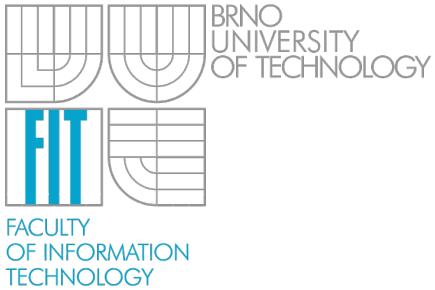
Anomaly Detection

Accelerated Network Technologies
Research Group

Václav Bartoš, Martin Žádník

Brno University of Technology, Faculty of Information Technology
Bozetechova 2, 612 00 Brno, CZ
http://merlin.fit.vutbr.cz/ant/



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Detekce anomálií

- *"Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior."*
- Obecný problém, nejen v počítačových sítích
- V oblasti sítí používáno k detekci útoků.
  - Na rozdíl od detekce útoků založené na vyhledávání vzorů dokáže odhalit i nové typy útoků

Detekce anomálií

- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Aplikace detekce anomálií

- Intrusion detection
  - Host based (system call based)
  - Network based
- Fraud detection
  - Credit cards, Mobile phones, Insurance claims, Insider trading
- Medical and public health
- Industrial damage detection
- Image processing, Anomalies in text data, Sensor networks, ...

Typy vstupních dat

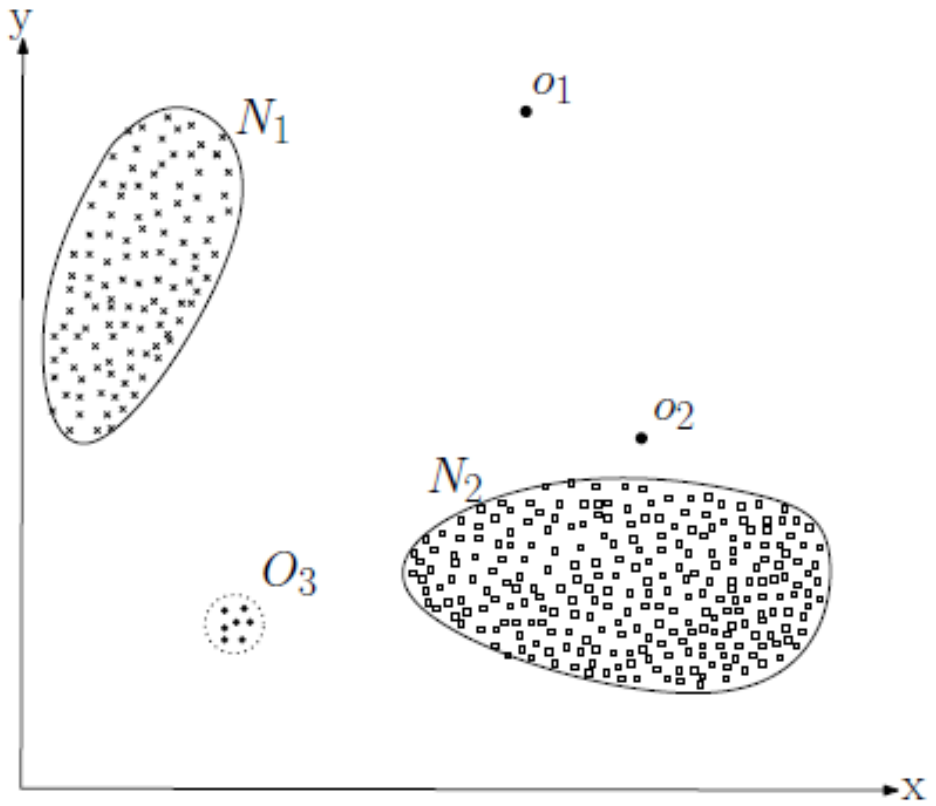
- Jednorozměrné
- Vícerozměrné
  - Stejné / různé datové typy

- Spojitá
- Diskrétní, kategorická

- Relace mezi instancemi
  - Body bez relací
  - Sekvence
  - Prostorová data
  - Grafy

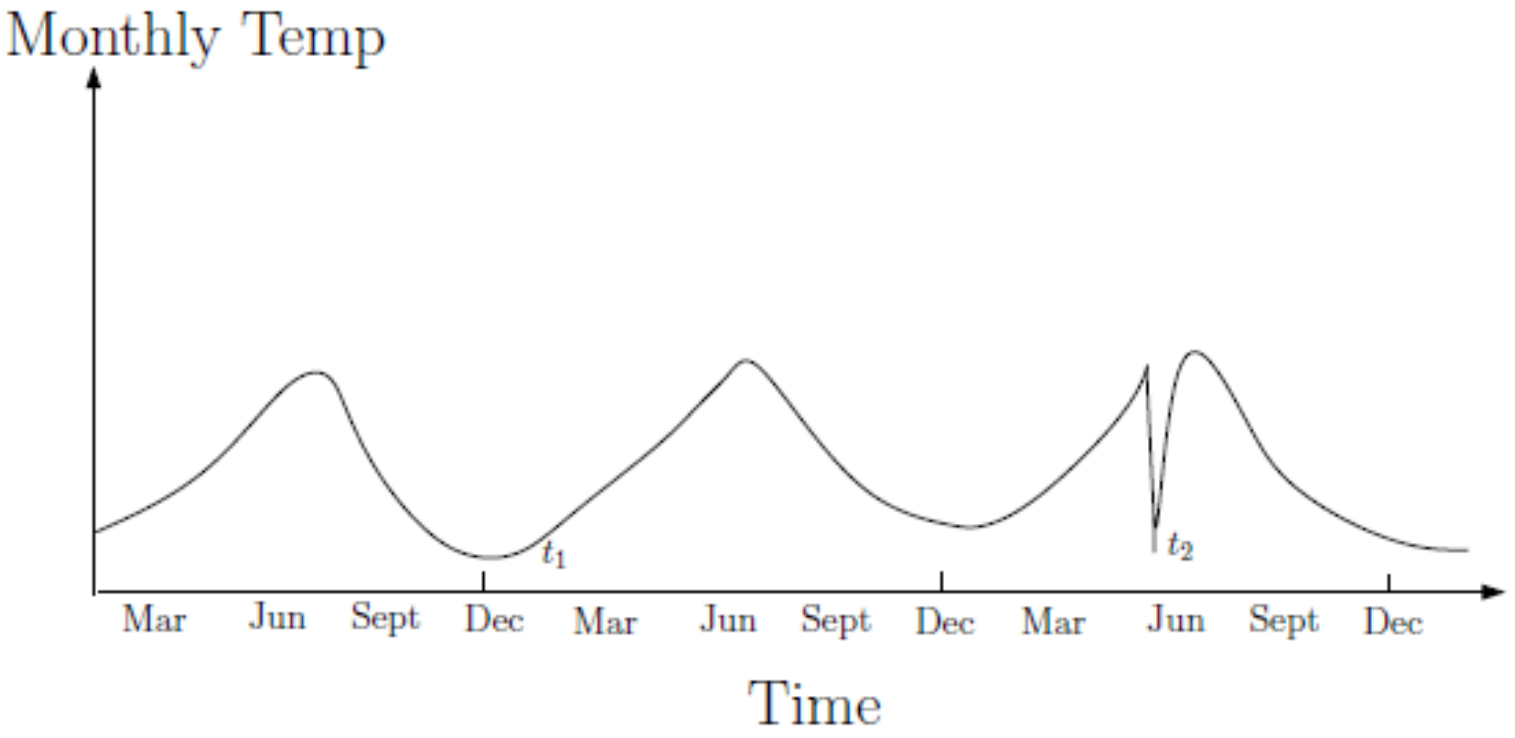
Typy Anomálií

- Bodové
  - Jedna instance dat je sama o sobě považována za anomálii
  - Záleží jen na jejích atributech / pozici v prostoru



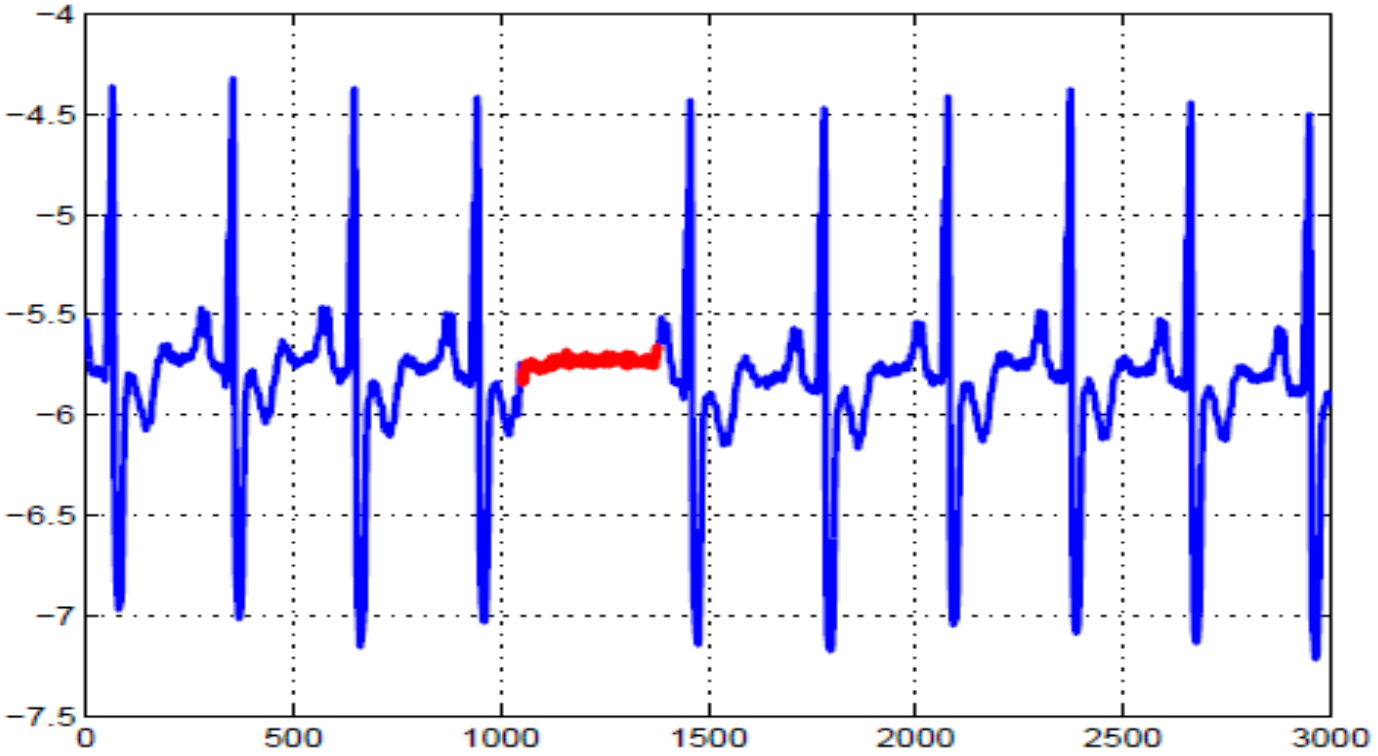
Typy Anomálií

- Kontextové
  - Instance se stává anomálií až díky nějakému kontextu
  - Kontextové atributy - např. pozice v prostoru, sekvenci



Typy Anomálií

- Kolektivní
  - Několik instancí je samo o sobě normální, ale jako skupina tvoří anomálii.



Typy učení

- Supervised AD
  - Jsou k dispozici popsaná data jak normální, tak anomálie, učení s učitelem
  - Data pro všechny anomálie ale obvykle nejsou k dispozici
- Semi-supervised AD
  - K dispozici jsou popsaná normální data, učení s učitelem
- Unsupervised AD
  - Nejsou k dispozici popsaná data, učení bez učitele

Výstup detekce anomálií

- Score
  - Pro každou instanci dat je určeno číslo vyjadřující, jak moc je daná instance anomální
  - Lze použít threshold, nebo analyzovat top *n* anomálií
- Label
  - Data se roztřídí do skupin normální / anomální

Detekce anomálií

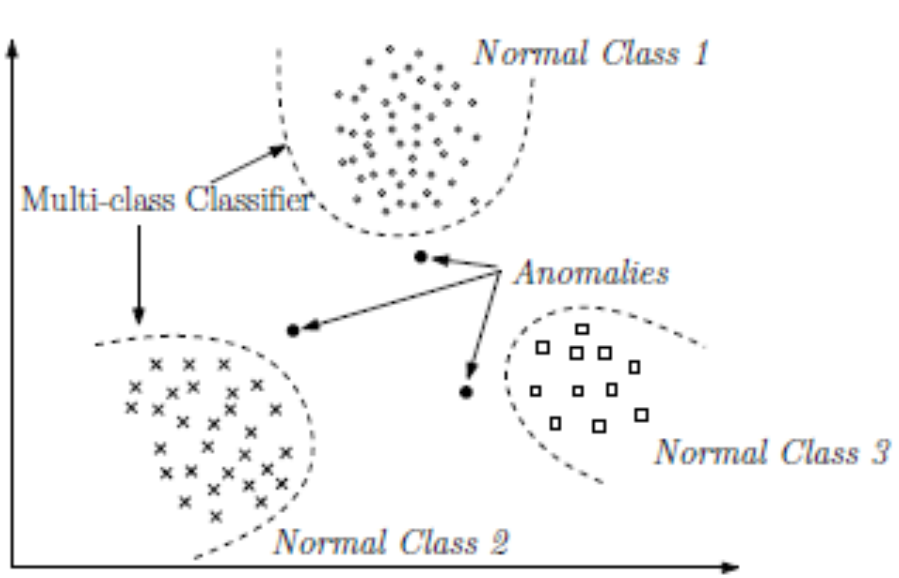
- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Classification based AD

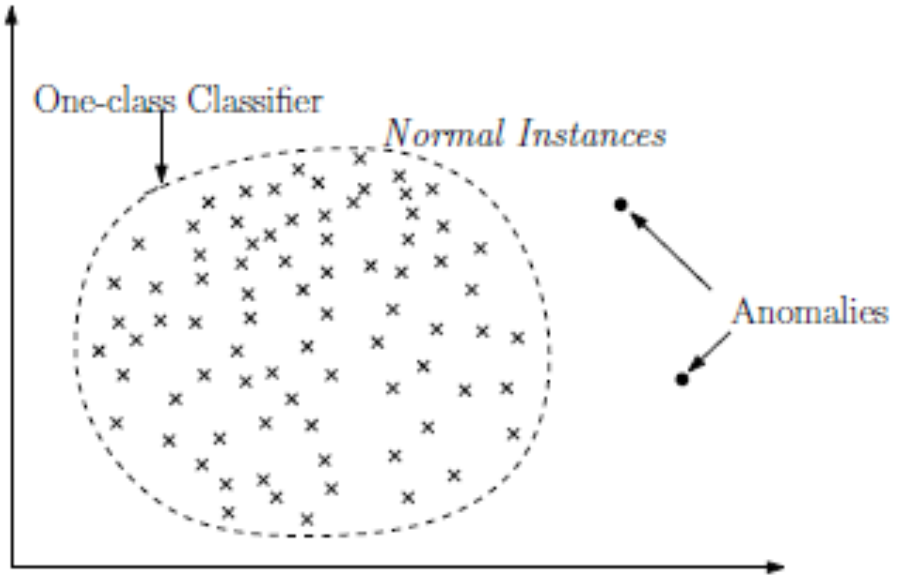
- Předpoklad:
  - *"A classifier that can distinguish between normal and anomalous classes can be learnt in the given feature space."*
- Klasifikátor, který třídí data do dvou tříd - normální nebo anomálie
- Učí se na popsaných datech

Classification based AD

- *Multi-class* klasifikátor
  - Popsaná data pro více normálních tříd
- *One-class* klasifikátor
  - Pouze jedna normální třída



(a) Multi-class Anomaly Detection



(b) One-class Anomaly Detection

Classification based AD

- Neural networks
  - Nejdříve se síť naučí na normálních datech
  - Pak testovací fáze - síť data přijme / nepřijme
- Bayesian networks
  - Multi-class
  - Pomocí Bayesovských sítí se určují pravděpodobnosti, že instance dat patří do jisté třídy (jedna z tříd znamená anomálii)
  - Několik variant použito pro Network Intrusion Detection
- Support Vector Machines
  - One-class
  - SVM se naučí oblast normálních dat
  - Pokud testovací instance leží mimo tuto oblast - anomálie
  - Robust SVM - odolné proti výskytu anomálií v testovacích datech

Classification based AD

- Rule based
  - One-class i multi-class
  - Algoritmy na odvozování pravidel (RIPPER, Decision trees, ...) z trénovacích dat
  - Každé pravidlo má "hodnotu důvěry"
    - Poměr množství trénovacích dat odpovídajících pravidlu ku celkovému množství dat
  - Pak se hledá pravidlo, které nejlépe odpovídá instanci dat, inverzní hodnota jeho důvěry je anomaly score instance
  - Pro one-class použito "Association rule mining"
    - Unsupervised technika
    - Používáno v Intrusion Detection

Classification based AD

- Výpočetní složitost
  - Trénovací fáze
    - velmi záleží na použité klasifikační metodě
  - Testovací fáze:
    - Většinou velmi rychlé
- Výhody:
  - Po naučení velmi rychlé
- Nevýhody
  - Multi-class metody potřebují popsaná data pro různé normální třídy, což často není k dispozici
  - Data se označují jako normální/anomální, není score

Nearest Neighbor based AD

- Předpoklad:
  - *"Normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors."*
- Definována míra vzdálenosti (podobnosti) mezi dvěma instancemi dat.
- Dvě skupiny, anomaly-score bodu vychází ze
  - Vzdálenosti k-tého nejbližšího souseda
  - Relativní hustoty bodů v okolí

Detekce anomálií

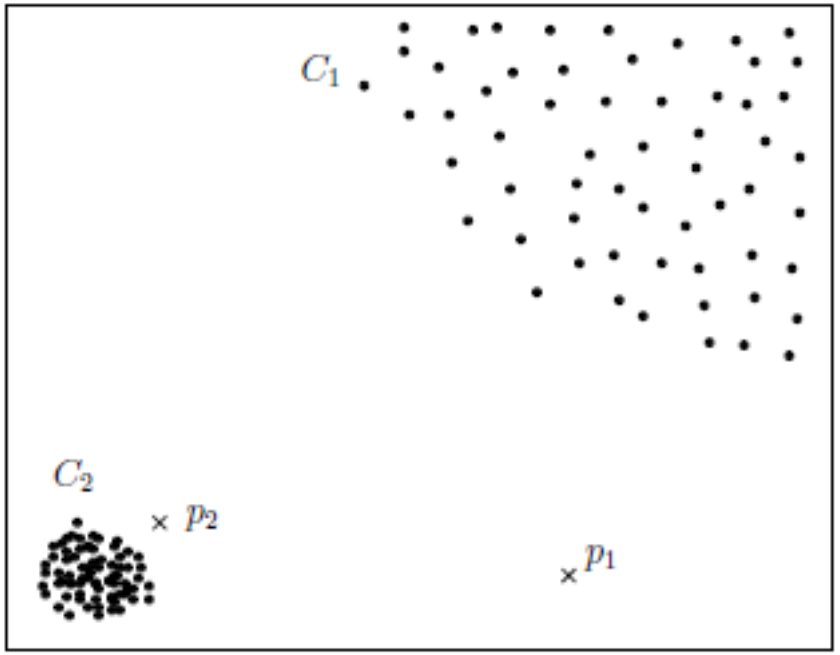
- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Nearest Neighbor based AD

- K-tý nejbližší soused
  - *Základ*: Anomaly score bodu je rovno vzdálenosti k jeho *k*-tému nejbližšímu sousedovi v dané množině dat
  - Součet vzdáleností *k* nejbližších sousedů
  - Počet sousedů do dané vzdálenosti (globální hustota)
  - Různé metody pro zpracování diskrétních datových typů
  - Techniky pro snížení výpočetní náročnosti
    - Různé prořezávání, lze snížit z $O(N^2)$ skoro na $O(N)$
    - Sampling

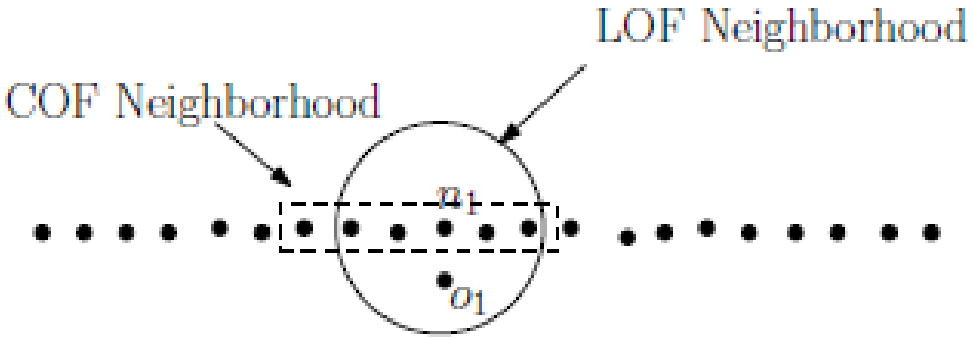
Nearest Neighbor based AD

- Relativní hustota
  - Malá hustota bodů v okolí - anomálie
  - *Základ:* Vzdálenost ke *k*-tému nejbližšímu sousedu = poloměr hyperkoule obsahující *k* nejbližších sousedů = převrácená hodnota hustoty okolí
  - Problém když mají normální data oblasti s různou hustotou
  - Řešení - *Local Outlier Factor*



Nearest Neighbor based AD

- *Local Outlier Factor (LOF)*
  - Lokální hustota = *k* / objem nejmenší hyperkoule obsahující *k* sousedů
  - LOF = Průměrná lok. hustota *k* sousedů / vlastní lok. hustota
  - Normální instance má lokální hustotu podobnou jako její sousedi, anomálie má tuto hustotu menší
- Opět mnoho variant vylepšujících základní LOF
  - Adaptace pro složitější datové typy
  - Snížení výpočetní složitosti
  - *Connectivity based Outlier Factor (COF)*
  - A další



Nearest Neighbor based AD

- Výpočetní složitost
  - Výpočet vzdáleností mezi všemi body - O(N^2)
  - Mnoho technik, jak složitost snížit
- Výhody
  - Učení bez učitele (nejsou třeba trénovací data) a žádné předpoklady o distribuci dat.
  - Lze i s učitelem, pak méně false negatives
  - Jednoduchá adaptace na různé datové typy, stačí definovat funkci pro určení vzdálenosti.

Nearest Neighbor based AD

- Nevýhody
  - Pro techniky bez učitele musí platit, že všechna normální data se vyskytují ve shlucích ale anomálie ne.
  - Pro techniky s učitelem je potřeba velké množství trénovacích dat, jinak je generováno mnoho false-positives.
  - Velká výpočetní složitost
  - Velmi záleží na použité funkci pro výpočet vzdálenosti dvou bodů - nalézt vhodnou funkci může být pro složitější datové struktury obtížné.

Detekce anomálií

- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Clustering based AD

Typ 1:

- Předpoklad:
  - *"Normal data instances belong to a cluster in the data, while anomalies either do not belong to any cluster."*
- Metoda:
  - Je použit nějaký běžný shlukovací algoritmus, který nevyžaduje aby každá položka nutně patřila do nějakého shluku.
  - Co nepatří do žádného shluku je anomálie.
- Nevýhoda:
  - Shlukovací algoritmy jsou optimalizovány pro hledání shluků, ne anomálií.

Clustering based AD

Typ 2:

- Předpoklad:
  - *"Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid."*
- Metoda:
  - Je použit nějaký běžný shlukovací algoritmus.
  - Anomaly score je vypočítáno jako vzdálenost od středu nejbližšího shluku.
- Může pracovat i s učitelem - clustering probíhá na trénovacích datech bez anomálií.
- Nefunguje, pokud i anomálie tvoří shluky.

Clustering based AD

Typ 3:

- Předpoklad:
  - *"Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters."*
- Shluky které jsou menší a/nebo řidší než daná mez jsou označeny jako anomálie.

Clustering based AD

- Výpočetní složitost:
  - Složitost učení závisí na použitém shlukovacím algoritmu
    - Většinou O(N^2)
    - Metody používající heuristiky i O(N)
  - Testovací fáze pak velmi rychlá
- Výhody:
  - Mohou pracovat bez učitele
  - Většinou lze adaptovat pro složité datové typy jen použitím vhodného shlukovacího algoritmu
  - Po naučení velmi rychlé

Clustering based AD

- Nevýhody:
  - Velmi záleží na schopnosti shlukovacího algoritmu zachytit strukturu shluků normálních dat.
  - Některé techniky detekují anomálie jako vedlejší produkt shlukování, nejsou optimalizovány pro AD.
  - Některé algoritmy vynucují, aby byla každá instance v nějakém shluku, což může způsobit že je anomálie přiřazena ke shluku, i když tam nepatří.
  - Některé metody fungují jen když anomálie netvoří shluky.
  - Velká výpočetní složitost shlukovaní.

Detekce anomálií

- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Statistical AD

- Předpoklad:
  - *"Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model."*
- Na data je namapován statistický model (obvykle normálního chování), poté se provádějí statistické testy pro určení, zda nová data modelu odpovídají.
- Instance, které jsou podle aplikovaného testu daným modelem generovány s malou pravděpodobností, jsou anomálie.

Statistical AD

- Parametric techniques
  - Předpokládají znalost typu distribuce dat, určují se její přesné parametry
  - Gaussian model based
    - Gausovská distribuce dat
    - Určují se parametry (*Maximum Likehood Estimates*)
    - Podle vzdálenosti od středu distribuce se určuje jestli jde o anomálii - různé metody určení hranice
    - Další: Grubb's test, *t*-test, X^2
  - Regression model based
    - Z dat je vytvořen regresní model, anomaly score jednotlivých instancí je pak vypočítáno na základě odchylky od tohoto modelu

Statistical AD

- Mixture of parametric distributions based
  - Používá více distribučních funkcí
  - Buď dva modely (distribuční funkce) - pro normální data a anomálie
  - Nebo několik modelů pro různá normální data

Statistical AD

- Non-parametric techniques
  - Struktura modelu je vzata ze samotných dat
  - Histogram based
    - Z trénovacích dat se vytvoří histogram, anomaly score pak podle existence/velikosti sloupce, kam testovací data padnou
    - Hlavní problém - vhodná šířka sloupců
    - Varianty pro vícerozměrná data často používány pro network intrusion detection
  - Kernel function based
    - Využívá "kernel functions" pro určení funkce distribuce pravděpodobnosti pro normální data

Statistical AD

- Výpočetní složitost
  - Záleží na použitém statistickém modelu.
  - Určení parametrů jednoduchých distribucí (Gaussova, Poissonova, ...) většinou O(N)
  - Pro složitější typy distribucí se používají iterační techniky, většinou lineární na iteraci, ale může pomalu konvergovat.
  - Kernel based - O(N^2)

Statistical AD

- Výhody
  - Spolu s anomaly score většinou dostaneme i tzv. interval spolehlivosti, což je další užitečná informace k rozhodování.
  - Pokud použité technice odhadu distribuce dat nevadí anomálie v trénovacích datech, lze použít učení bez učitele.
- Nevýhody
  - Předpokládá se, že data jsou generována nějakou distribuční funkcí, což často nemusí být pravda.
  - I když je předpoklad správný, existuje mnoho metod na testování hypotéz, výběr té správné není jednoduchý.
  - Techniky založené na histogramech jsou sice jednoduché, ale v případě vícerozměrných dat nedokážou zachytit vazby mezi jednotlivými dimenzemi.

Detekce anomálií

- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Information theoretic AD

- Předpoklad:
  - *"Anomalies in data induce irregularities in the information content of the data set."*
- Základní technika:
  - Nechť C(D) je složitost (complexity) množiny dat D.
  - Pak se snažíme najít nejmenší $I \subset D$ takové, aby C(D) − C($D - I$) bylo maximální.
  - Prvky v I jsou anomálie
  - Optimalizace na dva různé cíle - hledá se *pareto optimum*

Information theoretic AD

- Různé metody pro výpočet složitosti C(*D*)
  - Kolmogorov complexity
    - Kolik dat potřebujeme k popisu objektu
    - Velikost regulárního výrazu popisujícího data; komprese dat
  - Entropy, relative uncertainty, ...
- Lze modifikovat pro sekvenční data, grafy apod.
  - *I* je pak podsekvence, podgraf, ...
- Výpočetní složitost:
  - Nutno vyzkoušet všechny podmnožiny - Exponenciální časová složitost
  - Byly navrženy různé aproximační techniky pracující v lineárním čase

Information theoretic AD

- Výhody
  - Unsupervised, učení bez učitele.
  - Žádné předpoklady o distribuci dat.
- Nevýhody
  - Výkonnost velmi závisí na zvolené míře složitosti, často tyto míry dokážou detekovat anomálie jen když jich je v datech velké množství.
  - Techniky pro sekvence apod. závisí na volbě velikosti podsekvencí. Zjistit vhodnou velikost často není jednoduché.
  - Je obtížné získat anomaly score.

Detekce anomálií

- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Spectral AD

- Předpoklad:
  - *"Data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different."*
- Snaží se data aproximovat projekcí do takového méněrozměrného podprostoru, ve kterém by šly anomálie snadno detekovat.

File:GaussianScatterPCA.png

- Založeno na *Principal Component Analysis* (PCA)
  - Transformace několika (korelovaných) proměnných na menší počet nezávislých proměnných (*hlavní komponenty*).
- Varianta *robust PCA* použita v IDS.

Spectral AD

- Výpočetní složitost:
  - Většinou lineární vzhledem k množství dat, ale kvadratické k počtu dimenzí
  - Lze vylepšit na lineární k množství dimenzí, ale polynomiální k množství dat.
- Výhody
  - Redukují počet dimenzí - vhodné pro mnohorozměrná data.
  - Lze použít jako předzpracování pro jinou metodu.
  - Učení bez učitele.
- Nevýhody
  - Funguje jen pokud jsou normální data a anomálie oddělitelné i v méněrozměrném podprostoru.
  - Velká výpočetní složitost

Detekce anomálií

- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

Kontextové anomálie

- Doposud většinou bodové anomálie.
- Některá data jsou sama o sobě normální, až v kontextu s dalšími daty jsou anomální.
- Kontextová data
  - Pozice v prostoru, čas, pořadí v sekvenci, uživatel, ...
- Dvě základní metody
  - Redukce na detekci bodové anomálie
  - Přímé využití struktury dat

Kontextové anomálie

- Redukce na detekci bodové anomálie
  - Nejdříve se identifikuje kontext dané instance dat
  - Pak se pomocí běžného "bodového" algoritmu vypočítá anomaly score v tomto kontextu
  - Např.:
    - Podmíněná pravděpodobnost
    - Data pro každého uživatele/skupinu zvlášť
    - Převod časových dat na vektory ve fázovém prostoru
- Přímé využití struktury dat
  - Hlavně u časových dat a sekvencí událostí
  - Vytváří se model, který předvídá chování na základě kontextu. Pokud se skutečná data značně odchylují od tohoto odhadu - anomálie.

Kolektivní anomálie

- Několik instancí dat může být samo o sobě normální, ale až to, že se vyskytují společně, z nich dělá anomálii.
- Nutné relace mezi daty
  - Prostorové, **sekvenční**, grafy

- Anomální sekvence v množině sekvencí
- Anomální podsekvence v dlouhé sekvenci

Kolektivní anomálie

- Anomální sekvence v množině sekvencí
  - Redukce na detekci bodové anomálie
  - Převedení sekvence do nějakého vícerozměrného prostoru
  - Stejně dlouhé sekvence
    - Jednoduše se každý prvek sekvence vezme jako jedna dimenze vektoru a na tyto vektory se aplikuje algoritmus hledání bodových anomálií.
  - Různě dlouhé sekvence
    - Transformují se na struktury s konstantním počtem položek.
  - Nebo lze definovat funkci určující vzdálenost (podobnost) mezi dvěma sekvencemi různé délky.

Kolektivní anomálie

- Pro nezarovnané sekvence - *Modeling sequences*:
  - Semi-supervised
  - Učení pravidel
  - Konečné automaty
    - použito pro detekci anomálií na síti
    - sekvence je zpracovávána FSM, když se nedosáhne koncového stavu - anomálie
  - Markovovy modely, HMM
  - Probabilistic suffix trees

Kolektivní anomálie

- Anomální podsekvence v dlouhé sekvenci
  - Většinou unsupervised
  - Problémy:
    - Délka anomální podsekvence může být různá.
    - Vstupní data obsahují anomálie, je těžké získat model normálního chování.
  - Jedna z metod:
    - Rozdělit sekvenci na podsekvence tak, aby počet bitů potřebných k zakódování každé podsekvence byl minimální.
    - Sekvence s nejvyšším počtem bitů nutných k zakódování jsou anomálie.
  - Nebo např. posouvající se okno a porovnávání podsekvence v okně se zbytkem celé sekvence.
  - A další metody

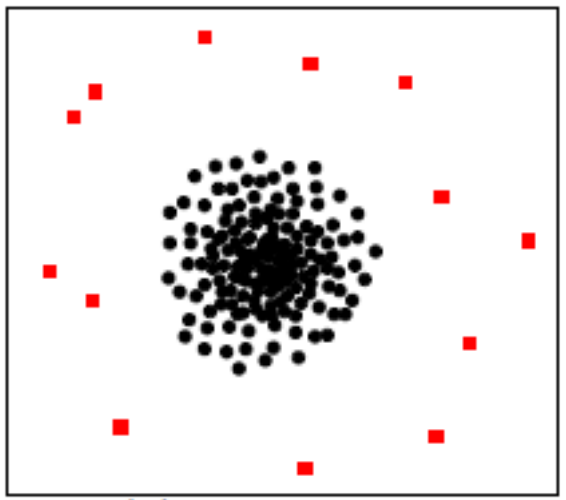
Detekce anomálií

- Úvodní informace
  - Oblasti aplikace, vstupní data, typy anomálií, ...
- Techniky detekce anomálií
  - Classification based
  - Nearest Neighbor based
  - Clustering based
  - Statistical
  - Information theoretic
  - Spectral
- Kontextové a kolektivní anomálie
- Shrnutí, srovnání

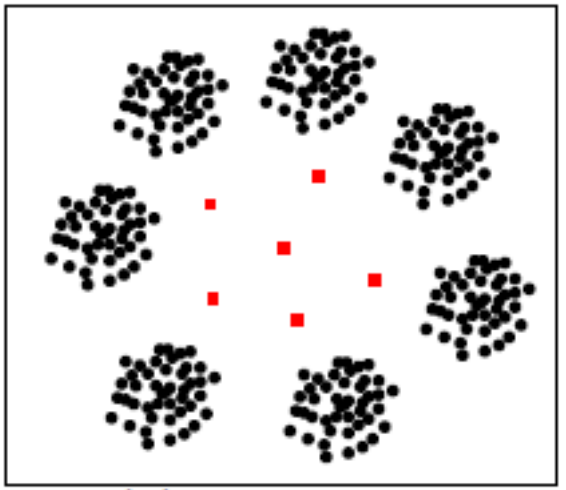
Shrnutí, srovnání

a)Jednoduchý případ

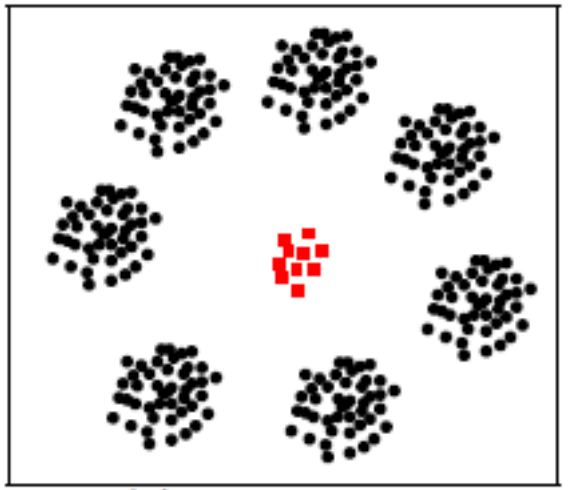
- Normální data generována gaussovskou distribuční funkcí
- Zvládnou všechny uvedené algoritmy



(a) Data Set 1



(b) Data Set 2

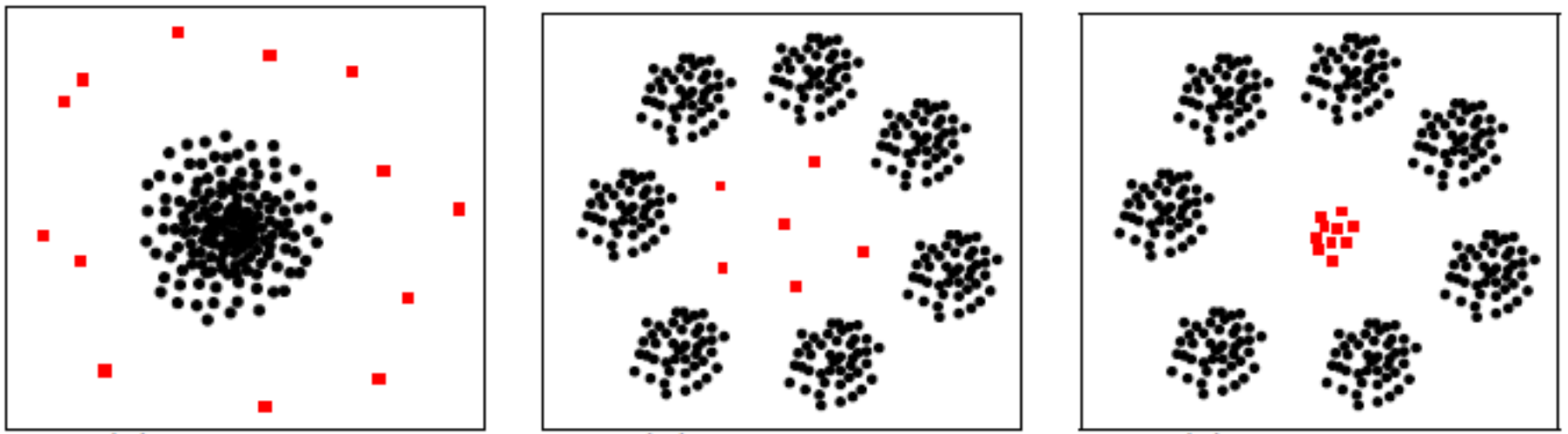


(c) Data Set 3

Shrnutí, srovnání

b)více tříd normálních dat

- Např. one-class klasifikátor nebo statistické metody používající jednu distribuční funkci nezvládnou
- Multi-class klasifikátor, k-nearest neighbor, clustering a další zvládnou

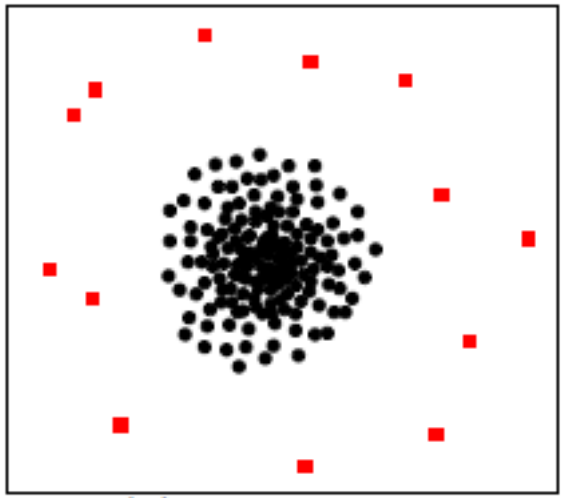


(a) Data Set 1 (b) Data Set 2 (c) Data Set 3

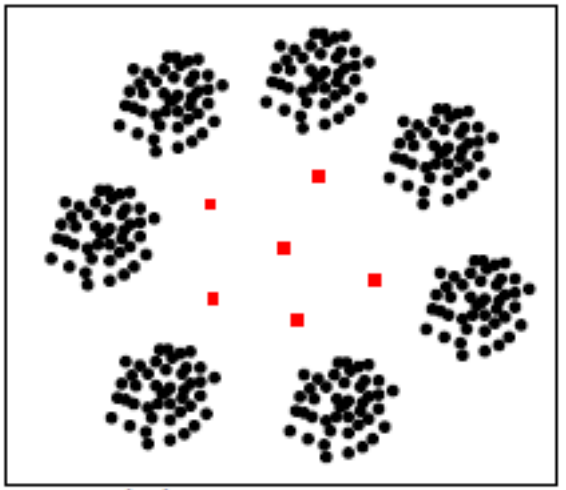
Shrnutí, srovnání

c) Anomálie tvoří malý shluk

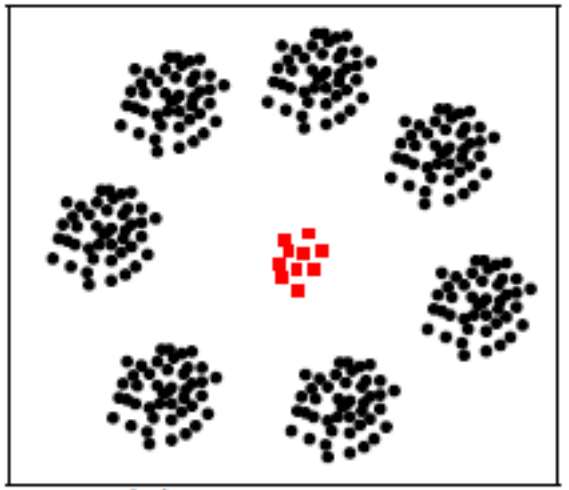
- K-nearest neighbor ani clustering nefungují
- Obecně zde mohou fungovat jen techniky s učitelem
-



(a) Data Set 1



(b) Data Set 2



(c) Data Set 3

Shrnutí, srovnání

- Množství dimenzí
  - K-nearest neighbor a clustering nevhodné pro velký počet dimenzí
    - Míry vzdálenosti ve vyšším množství dimenzí nedokážou rozlišit normální a anomální instance
  - Naopak spektrální metody jsou na to velmi vhodné
    - Ovšem nelze je použít vždy (podmínka, že anomálie musí být odlišitelné i v nižším počtu dimenzí)
  - Klasifikační techniky také vhodné pro velký počet dimenzí
    - Ale potřebují trénovací data, nejlépe normální i anomálie
    - I když jsou k dispozici, bývá problém rozdíl v jejich poměru
  - Statistické metody vhodné jen pro malý počet dimenzí

Shrnutí, srovnání

- Rychlost
  - Klasifikační metody, clustering a statistické metody mají pomalou fázi učení, ale potom jsou velmi rychlé.
  - Nearest-neighbor, information theoretic a spektrální metody nemají fázi učení, ale mívají pomalou testovací fázi.

- Ve výjimečných případech je anomálií více než normálních dat (detekce červů na poč. síti), pak lze použít jedině supervised nebo semi-supervised techniky.

Konec

Dotazy, diskuse ...