uskutočňujeme príkazom *sed s/;/,/g*. Teraz teda všetky .csv súbory používajú oddeľovač ','.

Poslednou potrebnou úpravou pre to, aby sme súbory dostali do jednotnej podoby čo sa týka formátu je konverzia .xlsx súborov na .csv súbory. Na túto úlohu som zvolil nástroj *ssconvert*1.2.1, ktorý už v pôvodnom nastavení konvertuje do formátu .csv s použitým oddeľovačom ','.

5.1.2 Dátumy

Podľa analýzy sme zvolili nový formát dátumov Y-m-d H:M:S. Na túto úlohu je ideálny modul *datetime* v jazyku python s jeho funkciami *strptime(format)* pre načítanie dátumu z textového reťazca v zadanom formáte a *strftime(format)* pre konverziu dátumu do formy textového reťazca v zadanom formáte. Takto vieme pomocou blokov *try-except* v pythone identifikovať jednotlivé formáty a následne ich konvertovať do jednotného.

5.1.3 Atribúty HeadGUID a BodyUID

Identifikovať, či súbor obsahuje atribúty HeadGUID a BodyUID je priamočiare a teda ich vymazanie spočíva v ignorovaní prvých dvoch stĺpcov (kedže sa nachádzajú vždy na rovnakom mieste). Realizácia prebiehala pythonovským skriptom za pomoci modulu *csv*, ktorý ponúka možnosti elegantnej práce s .csv súbormi.

5.1.4 Atribút Nationality

Chýbajúce hodnoty atribútu Nationality doplníme na hodnotu -1 pomocou pythonovského skriptu. Tejto hodnoty nenadobúda atribút za žiadnych iných okolností, preto je to dobre zvolená špeciálna hodnota.

5.1.5 Atribút Sex

Chýbajúce hodnoty atribútu Sex doplníme na hodnotu -1 pomocou pythonovského skriptu. Tejto hodnoty nenadobúda atribút za žiadnych iných okolností, preto je to dobre zvolená špeciálna hodnota.

5.1.5.1 Atribút DocumentType

Chýbajúce hodnoty atribútu DocumentType doplníme na hodnotu -1 pomocou pythonovského skriptu. Tejto hodnoty nenadobúda atribút za žiadnych iných okolností, preto je to dobre zvolená špeciálna hodnota.

5.1.5.2 Atribút DocumentIssued

Chýbajúce hodnoty atribútu DocumentIssued doplníme na hodnotu -1 pomocou pythonovského skriptu. Tejto hodnoty nenadobúda atribút za žiadnych iných okolností, preto je to dobre zvolená špeciálna hodnota.

5.1.5.3 Atribút DocumentNumber

Chýbajúce hodnoty atribútu DocumentNumber doplníme na hodnotu -1 pomocou pythonovského skriptu. Tejto hodnoty nenadobúda atribút za žiadnych iných okolností, preto je to dobre zvolená špeciálna hodnota.

Pri záznamoch, kde tento atribút nadobúda hodnôt desatinných čísel túto hodnotu nahradíme hodnotou -2. Tejto hodnoty taktiež atribút nenadobúda v žiadnom inom prípade.

5.1.6 Atribút Reservation

Chýbajúce hodnoty atribútu Reservation doplníme na hodnotu -1 pomocou pythonovského skriptu. Tejto hodnoty nenadobúda atribút za žiadnych iných okolností, preto je to dobre zvolená špeciálna hodnota.

5.1.7 Atribút HitType

Chýbajúce hodnoty atribútu HitType doplníme na hodnotu -1 pomocou pythonovského skriptu. Tejto hodnoty nenadobúda atribút za žiadnych iných okolností, preto je to dobre zvolená špeciálna hodnota.

5.1.8 Zjednotenie dát

Zjednotenie dát prebieha jednoduchým zreťazením jednotlivých súborov (upravených podľa predchádzajúcich bodov), pričom odmazávame hlavičku s menami atribútov. Dostávame tak jeden súbor s atribútmi:

1. FlightNumber
2. ScheduledArrival
3. Nationality
4. Surname
5. Names
6. BirthDate
7. Sex
8. DocumentType
9. DocumentIssued
10. DocumentNumber
11. FlightFrom
12. FlightTo
13. Reservation
14. HitType

Kde nemáme chýbajúce atribúty a ani iné nekonzistencie.

5.2 Anonymizácia dát

Anonymizácia dát prebieha jednoduchým vynechaním stĺpcov pomocou pythonovského skriptu.

5.3 Detekcia anomálií

Pri realizácii detekcie anomálií sa sústredíme na vytváranie profilov pre jednotlivých pasažierov.

Ako bolo spomenuté v analytickej sekcii, každého pasažiera budeme jednoznačne identifikovať národnosťou, menami a dátumom narodenia (číslo dokumentu nie je do tejto identifikácie zahrnuté, lebo človek môže dokument zmeniť, čo by spôsobilo, že by sa vytvoril nový profil). Keďže budeme aplikovať faktor rozpadu, pre jednoduchosť je dobré zoradiť si všetky záznamy podľa času príletu. Pri takto zoradenom data sete stačí ukladať dátum a čas posledného letu, podľa ktorého budeme pri novom zázname do profilu aplikovať faktor rozpadu na vektor reprezentujúci počet príletov z letísk.

Zjednotené dáta, ktoré sú zoradené podľa dátumu a času príletu (od najnižšieho po najvyšší) prechádzame záznam po zázname a pri každom zázname môžu nastať tri prípady:

1. Daný pasažier sa ešte v zozname nevyskytol - Pridávame do zoznamu pasažierov nový záznam, kde identifikátor je zreťazenie národnosti, mien a dátumu narodenia a tomuto pasažierovi pridávame prvé letisko (jeho trojpísmenové označenie a odpovedajúcu hodnotu tomuto letisku na 1) a uložíme si dátum a čas tohoto letu.

2. Daný pasažier sa vyskytol a teda uplatňujeme faktor rozpadu. Na základe dátumu a času posledného letu vyrátame počet mesiacov kedy tento pasažier letel naposledy a všetky zložky vektora v profile vynásobíme d^m, kde d je koeficient rozpadu a m počet mesiacov.

    - ale dané letisko ešte v jeho profile nemáme. Pasažierovi teda po uplatnení koeficientu rozpadu vytvoríme nový záznam pre toto letisko a nastavíme hodnotu na 1.
    - a záznam pre dané letisko už existuje. Takto len hodnotu zvýšime o 1.

Následne nastavíme záznam dátum a čas posledného letu tohoto profilu na dátum a čas príletu nového záznamu.

Pre detekciu anomálií budeme sledovať rozdiel hodnôt vektoru pred uplatnením faktoru rozpadu a pridaním nového letu a vektoru po uskutočnení zmien. Tento rozdiel budeme sledovať ako uhol, ktorý zvierajú tieto vektory. Pre označenie anomálnych záznamov môžeme nastaviť istú anomálnu hranicu a ak bude uhol väčší ako daná hranica, označíme ako anomálne.

5.4 Analytické otázky

Realizácia analytických otázok spočíva v definovaní spôsobu, akým na dané otázky odpovieme.

5.4.1 Klasifikácia nebezpečných pasažierov

Prvou otázkou bolo: „Je možné na základe poskytnutých dát zachytiť atribúty nebezpečného pasažiera? (označiť všetkých nebezpečných pasažierov)“. Odpovedať na otázku znamená vytvoriť prediktívny model, ktorý bude presne zachytávať pasažierov s HitType iným ako 1 (tie, ktorým tento atribút chýba nemôžeme do tejto klasifikácie zahrnúť, keďže nevieme do ktorej triedy patria). Model natrénujeme z data setu, kde budeme brať v úvahu len všeobecné atribúty pasažiera - národnosť, vek (získaný odčítaním dátumu narodenia od dátumu príletu) v mesiacoch, pohlavie, miesto vydania dokumentu, ktorým sa pasažier preukazuje, atribútom hovoriacim, či cestoval na rezerváciu alebo nie a v poslednom rade HitType aby sme boli schopní model natrénovať a následne zhodnotiť (supervised technika). Ako som už spomenul, potrebujeme model, ktorý bude mať vysokú hodnotu štatistiky recall (aký veľký pomer z celkového počtu nebezpečných pasažierov náš model reálne označil ako nebezpečných) pri triede odpovedajúcej nebezpečným pasažierom.

Prvou podotázkou je, s akou presnosťou vieme určiť týchto pasažierov. Na zhodnotenie presnosti modelu využijeme krížovú validáciu so stratifikovaným výberom vzoriek.

Druhou podotázkou je či dokážeme vymodelovať „bezpečného pasažiera“. To znamená, že sa pýtame, či náš model má vysokú hodnotu precision pri bezpečných pasažieroch (pomer koľko z tých, ktorých sme klasifikovali ako bezpečných pasažierov sú reálne bezpeční pasažieri).

Pre porovnanie, skúsim natrénovať rôzne prediktívne modely a následne budem porovnávať ako veľmi vyhovuje predpokladom, ktoré na daný klasifikátor kladieme. Budeme využívať desaťnásobnú krížovú validáciu a medzi porovnávané klasifikátory zahŕňam:

1. Bayesovský klasifikátor
2. Rozhodovacie stromy - tento model vyžaduje isté parametre pri trénovaní
    - Criterion - kritérium, podľa ktorého určujeme, či daný list zostane listom, alebo sa ďalej bude deliť. Jeho hodnotu nastavujeme na Gain ratio - vypočíta sa entropia každého atribútu. Takto vypočítaný informačný zisk je následne upravený vzhľadom na rozsah a rovnomernosť hodnôt atribútu a na delenie je zvolený ten s najnižším ziskom.
    - Maximal depth - maximálna hĺbka stromu. Nastavujeme na 200.

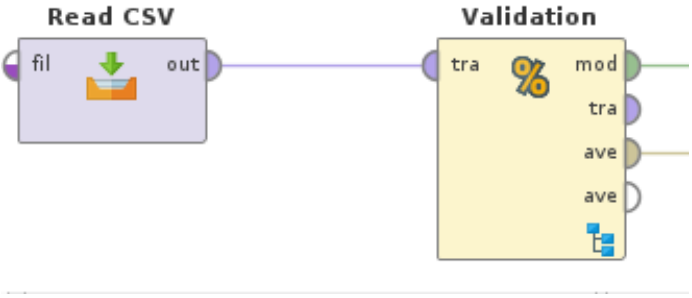
- Confidence - parameter, určujúci konfidenčnú úroveň pri pesimistickom odhade chyby prerezávania. Nastavujeme na 0.1
- Minimal gain - parameter určujúci hranicu, ktorú keď dosahuje hodnota kritéria na nahradenie listu uzlom (ďalšou podmienkou), toto nahradenie sa vykoná. Nastavujeme ho na 0.005.

Tieto parametre som zvolil za účelom maximalizácie spomínaných štatistických hodnôt.

Tieto klasifikátory sú schopné brať v úvahu atribúty Nationality, Sex, DocumentType a DocumentIssued (tzv. polynomiálne atribúty) bez zmeny a tým pádom vieme vytvoriť spoľahlivé modely bez pretvárania týchto atribútov. Na druhú stranu neurónové siete a SVM si s týmito atribútmi nie sú schopné poradiť keďže vyžadujú numerický vstup. Jednou z možností je zakódovať tieto atribúty unikátnymi celými číslami (napríklad CZE - 2, USA - 1 atp.). Tento spôsob však spôsobí, že vzdialenosť medzi dvoma národnosťami môže byť nízka napriek tomu, že si tieto štáty nie sú podobné v žiadnom ohľade. Tým pádom bude klasifikácia prebiehať nesprávne. Druhým spôsobom, ktorý by mohol vyriešiť tento problém je pre každú hodnotu atribútu vytvoriť nový atribút, ktorý by nadobúdal 1 ak pôvodný atribút nadobúda hodnotu danú hodnotu ([Nationality=CZE] 0). Keďže je 195 krajín a teda aj národností a štátov, kde by mohol byť vydaný dokument, znamenalo by to pre náš data set ~400 nových atribútov, čo pri ~2 000 000 pasažieroch spôsobí značné zväčšenie, s ktorým si RapidMiner nevie poradiť a tiež moje technické vybavenie neumožňuje takýto veľký data set spracovať.

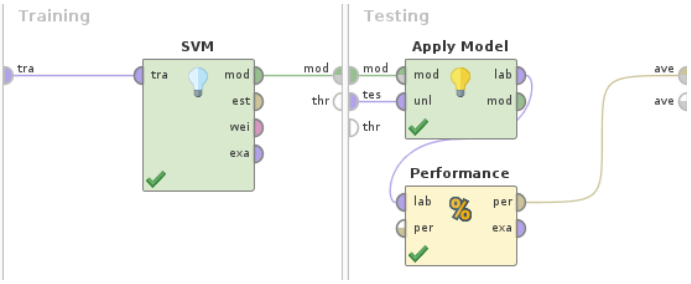
V RapidMineri sa nachádzajú operátory pre všetky tieto modely a tým pádom ich využijem. Pre názornú ukážku tiež preskúmam neurónové siete a SVM aby som ukázal, že ich výsledky nie sú vyhovujúce.

Pre Bayesovský klasifikátor a rozhodovacie stromy bude schéma vyzerať nasledovne:



Obr. 5.1: Zapojenie procesu

Schéma obsahuje dva operátory: jeden na načítanie data setu obsahujúceho pasažierov a druhý na krížovú validáciu.

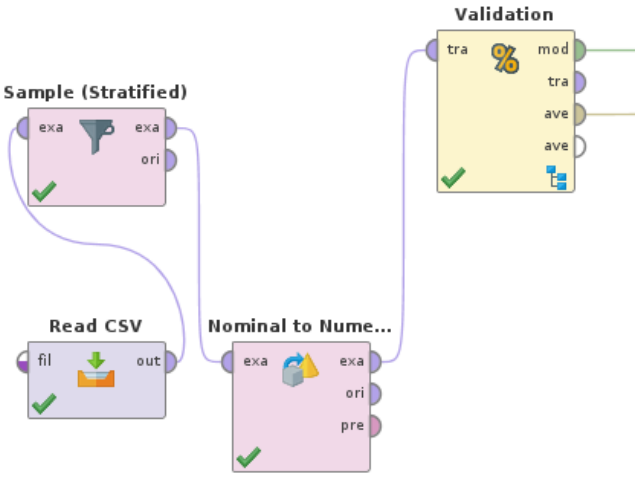


Obr. 5.2: Zapojenie vnútri operátoru validácie

Vnútro operátoru validácie máme rozdelené na dve časti a to:

1. Trénovacia časť - tu sa nachádza model, ktorý trénujeme (na obrázku SVM, ale pre jednotlivé modely mám v schéme odpovedajúce operátory)
2. Testovacia časť - v tejto časti natrénovaný model aplikujeme (vykonávame testovanie) a následne meriame spomínané miery presnosti.

Ako som spomenul, pre neurónové siete musíme niektoré atribúty prekonvertovať. Nasledujúca schéma ukazuje zapojenie procesu pri aplikovaní takejto konverzie. Používame unikátnu celočíselnú hodnotu pre každú hodnotu polynomiálneho atribútu. Ako si môžeme všimnúť, aplikujem ešte jeden operátor - Sample (Stratified), ktorý slúži na vybratie stratifikovaného vzorku z nášho data setu keďže trénovanie a testovanie na takom obsiahlom dátovom súbore trvalo príliš dlho (obzvlášť keď pri týchto modeloch len ukazujeme, že poskytujú nevyhovujúce výsledky).



Obr. 5.3: Zapojenie procesu s konverziou polynomiálnych atribútov

5.4.2 Nebezpečné lety a letiská

Obe skúmania prebiehajú pythonovským skriptom po vybraní atribútov spomenutých v 4.3.2, počítaním záznamov odpovedajúcim jednotlivým letiskám alebo letom, ktoré majú HitType rôzny od -1 (ktorým tento atribút nechýbal) a záznamov, ktoré majú HitType vyšší ako 1 (nebezpeční pasažieri).

Tieto hodnoty následne dáme do pomeru a sledujeme, ktoré lety alebo letiská majú najvyššiu pravdepodobnosť, že pasažier, ktorý týmito letmi alebo z týchto letísk letí je nebezpečný.

5.4.3 Neznámi spolucestujúci

Neznámych spolucestujúcich (pasažierov, ktorí spolu cestujú, ale nikdy nie na jednu rezerváciu) vieme identifikovať pythonovským skriptom. Najskôr si vytvoríme databázu, kde kľúčom je jednoznačná identifikácia pasažiera a hodnotou je množina jednoznačných identifikátorov letov. Každý z týchto záznamov záznamov porovnám proti ostatným záznamom v databázi a urobíme prienik jednotlivých letov, ktoré absolvovali. Tento prienik však nerobíme ak aspoň jeden z pasažierov absolvoval len jeden let (neodhalíme tým nič zaujímavé) alebo letia na jednu rezerváciu (rodina, kolegovia a podobne - ľudia, čo nemajú problém s tým, že je o nich známe, že cestujú spolu). Ak je tento prienik len jeden, nejedná sa o žiadnu anomáliu a teda ho ignorujeme.

Kapitola 6

Výsledky

V tejto sekcii hodnotím výsledky výstupov realizačnej časti.

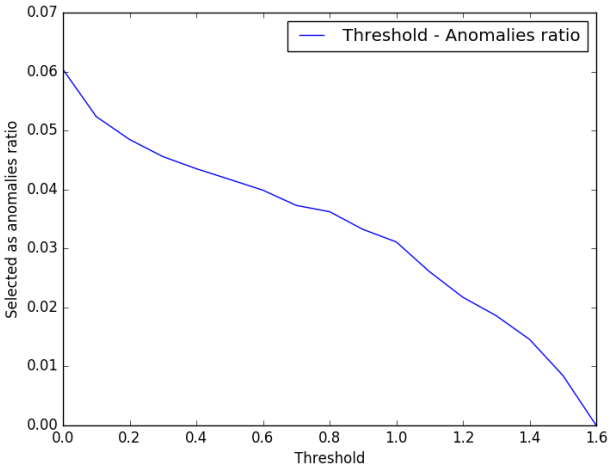
6.1 Detekcia anomálií

Výstupom detekcie anomálií popísanej v realizačnej časti je anomálne skóre zadaného nového letu daného pasažiera. Máme dva parametre, ktoré tu môžeme ladiť a to faktor rozpadu a hranica, ktorú keď prekročí anomálne skóre pre nový záznam budeme ho považovať za anomálny. Takto vieme upravovať citlivosť detekcie anomálií.

Avšak pre overenie úplnej správnosti tejto detekcie anomálií a nezostáva nič iné, len dané parametre nastavovať manuálne a skúmať, či výstup je zmysluplný. Pri nastavení faktoru rozpadu na 0.8 (ak človek lieta približne raz za rok z jedného letiska), tak hodnota zostáva približne rovnaká (odchýlka 0.6) a teda uhol medzi pôvodným vektorom a novým vektorom bude veľmi malý.

Naopak, ak daný pasažier neletel z daného letiska ešte nikdy, získava vektor nový rozmer a tým pádom je uhol medzi novým a pôvodným veľký a teda získavame anomáliu. Taktiež ak pasažier dlho nelieta z letiska, z ktorého zvykol lietavať bežne (vysoká hodnota vo vektore odpovedajúca danému letisku), faktor rozpadu zmenší túto hodnotu viac v porovnaní s ostatnými hodnotami, ktoré odpovedajú menej frekventovaným letiskám.

Na grafe 6.1 vidíme, ako klesá podiel detekovaných anomálií na základe voľby hranice anomálnosti. Samozrejme môžeme označovať za anomálnych pasažierov, ktorí letia prvýkrát (ktorých je drvivá väčšina), ale na tomto grafe som chcel ukázať, ako veľmi je podiel vybraných pasažierov ovplyvnený týmto parametrom.



Obr. 6.1: Ovplyvnenie podielu označených zvolenou hranicou anomálnosti

6.2 Analytické otázky

Odpovedať na analytické otázky znamená aj podložiť svoje odpovede faktami a číslami získanými z poskytnutých dát.

6.2.1 Klasifikácia nebezpečných pasažierov

Podľa realizácie som vykonal merania a táto sekcia zhŕňa ich výsledky.

6.2.1.1 Bayesovský klasifikátor

Prvým skúmaným modelom bol Bayesovský klasifikátor.

Tabuľka 6.1: Výsledky Bayesovského klasifikátoru

	true 1	true 0	class precision
pred. 1	841220	35170	95.99%
pred. 0	560918	489388	46.59%
class recall	60.00%	93.30%	

Ako vidíme, tento model je nadmieru úspešný vzhľadom na naše požiadavky. Pre schopnosť zachytiť nebezpečných pasažierov sme vyžadovali vysoký recall pri nebezpečných pasažieroch (trieda 0), čo tento model spĺňa - 93.30%.

Presnosťou, s ktorou ich vieme určiť budeme rozumieť hodnotu precision pri triede 0. Táto presnosť je teda 46.59%, čo síce nie je vysoká hodnota, ale

pri označovaní nebezpečných pasažierov je dôležitá práve hodnota recall, ktorá hovorí koľko z nebezpečných sme naozaj zachytili.

Či dokážeme vymodelovať „bezpečného pasažiera“ zase ukazuje hodnota precision pri triede 1. Táto hodnota je 95.99%, čo značí, že ich dokážeme vymodelovať (ak už niekoho označíme za bezpečného je 95.99% pravdepodobnosť, že naozaj je bezpečný).

6.2.1.2 Rozhodovacie stromy

Ako vidíme, výsledky rozhodovacieho stromu nie sú vyhovujúce, preto by som tento model nezahŕňal v ďalšej práci.

Tabuľka 6.2: Výsledky rozhodovacieho stromu

	true 1	true 0	class precision
pred. 1	1402054	524454	72.78%
pred. 0	84	104	55.32%
class recall	99.99%	0.02%	

Tieto výsledky som sa snažil optimalizovať voľbou iných parametrov modelu, ale žiadne rozhodovacie stromy neboli schopné dosiahnuť vysokých hodnôt tých štatistík, ktoré potrebujeme. Rozhodovacie stromy teda hodnotím ako nevhodné.

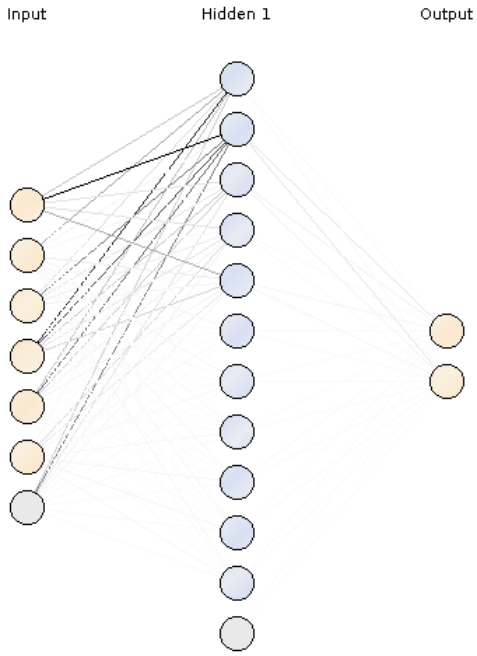
6.2.1.3 Neurónové siete

Ako som spomínal, neurónovú sieť zahŕňam len kvôli ukážke jej nesprávneho klasifikovania. V RapidMineri som využil operátor AutoMLP, ktorý neupravuje len váhy vrámci neurónovej siete, ale ju tiež počas učenia mutuje a kríži (pridávajú, respektíve odoberajú sa vrstvy a neuróny). Výstupom tohoto operátora bola neurónová sieť zobrazená na obrázku 6.2.1.3. Takto vyzerajúca neurónová sieť po uplatnení krížovej validácie dosahuje nasledovných výsledkov:

Tabuľka 6.3: Výsledky neurónových sietí

	true 1	true 0	class precision
pred. 1	13841	4999	73.47%
pred. 0	180	247	57.85%
class recall	98.72% 4.71	%	

Ako vidíme, výsledky neurónovej siete sú podľa očakávaní. Mapovanie atribútov na unikátne celé čísla zdeformovalo dáta a preto sme neboli schopní zachytiť takmer žiadne závislosti.



Obr. 6.2: Výsledná neurónová sieť

6.2.1.4 SVM

Tabuľka 6.4: Výsledky SVM

	true 1	true 0	class precision
pred. 1	13462	4962	73.07%
pred. 0	559	284	33.69%
class recall	96.01%	5.41%	

Podobne ako neurónová sieť, aj SVM neposkytuje dobré výsledky. Dôvod je taktiež rovnaký a to mapovanie atribútov na unikátne celé čísla.

6.2.2 Nebezpečné lety a letiská

Podľa postupu popísaného v realizácii som uskutočnil merania.

Stĺpec Airport odpovedá trojpísmenovej skratke jednotlivých letísk. Stĺpec Passengers odpovedá počtu pasažierov, ktorí odlietali z daného letiska do Českej republiky a zároveň mali určený HitType. Ratio odpovedá podielu nebezpečných pasažierov spomedzi všetkých pasažierov z daného letiska.

Tabuľka 6.5: Nebezpečné letiská

Airport	Passengers	Ratio	Airport	Passengers	Ratio
BUD	126	0.00	DME	37334	0.25
CMB	527	0.00	DWC	12	0.25
FUE	150	0.00	GYD	13029	0.25
LPA	317	0.00	SJJ	185	0.26
OTP	146	0.00	CEK	1649	0.28
TBJ	62	0.00	BEG	10295	0.30
ZAG	243	0.00	DOK	1283	0.31
DJE	2127	0.01	IST	145298	0.31
ICN	110954	0.01	GOJ	17296	0.32
OSL	155	0.01	KBP	95537	0.32
SID	3865	0.01	TBS	8875	0.32
ACE	3237	0.02	SVO	545198	0.34
AYT	175645	0.02	SAW	6125	0.35
DKR	540	0.02	MSQ	36694	0.36
GZP	101	0.02	UFA	16489	0.36
NBE	23624	0.02	ODS	7863	0.38
TIA	473	0.02	SVX	56625	0.38
MLA	2072	0.03	PEE	14409	0.41
TLV	70433	0.03	KUF	43594	0.44
AUH	14138	0.06	EVN	30747	0.45
TJM	471	0.11	ROV	32069	0.46
DXB	185838	0.12	ALA	15442	0.53
TAS	8161	0.15	VOZ	161	0.54
KZN	3331	0.18	LED	134020	0.64
KRR	12647	0.19	ADB	117	0.66
OVB	18679	0.21	YUL	2818	0.71
VKO	15470	0.24			

Ako vidíme isté letiská naozaj majú vyššiu pravdepodobnosť, že pasažier, ktorý z nich odlieta má HitType vyšší ako 1 a teda by bol zaradený ako nebezpečný pasažier (napríklad YUL - Montréal–Pierre Elliott Trudeau International Airport alebo LED - Pulkovo Airport). Tieto údaje nemusia viesť k presnému označeniu letísk, ktoré sú nebezpečné keďže vynechávame záznamy, ktoré HitType určený nemali.

Napriek tomu pri niektorých letiskách vidíme, že sú dostatočne frekventované, čo sa týka pasažierov letiacich do Českej republiky a tiež podiel nebezpečných pasažierov je vysoký.

Na druhú stranu rôznych letov je veľké množstvo. Spomedzi všetkých letov som vybral len tie s najvyšším podielom nebezpečných pasažierov:

Tabuľka 6.6: Nebezpečné lety

Flight	Passengers	Ratio	Flight	Passengers	Ratio
OK1893	589	0.31	OK913	21044	0.48
OK2897	175	0.31	OK947	6386	0.51
OK895	46283	0.31	OK181	15442	0.53
OK905	61575	0.31	OK863	2208	0.53
OK917	46641	0.31	7R5509	161	0.54
OK935	8875	0.32	8Q6017	247	0.57
OK893	66179	0.33	OK899	40293	0.59
U6701	20860	0.33	FV6221	19598	0.61
B2861	30107	0.34	OK887	46773	0.61
OK921	1147	0.34	FV221	36896	0.62
PC453	6019	0.34	OK889	7724	0.62
PS807	48763	0.34	QS2653	117	0.66
SU2012	69023	0.34	OK891	10979	0.67
OK255	11056	0.35	TS700	2563	0.69
OK865	4334	0.35	UN9797	53	0.70
OK251	16489	0.36	FV223	1346	0.77
SU2014	47416	0.36	TK3306	106	0.83
JU610	5192	0.37	UN461	10124	0.83
OK907	1441	0.37	FV6715	513	0.87
TK1767	75723	0.37	JA1002	54	0.89
OK923	7863	0.38	CAI755	150	0.91
OK257	13849	0.41	8Q781	94	0.95
SU2010	65089	0.42	6W2951	38	0.97
OK911	40585	0.45	FV5929	67	0.99
OK931	30747	0.45	TS798	70	0.99
OK915	32069	0.46	TS690	159	1.00
UN8361	324	0.46	U6500	3	1.00

Stĺpce sú označené podobne ako pri letiskách, Flight označuje číslo letu, ktorého podiel nebezpečných pasažierov skúmame. Zase musím podotknúť, že vynechávame záznamy pasažierov, ktorí nemali pôvodne atribút HitType.

Ako vidíme, niektoré lety sú príliš malé - nie sú pravidelné, nesú málo pasažierov alebo sme vyfiltrovali veľké množstvo záznamov tým, že sme ignorovali záznamy s chýbajúcim

Zase sa však vyskytujú isté lety, ktoré sú dostatočne veľké a ich podiel nebezpečných pasažierov je vysoký - napríklad UN461. Čo je však ešte zaujímavejšie, let UN461 je pravidelný let medzi letiskami LED - Pulkovo Airport a letiskom v Pardubiciach. Ako sme si mohli všimnúť, letisko LED - Pulkovo Airport sa vyskytlo aj medzi nebezpečnými letiskami.

6.2.3 Neznámi spolucestujúci

Výstupom je zoznam dvojíc pasažierov a letov, na ktorých boli obaja títo pasažieri a zároveň necestovali na jednu rezerváciu. Tento výstup je formátovaný do .json-ovej databázy, ktorej obsah môže byť ďalej spracovaný podľa počtu spoločných letov bez spoločnej rezervácie.

Tabuľka 6.7: Neznámi spolucestujúci

Flights	Pairs
2	101226
3	5300
4	1197
5	410
6	140
7	73
8	31
9	23
10	12
11	12
12	6
13	3
14	3
16	2
18	2
21	1
24	1

V tejto tabuľke Flights odpovedá počtu spoločných letov bez spoločnej rezervácie a Pairs odpovedá počtu dvojíc pasažierov, ktoré tento počet spoločných letov má.

Ako vidíme, pasažierov, ktorí majú spoločné 2 lety a zároveň neletia na jednu rezerváciu je obrovské množstvo (tých, čo majú spoločný len jeden let by bolo ešte viac, ale tých sme nezahrnuli). Zaujímavejším faktom je však to, že sa vyskytujú pasažieri, ktorí spolu lietajú (nie však na jednu rezerváciu) pravidelne. Títo pasažieri by mohli byť označení na podrobnejšie preskúmanie, môže sa však jednať o obchodných cestujúcich so zdieľaným štátom/mestom obchodných ciest. Pre porovnanie dvojice, ktorá dosiahla 24 spoločných letov, jeden pasažier má celkovo 69 a druhý 66 letov a z nich drvivá väčšina začína na rovnakom letisku a teda fakt, že sa vyskytli v 24 spoločných letoch nie je až tak prekvapivý.

Kapitola 7

Budúce práce

Ako som už načrtol v analytickej časti svojej práce, je mnoho rôznych spôsobov náhľadu na problém detekcie anomálií v dátach, ktoré nám boli poskytnuté. Z mnohých prístupov som sa v tejto práci sústredil na jeden a preto v náväznosti na túto prácu by sa mohli spracovať aj ďalšie prístupy.

7.1 Voľba kontextu

Prvou možnosťou pokračovania je zvoliť iný spôsob definície kontextu v dátach. Pri rôzne vymedzených kontextoch sme schopní skúmať iné závislosti a tým pádom detekovať rôzne anomálie.

7.2 Voľba techniky detekcie anomálií

Ďalším spôsobom akým analyzovať a spracovávať dáta by mohlo byť skúmať ako supervised techniky detekcie anomálií, tak aj unsupervised techniky. Predpokladám, že pri oboch spôsoboch by bolo možné vytvoriť zaujímavý prístup k detekcii anomálií.

7.3 Skúmanie regiónov

Skúmanie politickej a kultúrnej podobnosti jednotlivých národností alebo regiónov, do ktorých patria dané letiská a tým aj definovanie vzdialenosti (dvoch hodnôt atribútu, nie geografickej vzdialenosti) je aj v spolupráci s Políciou ČR obtiažnou úlohou. Napriek tomu je to veľmi zaujímavá úloha, ktorá by mohla viesť k zdokonaleniu detekcie anomálií ako aj konštrukcie prediktívneho modelu pre určovanie nebezpečnosti pasažiera.

Záver

V tejto práci som preskúmal teoretický základ potrebný pre spracovanie leteckých dát poskytnutých Políciou ČR. Po analýze zadania a návrhu postupov na splnenie jednotlivých úloh som zrealizoval predspracovanie týchto dát do formy vhodnej na prácu s nimi - nahradil chýbajúce atribúty, odstránil prebytočné atribúty a zbavil sa ďalších nekonzistencií.

Ďalšou úlohou bolo detekovať anomálie v týchto predspracovaných dátach. Detekcia anomálií bola zrealizovaná vytvorením letových profilov jednotlivým pasažierom, uplatňovaním faktoru rozpadu na počty odletov z daného letiska a sledovaním uhlov vektorov reprezentujúcich profil pred pridaním a po pridaní nového záznamu.

Posledným bodom bolo dohodnutie analytických otázok v spolupráci s Políciou ČR. Týmito otázkami boli:

1. Je možné na základe poskytnutých dát zachytiť atribúty nebezpečného pasažiera? (označiť všetkých nebezpečných pasažierov)
   - Ak áno, s akou presnosťou vieme určiť týchto pasažierov?
   - Dokážeme vymodelovať „bezpečného pasažiera“?
2. Dajú sa určiť na základe týchto dát celé lety (alebo letiská), ktoré majú oproti ostatným vyššiu pravdepodobnosť, že v nich budú nebezpeční pasažieri?
3. Existujú ľudia, čo stále cestujú spolu v lietadle, ale nikdy nie na jednu rezerváciu?

Aby som bol schopný na tieto otázky odpovedať, najskôr som preskúmal akým spôsobom je vhodné tak učiniť a ako budeme schopní svoje tvrdenia podložiť. Podľa prvej otázky a jej podotázok som vykonal analýzu pomocou natrénovania rôznych modelov z dát reprezentujúcich pasažierov. Výsledky boli zachytené pomocou úspešnosti modelu overenou krížovou validáciou.

Pre odpoveď na druhú otázku som zvolil skúmanie pomerov bezpečných a nebezpečných pasažierov pre jednotlivé letiská a lety. Výsledky sú zhrnuté vo forme tabuliek s jednotlivými pomermi.

Na poslednú otázku som odpovedal vytvorením zoznamu spolucestujucich, ktorí necestujú na jednu rezerváciu a následným preskúmaním počtu spoločných letov. Výstupom sú dvojice pasažierov a ich spoločných letov. Vo výsledkovej časti sú zhodnotené jednotlivé počty spoločných letov.

Keďže dáta poskytnuté Políciou ČR sú zaujímavé z mnohých perspektív a poskytujú mnohé možnosti analýzy a spracovania, v sekcii 7 opisujem, akým smerom by sa mohli uberať ďalšie práce.

Literatúra

[1] Chandola, V.; Banerjee, A.; Kumar, V.: Anomaly Detection: A Survey. *ACM Comput. Surv.*, ročník 41, č. 3, Júl 2009: s. 15:1–15:58, ISSN 0360-0300, doi:10.1145/1541880.1541882. Dostupné z: http://doi.acm.org/10.1145/1541880.1541882

[2] Gnumeric Spreadsheet. http://www.gnumeric.org/, 12 2001.

[3] Hastie, T.; Tibshirani, R.; Friedman, J.: *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

[4] Clifton, C.: Data mining. https://www.britannica.com/technology/data-mining, 11 2009.

[5] RapidMiner. Online, Apríl 2012. Dostupné z: http://rapid-i.com/content/view/181/190/

[6] Joshi, M. V.; Agarwal, R. C.; Kumar, V.: Mining Needle in a Haystack: Classifying Rare Classes via Two-phase Rule Induction. *SIGMOD Rec.*, ročník 30, č. 2, Máj 2001, ISSN 0163-5808, doi:10.1145/376284.375673. Dostupné z: http://doi.acm.org/10.1145/376284.375673

[7] Abe, N.; Zadrozny, B.; Langford, J.: Outlier Detection by Active Learning. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, New York, NY, USA: ACM, 2006, ISBN 1-59593-339-5, doi:10.1145/1150402.1150459. Dostupné z: http://doi.acm.org/10.1145/1150402.1150459

[8] Fan, W.; Miller, M.; Stolfo, S. J.; aj.: Using Artificial Anomalies to Detect Unknown and Known Network Intrusions. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, Washington, DC, USA: IEEE Computer Society, 2001, ISBN 0-7695-1119-8. Dostupné z: http://dl.acm.org/citation.cfm?id=645496.658057

[9] Vasconcelos, G. C.; Fairhurst, M. C.; Bisset, D. L.: Investigating Feedforward Neural Networks with Respect to the Rejection of Spurious Patterns. *Pattern Recogn. Lett.*, ročník 16, č. 2, Február 1995, ISSN 0167-8655, doi:10.1016/0167-8655(94)00092-H. Dostupné z: http://dx.doi.org/10.1016/0167-8655(94)00092-H

[10] Agarwal, D.: An Empirical Bayes Approach to Detect Anomalies in Dynamic Multidimensional Arrays. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, Washington, DC, USA: IEEE Computer Society, 2005, ISBN 0-7695-2278-5, doi:10.1109/ICDM.2005.22. Dostupné z: http://dx.doi.org/10.1109/ICDM.2005.22

[11] Das, K.; Schneider, J.: Detecting Anomalous Records in Categorical Datasets. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, New York, NY, USA: ACM, 2007, ISBN 978-1-59593-609-7, doi:10.1145/1281192.1281219. Dostupné z: http://doi.acm.org/10.1145/1281192.1281219

[12] Vapnik, V. N.: *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995, ISBN 0-387-94559-8.

[13] Li, Y.; Pont, M. J.; Jones, N. B.: Improving the Performance of Radial Basis Function Classifiers in Condition Monitoring and Fault Diagnosis Applications Where ‘Unknown’ Faults May Occur. *Pattern Recogn. Lett.*, ročník 23, č. 5, Marec 2002, ISSN 0167-8655, doi:10.1016/S0167-8655(01)00133-7. Dostupné z: http://dx.doi.org/10.1016/S0167-8655(01)00133-7

[14] Joachims, T.: Training Linear SVMs in Linear Time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, New York, NY, USA: ACM, 2006, ISBN 1-59593-339-5, doi:10.1145/1150402.1150429. Dostupné z: http://doi.acm.org/10.1145/1150402.1150429

[15] Mahoney, M. V.; Chan, P. K.: Learning Rules for Anomaly Detection of Hostile Network Traffic. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, Washington, DC, USA: IEEE Computer Society, 2003, ISBN 0-7695-1978-4. Dostupné z: http://dl.acm.org/citation.cfm?id=951949.952127

[16] Hautamaki, V.; Karkkainen, I.; Franti, P.: Outlier Detection Using k-Nearest Neighbour Graph. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, Washington, DC, USA: IEEE Computer Society, 2004, ISBN 0-7695-2128-2, doi:10.1109/ICPR.2004.671. Dostupné z: http://dx.doi.org/10.1109/ICPR.2004.671

[17] Ramaswamy, S.; Rastogi, R.; Shim, K.: Efficient Algorithms for Mining Outliers from Large Data Sets. *SIGMOD Rec.*, ročník 29, č. 2, Máj 2000, ISSN 0163-5808, doi:10.1145/335191.335437. Dostupné z: `http://doi.acm.org/10.1145/335191.335437`

[18] Knorr, E. M.; Ng, R. T.: A Unified Approach for Mining Outliers. In *Proceedings of the 1997 Conference of the Centre for Advanced Studies on Collaborative Research*, CASCON '97, IBM Press, 1997. Dostupné z: `http://dl.acm.org/citation.cfm?id=782010.782021`

[19] Knorr, E. M.; Ng, R. T.: Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, ISBN 1-55860-566-5. Dostupné z: `http://dl.acm.org/citation.cfm?id=645924.671334`

[20] Knorr, E. M.; Ng, R. T.: Finding Intensional Knowledge of Distance-Based Outliers. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, ISBN 1-55860-615-7. Dostupné z: `http://dl.acm.org/citation.cfm?id=645925.671529`

[21] Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; aj.: LOF: Identifying Density-based Local Outliers. *SIGMOD Rec.*, ročník 29, č. 2, Máj 2000, ISSN 0163-5808, doi:10.1145/335191.335388. Dostupné z: `http://doi.acm.org/10.1145/335191.335388`

[22] Tang, J.; Chen, Z.; Fu, A. W.-C.; aj.: Enhancing Effectiveness of Outlier Detections for Low Density Patterns. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '02, London, UK, UK: Springer-Verlag, 2002, ISBN 3-540-43704-5. Dostupné z: `http://dl.acm.org/citation.cfm?id=646420.693665`

[23] Papadimitriou, S.; Kitagawa, H.; Gibbons, P. B.; aj.: LOCI: Fast Outlier Detection Using the Local Correlation Integral. In *ICDE*, 2003.

[24] Jain, A. K.; Dubes, R. C.: *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988, ISBN 0-13-022278-X.

[25] Guha, S.; Rastogi, R.; Shim, K.: ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Inf. Syst.*, ročník 25, č. 5, Júl 2000, ISSN 0306-4379, doi:10.1016/S0306-4379(00)00022-3. Dostupné z: `http://dx.doi.org/10.1016/S0306-4379(00)00022-3`

[26] Kohonen, T. (editor): *Self-organizing Maps*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997, ISBN 3-540-62017-6.

[27] He, Z.; Xu, X.; Deng, S.: Discovering Cluster-based Local Outliers. *Pattern Recogn. Lett.*, ročník 24, č. 9-10, Jún 2003, ISSN 0167-8655, doi:10.1016/S0167-8655(03)00003-5. Dostupné z: http://dx.doi.org/10.1016/S0167-8655(03)00003-5

[28] Jaing, M. F.; Tseng, S. S.; Su, C. M.: Two-phase Clustering Process for Outliers Detection. *Pattern Recogn. Lett.*, ročník 22, č. 6-7, Máj 2001, ISSN 0167-8655, doi:10.1016/S0167-8655(00)00131-8. Dostupné z: http://dx.doi.org/10.1016/S0167-8655(00)00131-8

[29] Soule, A.; Salamatian, K.; Taft, N.: Combining Filtering and Statistical Methods for Anomaly Detection. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*, IMC '05, Berkeley, CA, USA: USENIX Association, 2005. Dostupné z: http://dl.acm.org/citation.cfm?id=1251086.1251117

[30] Eskin, E.: Anomaly Detection over Noisy Data Using Learned Probability Distributions. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, ISBN 1-55860-707-2. Dostupné z: http://dl.acm.org/citation.cfm?id=645529.658128

[31] Aggarwal, C. C.; Yu, P. S.: Outlier Detection for High Dimensional Data. *SIGMOD Rec.*, ročník 30, č. 2, Máj 2001, ISSN 0163-5808, doi:10.1145/376284.375668. Dostupné z: http://doi.acm.org/10.1145/376284.375668

[32] Abraham, B.; Chuang, A.: Outlier Detection and Time Series Modeling. *Technometrics*, ročník 31, č. 2, Máj 1989, ISSN 0040-1706, doi:10.2307/1268821. Dostupné z: http://dx.doi.org/10.2307/1268821

[33] Lane, T.; Brodley, C. E.: Temporal Sequence Learning and Data Reduction for Anomaly Detection. *ACM Trans. Inf. Syst. Secur.*, ročník 2, č. 3, August 1999, ISSN 1094-9224, doi:10.1145/322510.322526. Dostupné z: http://doi.acm.org/10.1145/322510.322526

[34] Lee, W.; Xiang, D.: Information-Theoretic Measures for Anomaly Detection. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, SP '01, Washington, DC, USA: IEEE Computer Society, 2001. Dostupné z: http://dl.acm.org/citation.cfm?id=882495.884435

[35] Ando, S.: Clustering Needles in a Haystack: An Information Theoretic Analysis of Minority and Outlier Detection. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, Washington, DC, USA: IEEE Computer Society, 2007, ISBN 0-7695-3018-4, doi:10.1109/ICDM.2007.53. Dostupné z: http://dx.doi.org/10.1109/ICDM.2007.53

[36] Günter, S.; Schraudolph, N. N.; Vishwanathan, S. V. N.: Fast Iterative Kernel Principal Component Analysis. *J. Mach. Learn. Res.*, ročník 8, December 2007, ISSN 1532-4435. Dostupné z: http://dl.acm.org/citation.cfm?id=1314498.1314562

[37] IDÉ, T.; KASHIMA, H.: Eigenspace-based Anomaly Detection in Computer Systems. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, New York, NY, USA: ACM, 2004, ISBN 1-58113-888-1, s. 440–449, doi: 10.1145/1014052.1014102. Dostupné z: http://doi.acm.org/10.1145/1014052.1014102

[38] Lu, C.-T.; Chen, D.; Kou, Y.: Algorithms for Spatial Outlier Detection. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, Washington, DC, USA: IEEE Computer Society, 2003, ISBN 0-7695-1978-4. Dostupné z: http://dl.acm.org/citation.cfm?id=951949.952103

[39] Basu, S.; Meckesheimer, M.: Automatic Outlier Detection for Time Series: An Application to Sensor Data. *Knowl. Inf. Syst.*, ročník 11, č. 2, Február 2007, ISSN 0219-1377, doi:10.1007/s10115-006-0026-6. Dostupné z: http://dx.doi.org/10.1007/s10115-006-0026-6

[40] Keogh, E.; Lin, J.; Lee, S.-H.; aj.: Finding the Most Unusual Time Series Subsequence: Algorithms and Applications. *Knowl. Inf. Syst.*, ročník 11, č. 1, December 2006, ISSN 0219-1377, doi:10.1007/s10115-006-0034-6. Dostupné z: http://dx.doi.org/10.1007/s10115-006-0034-6

[41] Debar, H.; Dacier, M.; Nassehi, M.; aj.: Fixed vs. Variable-Length Patterns for Detecting Suspicious Process Behavior. In *Proceedings of the 5th European Symposium on Research in Computer Security*, ESORICS '98, London, UK, UK: Springer-Verlag, 1998, ISBN 3-540-65004-0. Dostupné z: http://dl.acm.org/citation.cfm?id=646647.699202

[42] Keogh, E.; Lonardi, S.; Chiu, B. Y.-c.: Finding Surprising Patterns in a Time Series Database in Linear Time and Space. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, New York, NY, USA: ACM, 2002, ISBN 1-58113-567-X, doi:10.1145/775047.775128. Dostupné z: http://doi.acm.org/10.1145/775047.775128

[43] Gwadera, R.; Atallah, M. J.; Szpankowski, W.: Reliable Detection of Episodes in Event Sequences. *Knowl. Inf. Syst.*, ročník 7, č. 4, Máj 2005, ISSN 0219-1377, doi:10.1007/s10115-004-0174-5. Dostupné z: http://dx.doi.org/10.1007/s10115-004-0174-5

no

Dodatok **A**

Zoznam použitých skratiek

ŘSCP Ředitelství služby cizinecké policie

AGPL Affero General Public License

API Application programming interface

CSV Comma separated values

JSON JavaScript Object Notation

NN Nearest Neighbour

LOF Local Outlier Factor

COF Connectivity-based Outlier Factor

ODIN Outlier Detection using In-Degree Number

MDEF Multi-granularity Deviation Factor

SOM Self-Organizing Maps

PCA Principal component analysis

RBF Radial Basis Function

SVM Support vector machine

Dodatok **B**

Obsah priloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `impl` ..... zdrojové kódy skriptov na spracovanie dát
  - `tables` ..... výsledkové tabuľky
  - `thesis` ..... zdrojová forma práce vo formáte LaTeX
- `text` ..... text práce
  - `tothmatu.pdf` ..... text práce vo formáte PDF