

# Откривање аномалија

Ненад Митић

Математички факултет  
`nenad@matf.bg.ac.rs`







# Шум и аномалије

## Шум

- погрешна вредност или догађај са грешком. Нпр.
  - тежина је погрешно записана
  - мерење тежине лимуна /лимета
- случајан догађај
- не мора да произведе неубичајене вредности/објекте
- није од интереса у истраживању

Аномалије јесу од интереса ако нису резултат шума

# Технике откривања аномалија

Претпоставка: постоји значајно већи број "нормалних" него "ненормалних" података (аномалија) у посматраном материјалу

## Технике

- засноване на формирању модела
- са визуелизацијом
- засноване на статистици
- засноване на одређивању растојања
- засноване на одређивању густине

# Карактеристике процеса откривања

## Број атрибута

- један (униваријантне методе)
- више атрибута (мултиваријантне методе)
- теже за откривање ако се користе сви атрибути
- шум / неупотребљиви атрибути
- аномалија само у односу на неке од атрибута
- ни један од атрибута нема аномалију али комбинација има (нпр. тежина x висина)

# Карактеристике процеса откривања (наставак)

- Глобална /локална перспектива посматрања
- Величина аномалије
- Истовремено одређивање једне или више аномалија
- Ефикасност



# Величина аномалије

- Методе које дају само бинарну карактеризацију (јесте/није)
  - најчешће засноване на класификацији
- Методе које свакој тачки додељују скор / величину аномалије
  - Величина аномалије представља степен по коме је објекат рангиран као аномалија
  - Праг величине
  - Број аномалија зависи од прага, контекста у коме се посматрају подаци, ...

# Варијанте проблема откривања аномалија

- За дати skup  $D$  наћи све тачке  $x \in D$  чија је величина аномалије већа од неког прага  $t$
- За дати skup  $D$  наћи све тачке  $x \in D$  које имају  $n$  највећих вредности величине аномалије
- За дати skup  $D$  који највећим делом садржи нормалне али неозначене тачке и тестну тачку  $x$ , одредити њену величину аномалије у односу на skup  $D$

# Методе засноване на формирању модела

Два корака:

- 1 Направи се модел са 'нормалним' понашањем на изабраном скупу
  - Са надгледањем
    - Аномалије су тачке које се не уклапају добро у карактеристике
    - Аномалије су тачке које нарушавају изглед модела
  - Ненадгледани модели
    - Аномалије су тачке које припадају ретким класама
- 2 Користећи направљен модел налазе се подаци који одскачу

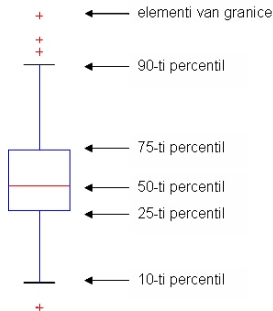
Избор и прецизно одређивање подскупа је захтевно ако је скуп података јако велики

# Методе засноване на визуелизацији

Корисне ако се подаци представљају у мањем броју димензија

Проблем визуелизације мултидимензионих података

Ограничење: подложне субјективној оцени података



# Методе засноване на статистици

Елемент ван граница је објекат који има мању вероватноћу у односу на вероватноћу у односу на дистрибуцију вероватноћа у моделу података

- Претпоставља се познавање дистрибуције података
- Статистички тест зависи од саме дистрибуције, њених параметара, и постављеног прага поузданости
- Проблем: дистрибуција је често непозната, или подаци имају мешавину дистрибуција

# Методе засноване на статистици

Најједноставнији примери: унимодална статистика

Пример: број година:

године = {3, 56, 23, 39, 156, 52, 41, 22, 9, 28, 139, 31, 55, 20, -67, 37, 11, 55, 45, 37}

- Статистички параметар: средина  $m = 39.9$ , стандардна девијација  $\sigma = 45.65$
- Избором прага:  $m \pm 2 \times \sigma$  добија се да су сви подаци ван скупа  $[-54.1, 131.2]$  потенцијални елементи ван граница
- Са великом вероватноћом добија се да су подаци ван граница -67, 139 и 156.

# Z-вредност

Унимодалне статистике користе тест поузданости крајева, односно веровантоћу да се елемент налази на крајевима

Функција густине за нормалну расподелу

$$f_X(x) = \frac{1}{\sigma \times \sqrt{2 \times \pi}} \times e^{-\frac{(x-\mu)^2}{2 \times \sigma^2}}$$

- Стандардна нормална расподела има средину 0 и девијацију 1.
- У одређеним случајевима средина и девијација могу да буду познате унапред

# Z-вредност

- Алтернативно, код велике количине података,  $\mu$  и  $\sigma$  могу имају високу поузданост и могу да послуже за рачунање Z вредности за случајну променљиву.
- вредност за посматрани податак  $x_i$  је  $z_i = (x_i - \mu) / \sigma$
- Велике апсолутне вредности  $z_i$  одговарају горњим и доњим границама
- Нормална расподела може да се прикаже преко Z вредности јер у том случају одговара скалираној и транслираној случајној променљивој са средном 0 и девијацијом 1





# Рачунање изгледних вероватноћа

- Претпоставка је да skup  $D$  садржи примерке са мешавином две расподеле
  - $M$  (расподела већине 'нормалних' података)
  - $A$  (расподела података са аномалијом)
- Приступ:
  - Иницијално претпоставка је да сви подаци имају расподелу  $M$
  - Нека је  $L_t(D)$  претпостављена вероватноћа припадности за  $D$  у тренутку  $t$
  - Сваку тачку  $x_t \in M$  преместити у  $A$  и одредити  $L_{t+1}(D)$  и  $\Delta = L_t(D) - L_{t+1}(D)$
  - Ако је разлика  $>$  прага тада је  $x_t$  аномалија

# Особине метода заснованих на статистици

- Строга математичка заснованост
- Велика ефикасност
- Добри резултати ако је позната расподела
- Проблеми у процени код вишедимензионих података
- Аномалије могу да утичу на параметре расподеле

# Методе засноване на одређивању растојања

Објекат је аномалија/ елемент ван граница ако је цео објекат или његов део удаљен више од предвиђене границе

Више техника

- К-најближих суседа (важан избор  $k$ )
- Рачунање растојања - Махаланобисово растојање
- Растојање се одређује између тачке  $x$  и средине  $\bar{x}$  скупа података

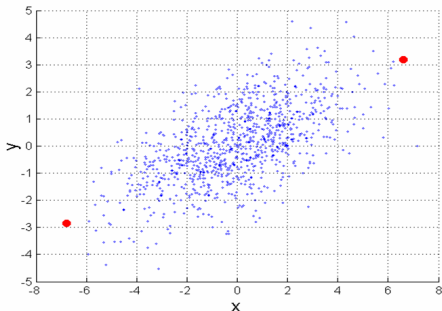
$$Mahalanobis(x, \bar{x}) = \sqrt{(x - \bar{x})\Sigma^{-1}(x - \bar{x})^T}$$

где је  $\Sigma^{-1}$  инверзна матрица матрице коваријанси података

- ...

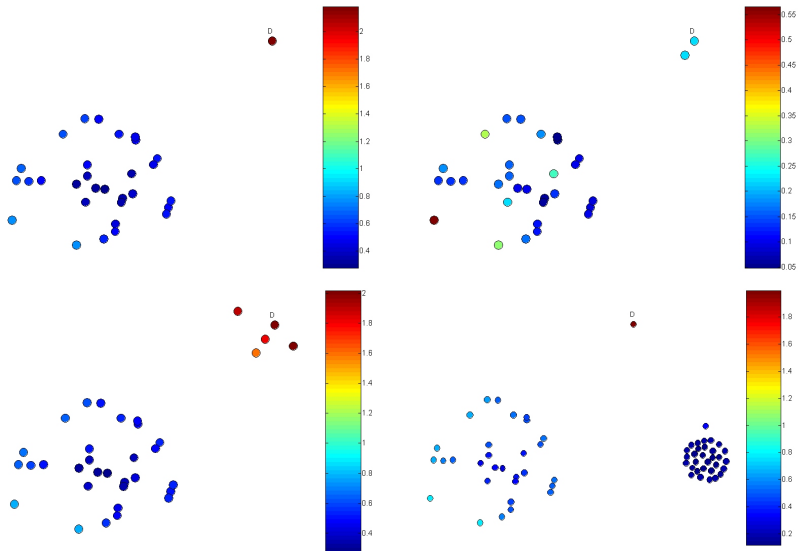
# Утицај расподеле на растојање

Међусобно растојање тачака  $A(-6.8, -2.9)$  и  $B(6.8, 3.1)$



Еуклидско растојање тачака је 14.7, а Махаланобисово 6

# Одређивање растојања K-нн



# Особине метода заснованих на растојању

- Једноставне су за примену
- Рачунарски захтевне -  $O(n^2)$
- Осетљиве на промене параметара
- Проблем са одређивањем растојања у вешедимензионом простору

# Методе засноване на одређивању густине

Величина аномалије објекта је обрнуто пропорционална густини елемената у његовом окружењу

Више техника

- $K$ -најближих суседа (важан избор  $k$ ) - инверзно од растојања до  $k$  суседа
- Инверзно просечном растојању до  $k$  суседа
- DBSCAN
- друге методе кластеровања

Проблем код региона са различитом густином



# Густина према Кнн суседа

Густина = инверзно од растојања до Кнн суседа

$$gustina(x, k) = \left( \frac{\sum_{y \in N(x, k)} \text{rastojanje}(x, y)}{|N(x, k)|} \right)^{-1}$$

где је  $N(x, k)$  скуп који садржи  $k$  најближих суседа од  $x$ ,  
 $|N(x, k)|$  је величина тог скупа, а  $y$  је најближи сусед.

# Релативна густина према Кнн суседа

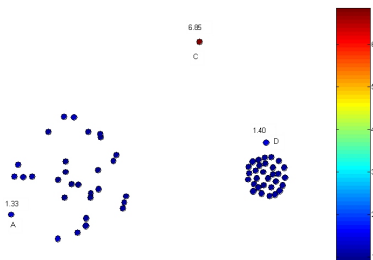
Ако су скупови различите густине тада може да се примени одређивање просечне релативне густине  $prg$

$$prg(x, k) = \frac{gustina(x, k)}{\sum_{y \in N(x, k)} gustina(y, k) / |N(x, k)|}$$

где је  $N(x, k)$  скуп који садржи  $k$  најближих суседа од  $x$ ,  $|N(x, k)|$  је величина тог скупа, а  $y$  је најближи сусед.

# Релативна густина према Кнн суседа

Употреба релативне густине омогућава боље одређивање аномалија у случају скупова са различитом густином





## Приступ заснован на кластерованју

Објекат је аномалија/елемент ван граница ако је очигледно да не припада ни једном кластеру

Објекат је елемент ван граница/аномалија

- Код метода кластерованња заснованих на прототиповима, ако није близу центру ни једног од кластера
- Код кластера заснованих на густини, ако је његова густина мала
- Код метода заснованих на графовима, ако није добро повезан

Проблем: Неке методе кластерованња формирају кластере са малим бројем елемената



# Особине метода заснованих на густини

- Једноставне су за примену
- Рачунарски захтевне -  $O(n^2)$
- Осетљиве на промене параметара
- Проблем са одређивањем густине у вишедимензионом простору
- Проблем у одређивању технике кластеровања и броја кластера