SPAMIA: filtrovanie spamu pomocou kvantitatívnych profilov a adaptívneho klastrovania

Marián Grendár, Jana Škutová, Vladimír Špitalský

Slovanet a.s., Záhradnícka 151, 821 08 Bratislava

Bez(a)Dis, PF UPJŠ, 9. 5. 2012

Dosiahnuté výsledky vznikli v rámci riešenia projektu Výskum efektivity algoritmov pre inteligentné rozpoznávanie nevyžiadanej elektronickej komunikácie, návrh teoretických modelov nových algoritmov a posúdenie ich účinnosti, ktorý je podporovaný Ministerstvom školstva, vedy, výskumu a športu SR v rámci poskytnutých stimulov pre výskum a vývoj zo štátneho rozpočtu v zmysle zákona č. 185/2009 Z.z. o stimuloch pre výskum a vývoj.

Obsah

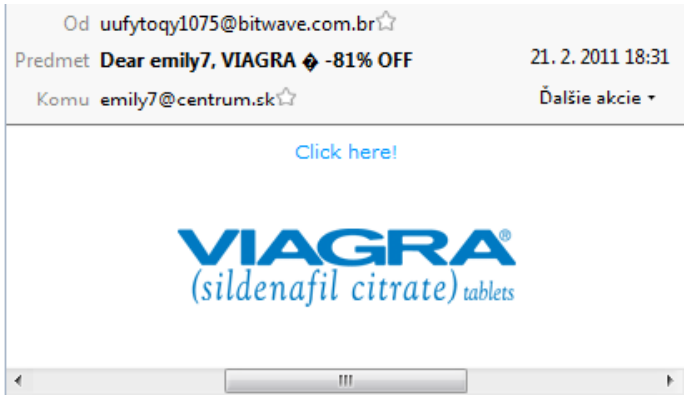
Problematika spamu

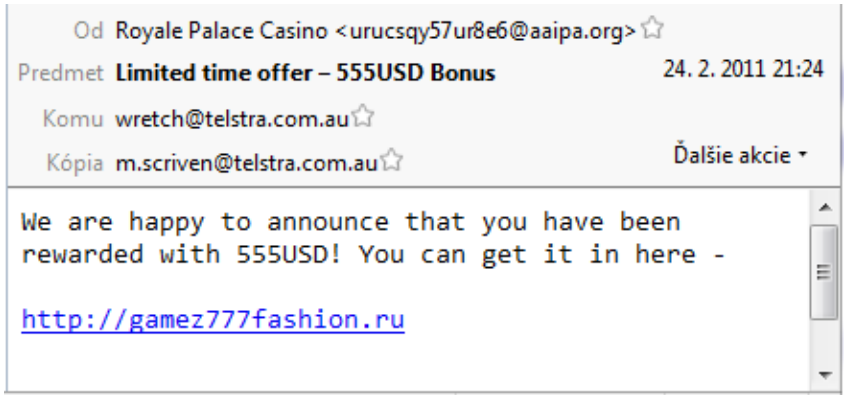
Spam

- je nevyžiadaná emailová správa často rozposielaná hromadne

Účel

- reklama, phishing, scam, vírusy, verifikácia emailu, . . .





Existujúce riešenia na filtrovanie spamu

Open-source produkty

SpamAssassin

Bogofilter

DSPAM

...

Komerčné produkty

Metódy

- blacklisty, whitelisty, greylisty, SPF, ...
- heuristické pravidlá
- textmining

Efektívnosť existujúcich riešení

Udávaná úspešnosť $> 99\%$

- Virus Bulletin, TREC, CEAS

G. Hulten, J. Goodman (Microsoft Research, 2004):
Junk e-mail filtering
"Why you won't get 99% in real life"

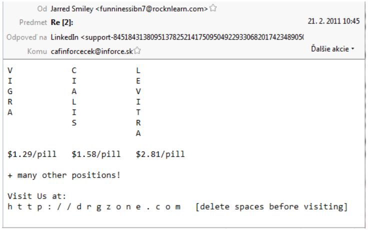
- testovacie korpusy sú nerealisticky jednoduché, ...

Výsledky Bogofiltra

- verejný korpus — viac ako 99.5% (1 chyba z 200)
- privátny korpus — cca. 95% (1 chyba z 20)

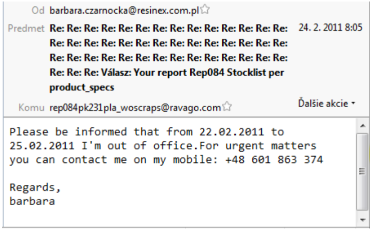
Nevýhody existujúcich riešení

- vysoká vulnerabilita



Od Jarred Smiley <funninessibn7@rocknlearn.com>
Predmet **Re [2]:** 21. 2. 2011 10:45
Odpoveď na LinkedIn <support-8451843138095137825214175095049229330682017423489050
Komu cafinforcecek@inforce.sk
Ďalšie akcie

```
V           C           L
I           I           E
G           A           V
R           L           I
A           I           T
            S           R
                        A

$1.29/pill  $1.58/pill  $2.81/pill

+ many other positions!

Visit Us at:
h t t p : / / d r g z o n e . c o m   [delete spaces before visiting]
```



Od barbara.czarnocka@resinex.com.pl
Predmet **Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Re: Válasz: Your report Rep084 Stocklist per product_specs** 24. 2. 2011 8:05
Komu rep084pk231pla_woscraps@ravago.com
Ďalšie akcie

```
Please be informed that from 22.02.2011 to
25.02.2011 I'm out of office.For urgent matters
you can contact me on my mobile: +48 601 863 374

Regards,
barbara
```

Nevýhody existujúcich riešení

- fixovanosť heuristických pravidiel

- závislosť na jazyku
- viazanosť na binárnu klasifikáciu
- rigidnosť výsledného hodnotenia spamovosti

Návrh nového algoritmu

Východiská návrhu nového algoritmu

Kvantitatívne profily

- vektor reálnych/celých čísel pevnej dimenzie
- zachytenie doposiaľ nevyužívaných charakteristík emailov

Adaptívne klastrovanie

- hierarchické rozčlenenie emailov na homogénne skupiny a zvyšok
- rozpoznanie emailových kampaní

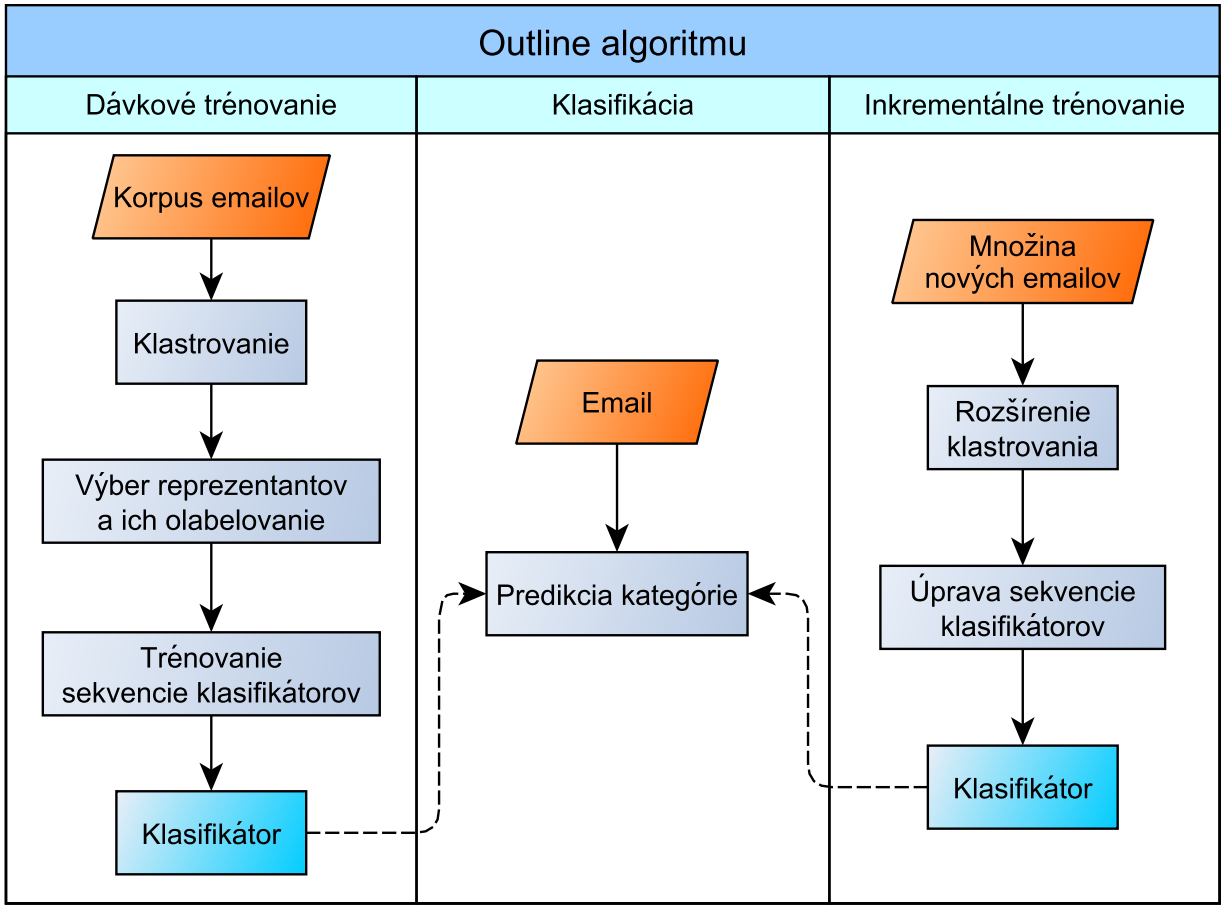
Sekvenčné klasifikovanie

- viacstupňová predikcia

Inkrementálne trénovanie

- rýchla a efektívna aktualizácia klastrovania a klasifikátorov

Outline algoritmu



Základné kvantitatívne profily

Binárny profil: vzdialenosti medzi výskytmi daného znaku/znakov

- LP riadkový: dĺžky riadkov
- SNP vetný: dĺžky viet
- WP slovný: dĺžky slov
- UP upper-case: vzdialenosti medzi veľkými písmenami
- STAP: vzdialenosti medzi výskytmi $*, \tilde{}, >$
- ...

Histogramový binárny profil:

- HWP: histogram dĺžok slov
- HUP: histogram vzdialeností medzi veľkými písmenami
- ...

Základné kvantitatívne profily

Znakový profil: početnosti znakov

- CP: ASCII profil

Zoskupený znakový profil: početnosti skupín znakov

- CPG9: čísla, medzery, zátvorky, operátory, separátory, veľké/malé písmená, nedovolené znaky, ostatné
- CPG11: z ostatných — samostatne ! a $

d-gramový zoskupený znakový profil:

- 2CPG9: dvojice skupín znakov
- 3CPG9: trojice skupín znakov
- ...

Základné kvantitatívne profily

Moving window profil: CPG za jednotlivé "časti" emailu

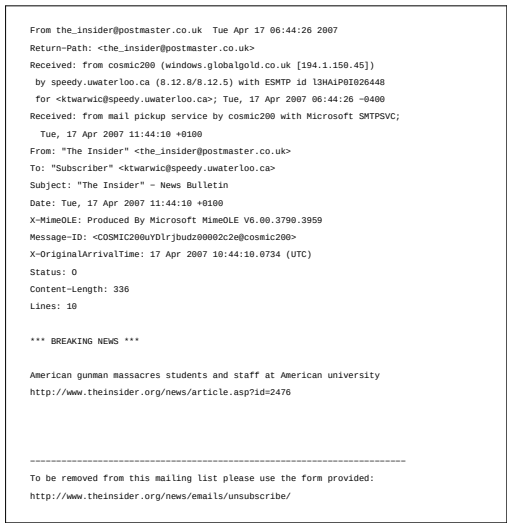
- MWP-CPG9
- MWP-CPG11

Veľkostný profil:

- veľkosť emailu
- veľkosti vybraných hlavičiek
- veľkosti častí mailu podľa content-type
- (voliteľne) CPG hlavičiek / častí

- SP
- SP-CPG9
- SP-CPG11

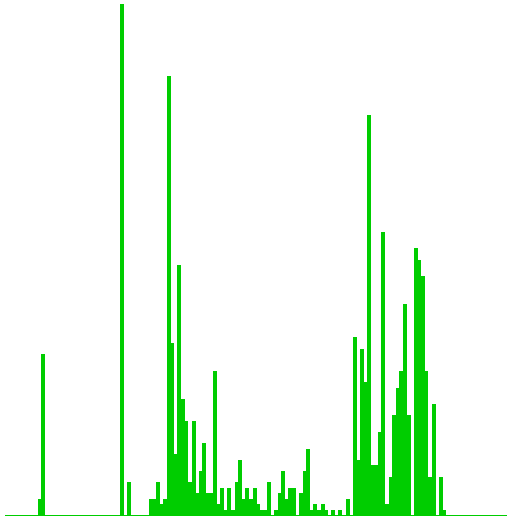
Základné kvantitatívne profily



```
From the_insider@postmaster.co.uk  Tue Apr 17 06:44:26 2007
Return-Path: <the_insider@postmaster.co.uk>
Received: from cosmic200 (windows.globalgold.co.uk [194.1.150.45])
 by speedy.uwaterloo.ca (8.12.8/8.12.5) with ESMTP id l3HAiP0I026448
 for <ktwarwic@speedy.uwaterloo.ca>; Tue, 17 Apr 2007 06:44:26 -0400
Received: from mail pickup service by cosmic200 with Microsoft SMTPSVC;
  Tue, 17 Apr 2007 11:44:10 +0100
From: "The Insider" <the_insider@postmaster.co.uk>
To: "Subscriber" <ktwarwic@speedy.uwaterloo.ca>
Subject: "The Insider" - News Bulletin
Date: Tue, 17 Apr 2007 11:44:10 +0100
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.3790.3959
Message-ID: <COSMIC200uYDlrjbudz00002c2e@cosmic200>
X-OriginalArrivalTime: 17 Apr 2007 10:44:10.0734 (UTC)
Status: O
Content-Length: 336
Lines: 10

*** BREAKING NEWS ***

American gunman massacres students and staff at American university
http://www.theinsider.org/news/article.asp?id=2476



--------------------------------------------------------------------------
To be removed from this mailing list please use the form provided:
http://www.theinsider.org/news/emails/unsubscribe/
```

(a) Email

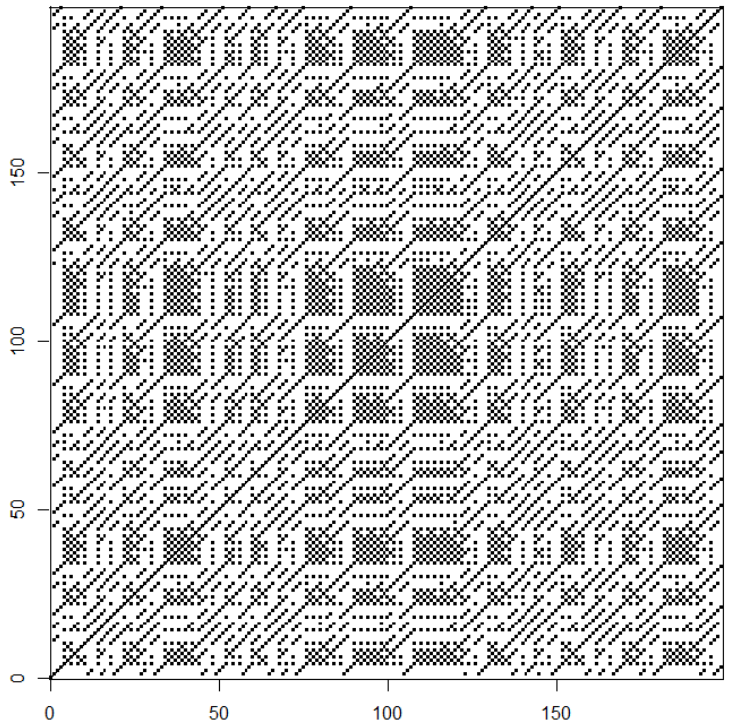
(b) Riadkový profil



(c) Znakový profil

Rekurenčné kvantitatívne profily

- založené na symbolicko-dynamickom prístupe k emailu
- charakteristiky získané z rekurenčnej kvantifikačnej analýzy



Adaptívne klastrovanie

Motivácia

- nutnosť tvorby a aktualizácie trénovacieho korpusu emailov
- permanentná potreba nových olabelovaných emailov

Adaptívne klastrovanie

- hierarchické rozčlenenie emailov na homogénne skupiny a zvyšok

Výhody

- redukcia korpusu emailov s minimálnou stratou informácie
- detekovanie hromadných kampaní (spamových, reklamných a iných)
- zohľadnenie informácie aj z nelabelovaných emailov
- efektívne využitie spätnej väzby o labeli emailu

Sekvenčné klasifikovanie

Účel

- zníženie časovej náročnosti klasifikovania nového emailu

Prostriedok

- najskôr jednoducho získateľné a nízko-dimenzionálne profily
- zložitejšie profily iba v prípade nejednoznačnej klasifikácie

Inkrementálne trénovanie

Účel

- aktualizácia klastrovania/klasifikátorov
- s nízkymi výpočtovými nárokmi

Predbežné meranie efektívnosti

Efektívnosť profilov

Privátny multilingválny korpus:

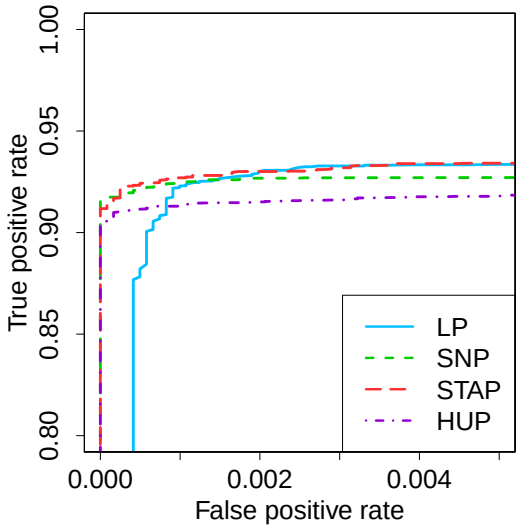
- train: Sep 2010 (11 050, 12% spam)
- test: Feb 2011 (17 248, 30% spam)

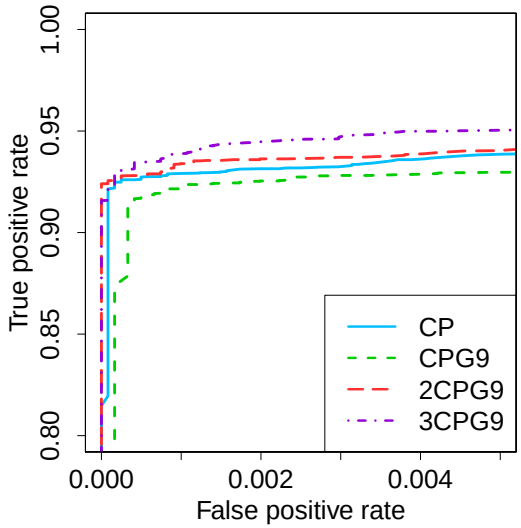
Chyba *fnr* (%) pri fixnom *fpr* = 0.1%

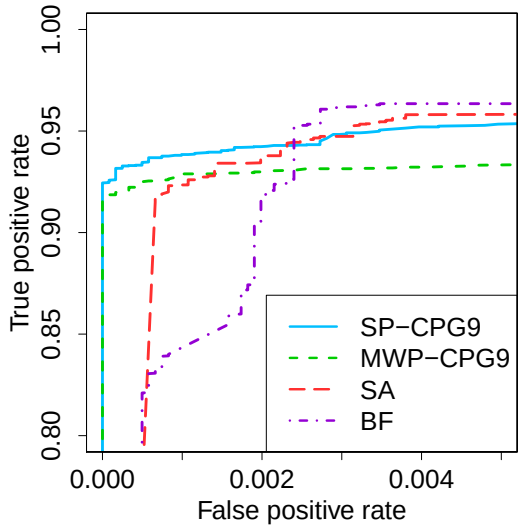
filter	*fnr*	filter	*fnr*
LP	7.7	CP	7.1
SNP	7.5	CPG9	7.8
STAP	7.3	2CPG9	6.6
HUP	8.7	3CPG9	6.1
SP-CPG9	6.2	MWP-CPG9	7.1
SpamAssassin	7.6	*Bogofilter*	15.6

Efektívnosť profilov

ROC krivky







Efektívnosť klastrovania

Korpus

- privátny (172 650, 30% spam)

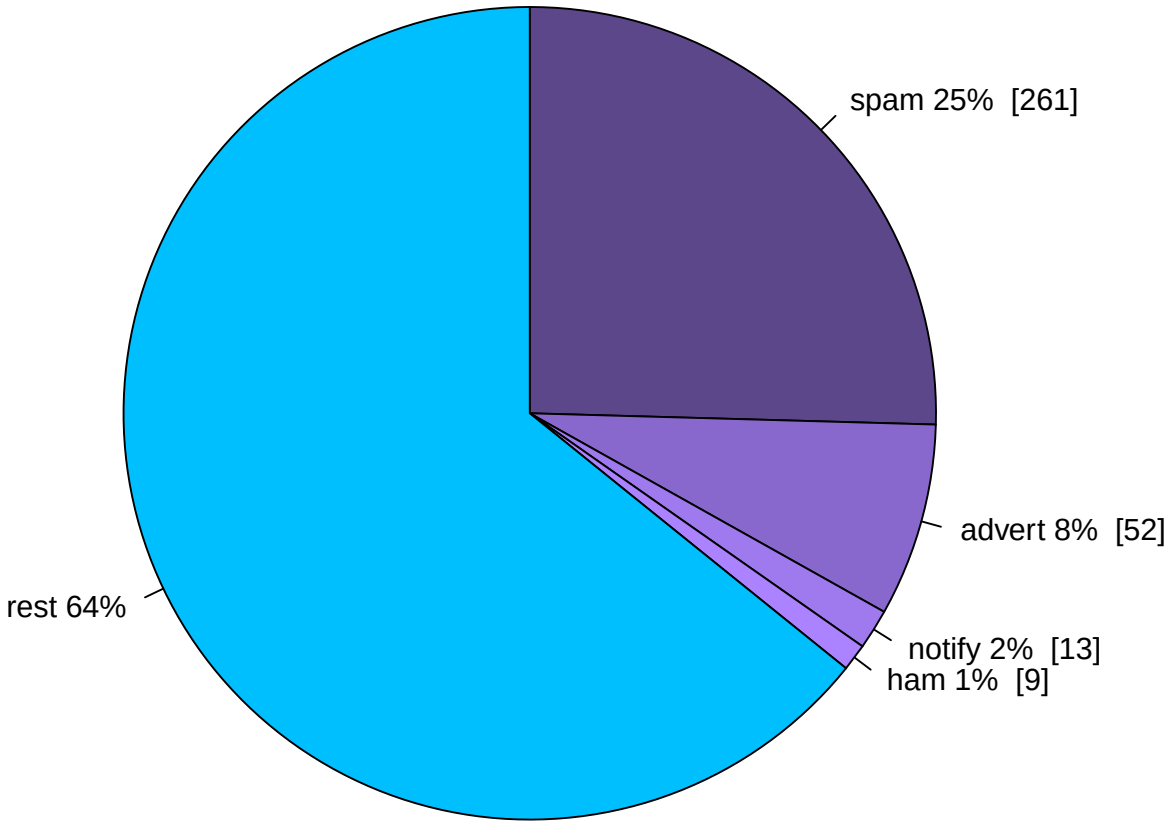
Adaptívne klastrovanie

- úroveň 1: entropia dĺžok riadkov (dim = 1)
- úroveň 2: riadkový profil (dim = 20)

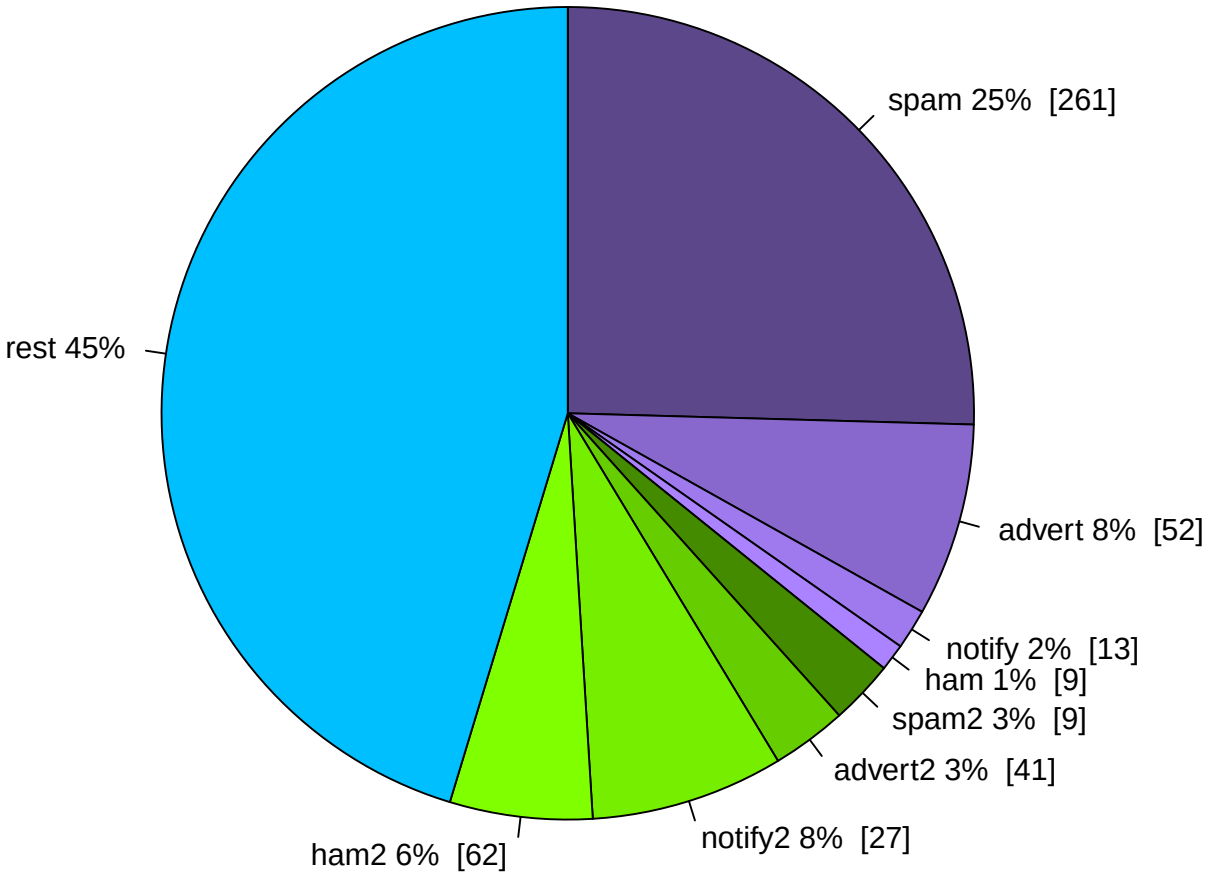
Chyba *fpr*, *fnr* (%) pri zaradení zvyšku do hamu

	Chyba		Klastre			
	fpr	fnr	spam	advert	notify	ham
úroveň 1	0.12	14.2	261	52	13	9
úroveň 2	0.26	5.7	270	93	40	71

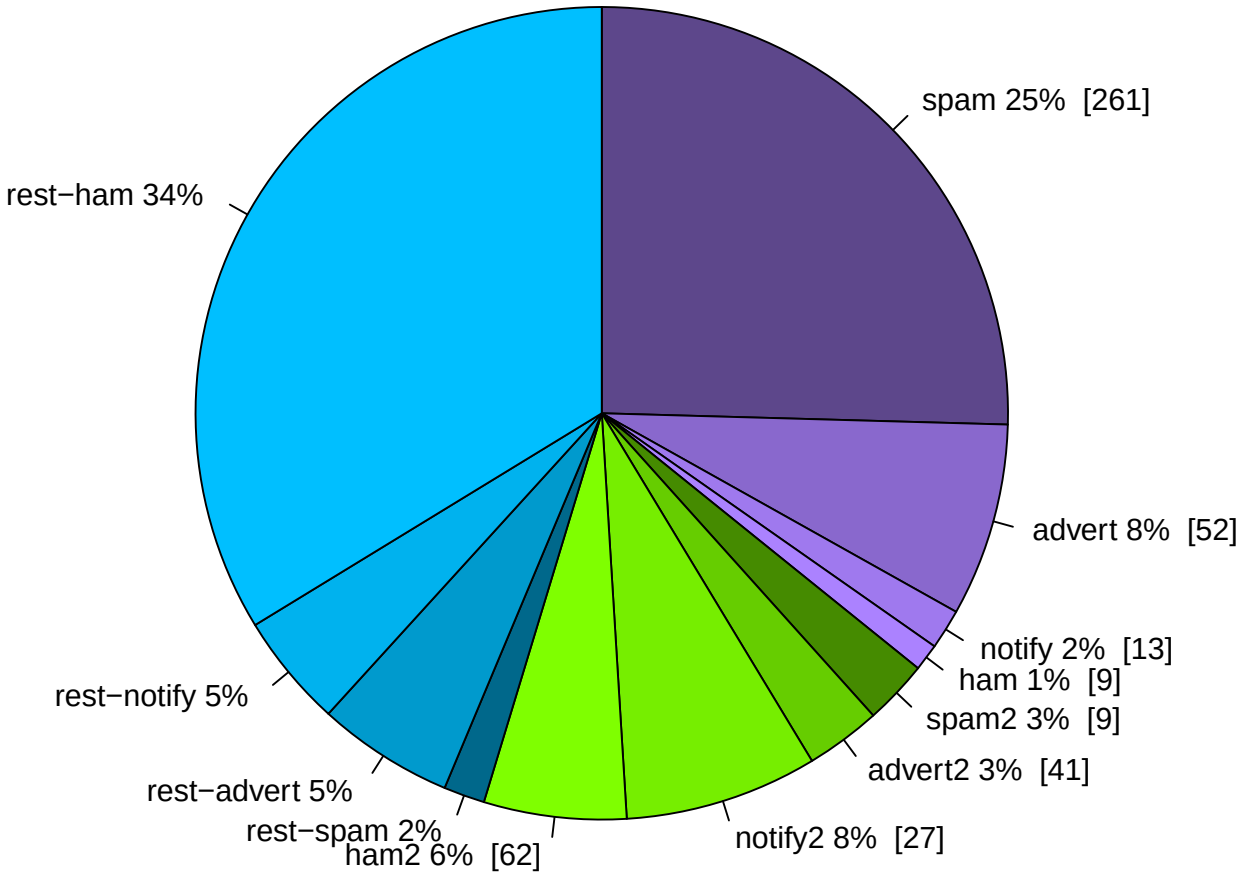
Efektívnosť klastrovania — úroveň 1



Efektívnosť klastrovania — úroveň 2



Efektívnosť klastrovania — úroveň 2



Zhrnutie

Hlavné charakteristiky nového algoritmu na filtrovanie spamu

- email je charakterizovaný kvantitatívnymi profilmi
- elementárne profily sa z hľadiska výkonnosti vyrovnajú alebo predčia heuristické pravidlá a Bayesov filter
- ďalšie výhody
  - škálovateľnosť, nízka výpočtová zložitosť, paralelizovateľnosť, robustnosť
  - nízka vulnerabilita
  - kombinovateľnosť s existujúcimi riešeniami
  - flexibilnosť a rozšíriteľnosť algoritmu
  - možnosť kategorizácie emailov
  - nezávislosť na jazyku
- efektívna tvorba a aktualizácia korpusu prostredníctvom adaptívneho klastrovania

Klastrovacie algoritmy

Klastrovacie algoritmy

Zmes normálnych rozdelení/EM-algoritmus

Hierarchické klastrovanie

- AGNES, DIANA, ...

Deliace klastrovanie

- K-Means, PAM, CLARA, ...

Klastrovanie založené na hustote

- DBSCAN, OPTICS, ...

a mnoho ďalších ...

Klastrovacie algoritmy

K-Means

- Forgy (1965), MacQueen (1967)

DBSCAN (Density-Based Spatial Clustering of Apps. with Noise)

- Ester, Kriegel, Sander, Xu (1996)

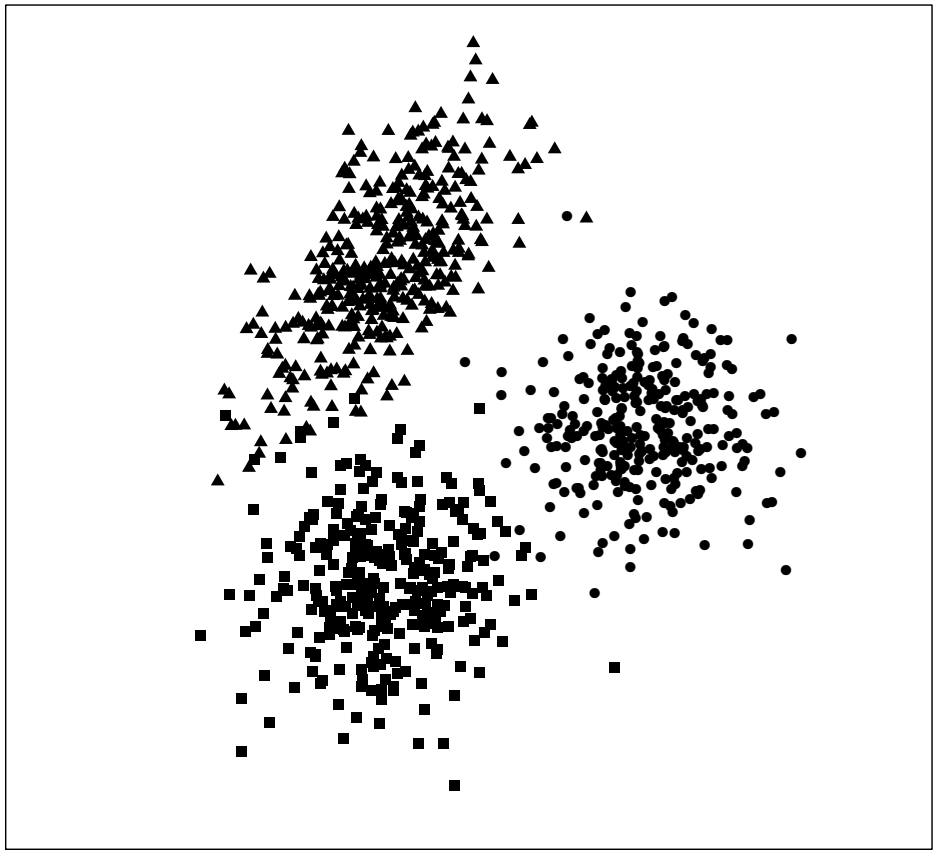
OPTICS (Ordering Points To Identify the Clustering Structure)

- Ankerst, Breuning, Kriegel, Sander (1999)

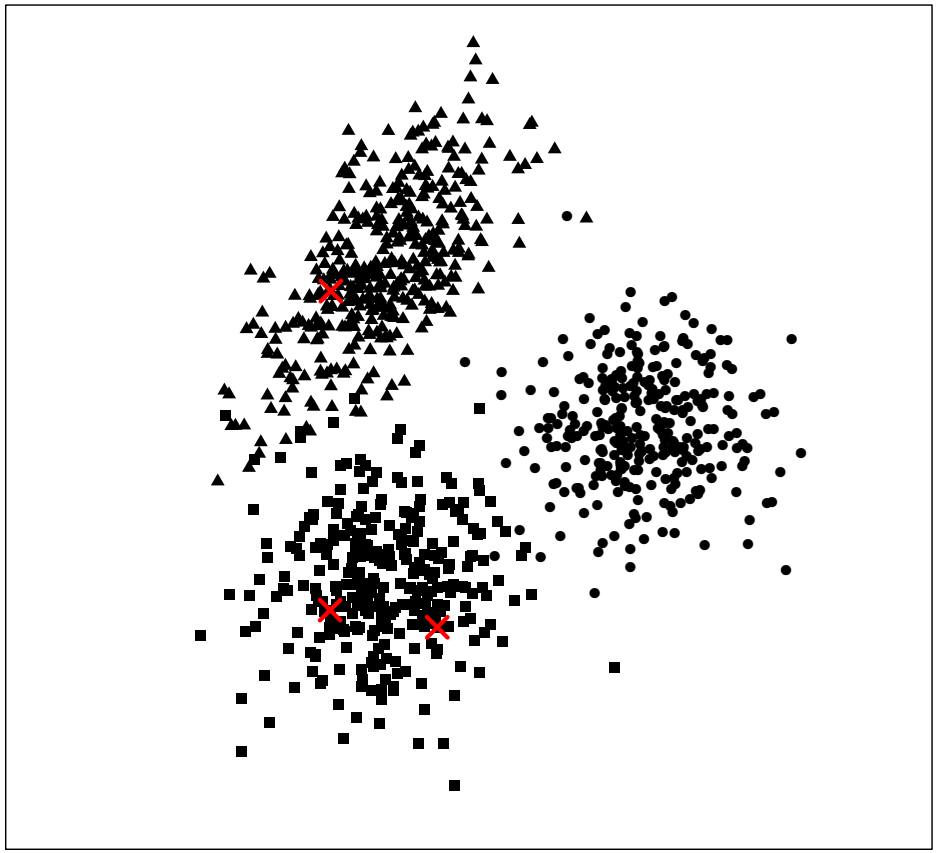
K-Means: algoritmus

1. určenie počtu hľadaných klastrov k
2. výber k náhodných objektov z množiny klastrovaných objektov — reprezentantov klastrov
3. priradenie všetkých objektov k najbližšiemu z vybraných k reprezentantov
4. určenie nových reprezentantov klastrov — sú nimi priemerné objekty (means) v jednotlivých klastroch
5. ak sa reprezentanti klastrov zmenia (vzhľadom k zvolenému kritériu), algoritmus pokračuje bodom 3.;
   ak sa reprezentanti nezmenia, algoritmus končí

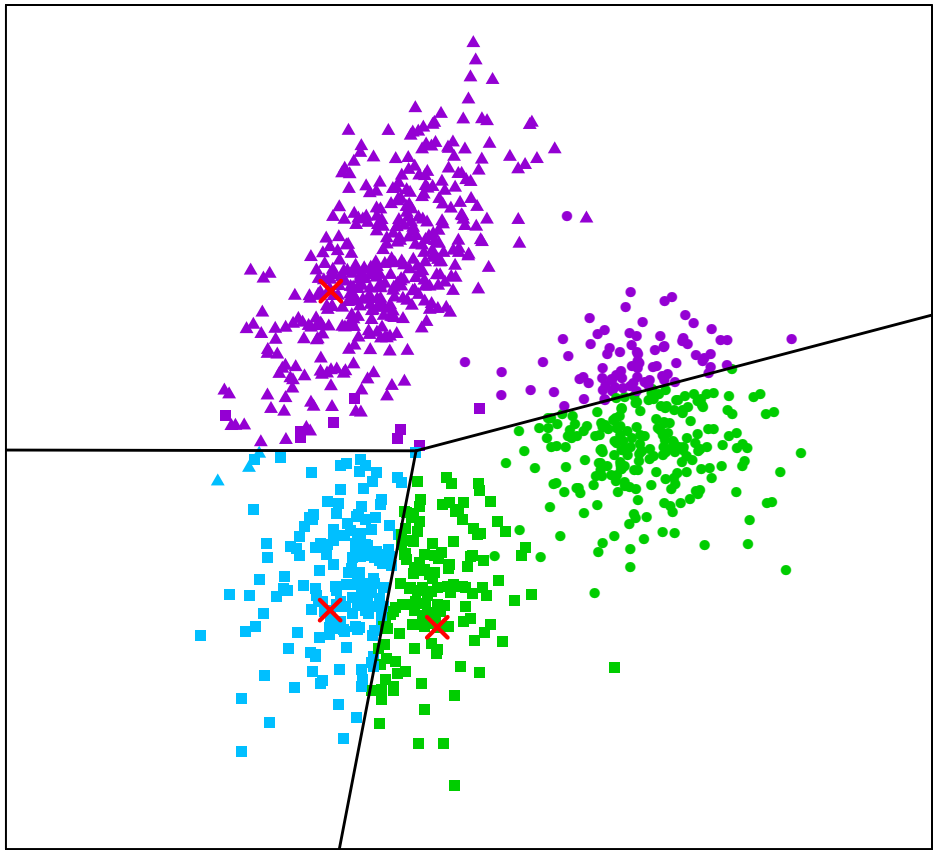
K-Means: množina objektov 1



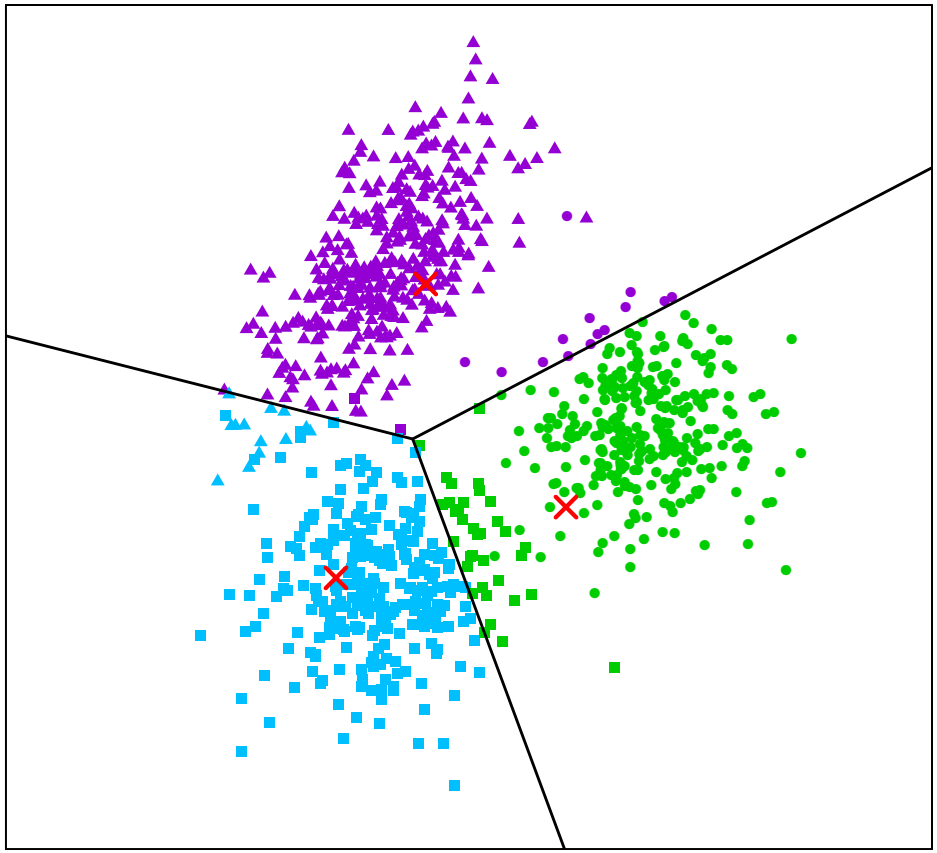
K-Means: prvý výber reprezentantov (náhodný)



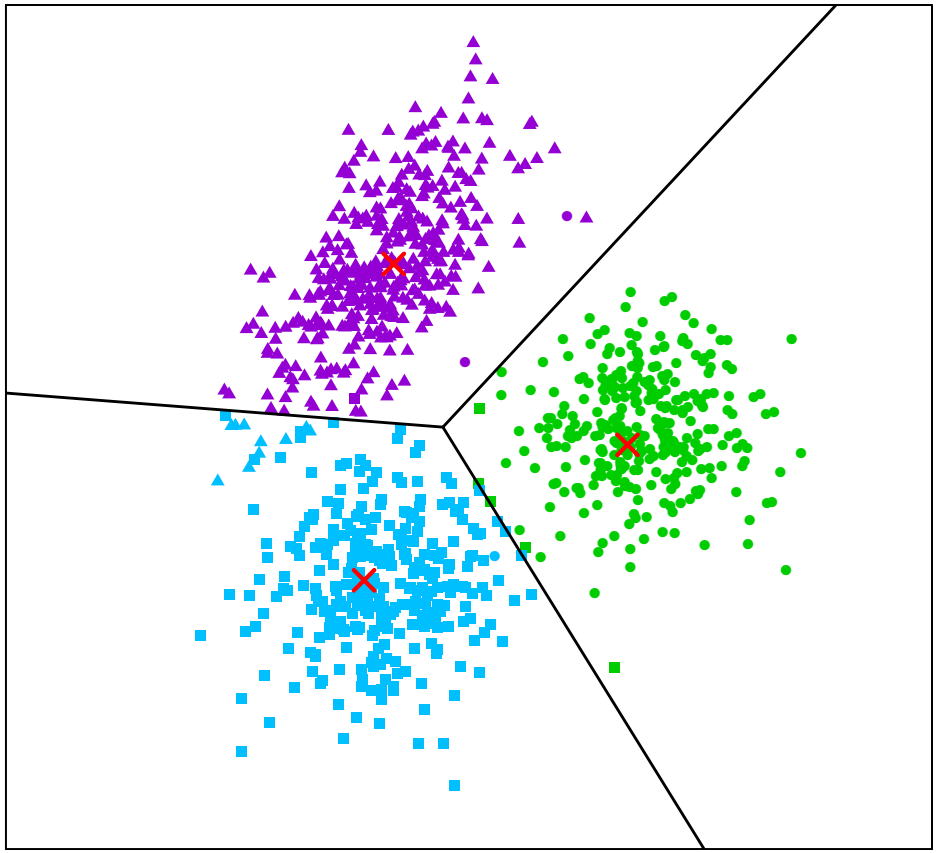
K-Means: priradenie objektov k reprezentantom



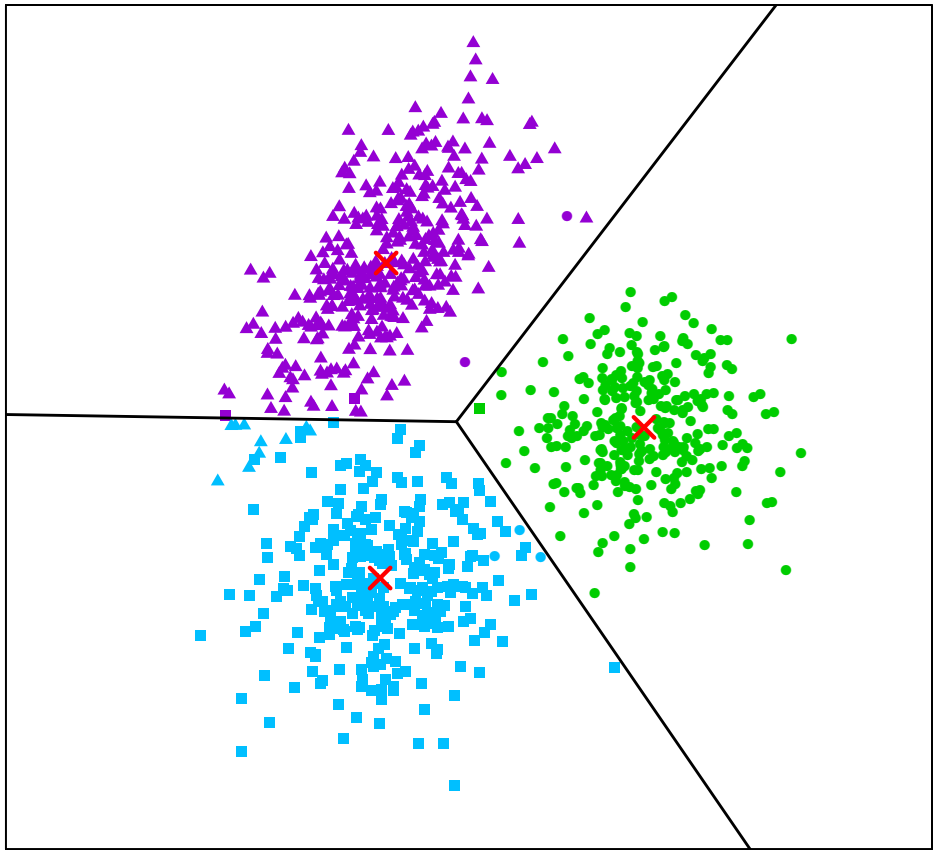
K-Means: druhá iterácia



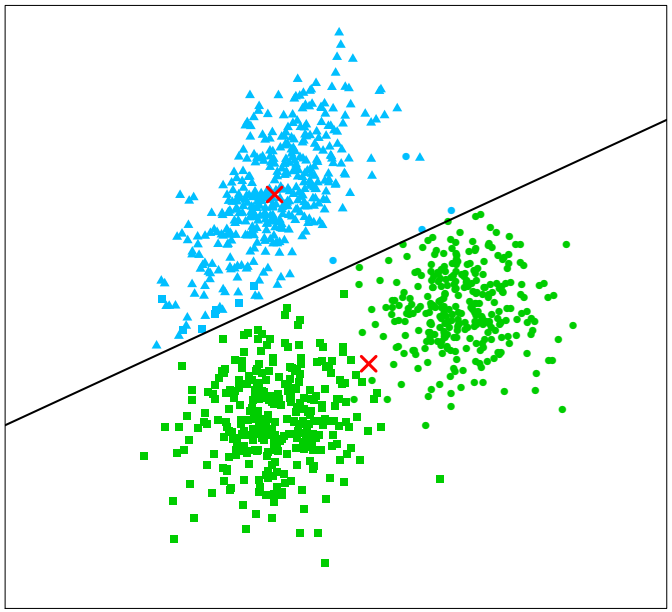
K-Means: tretia iterácia



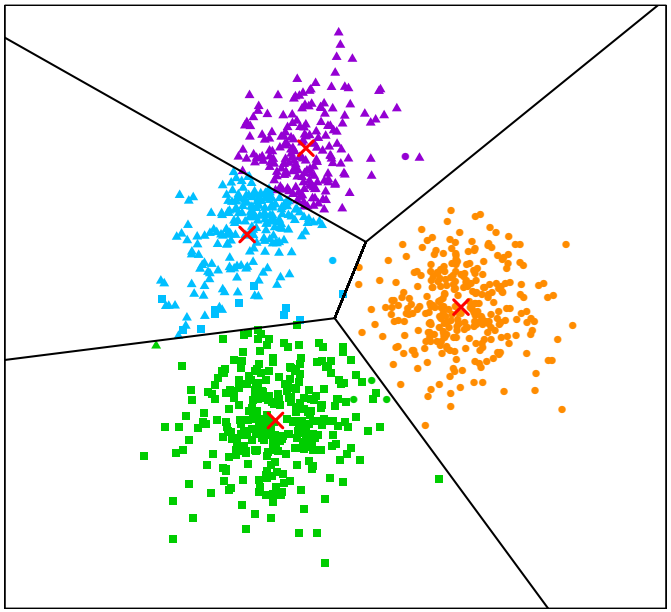
K-Means: výsledné klastrovanie



K-Means: nesprávne určený počet klastrov

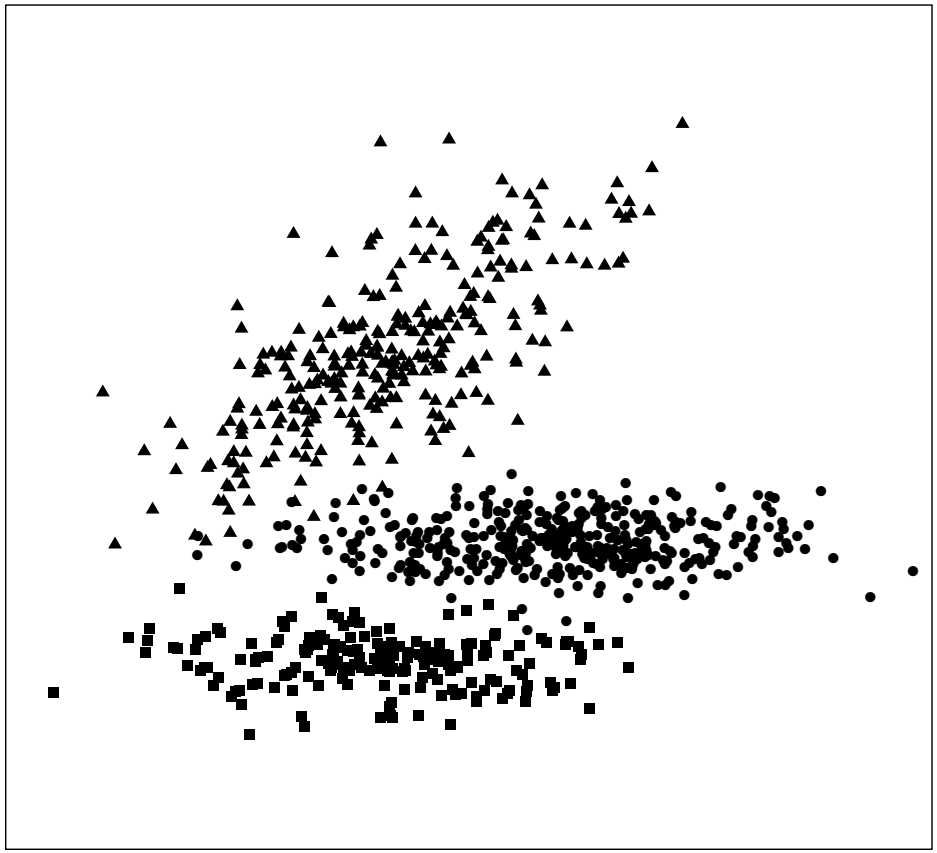


(a) $k = 2$

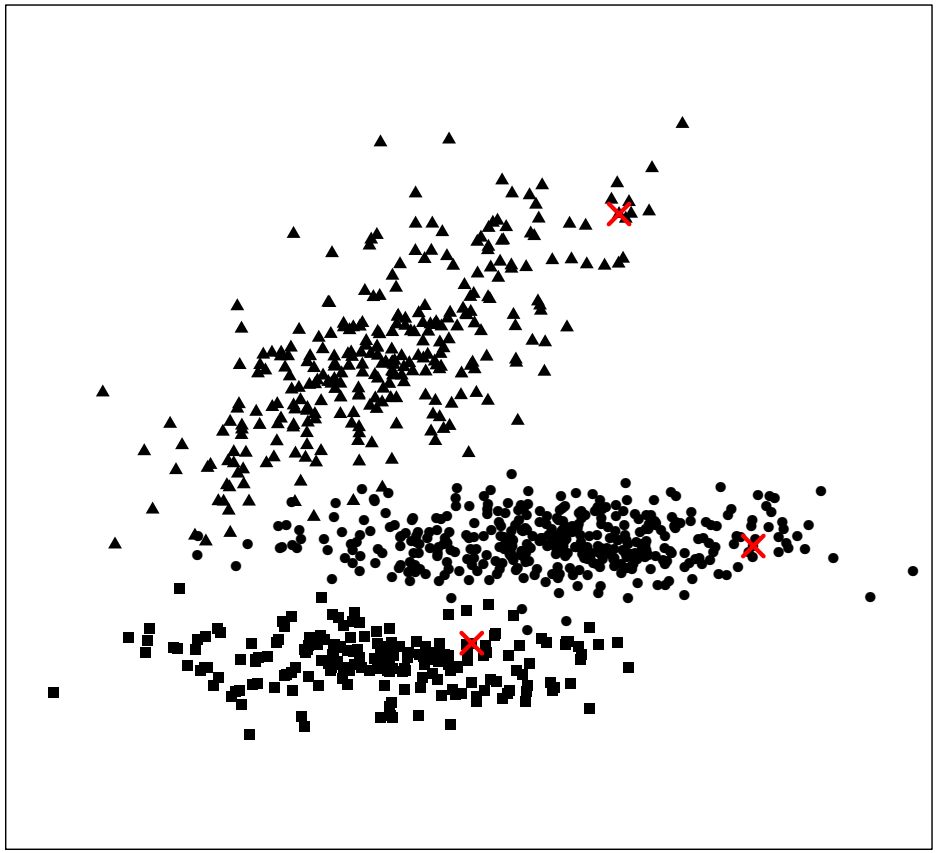


(b) $k = 4$

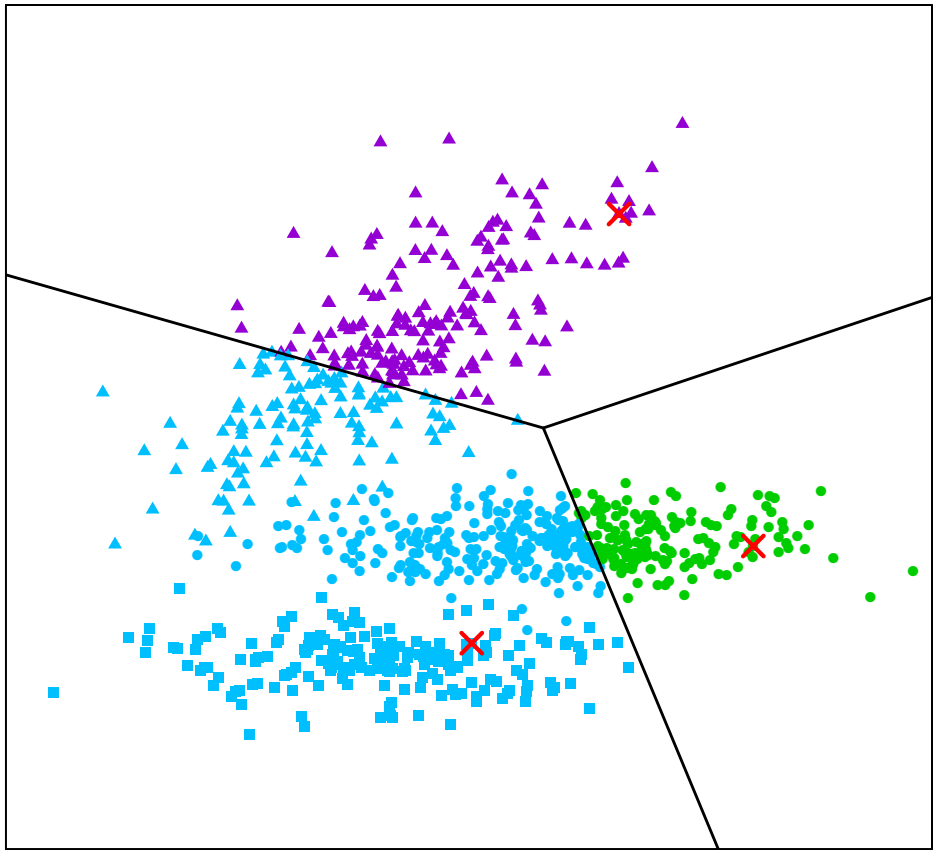
K-Means: množina objektov 2



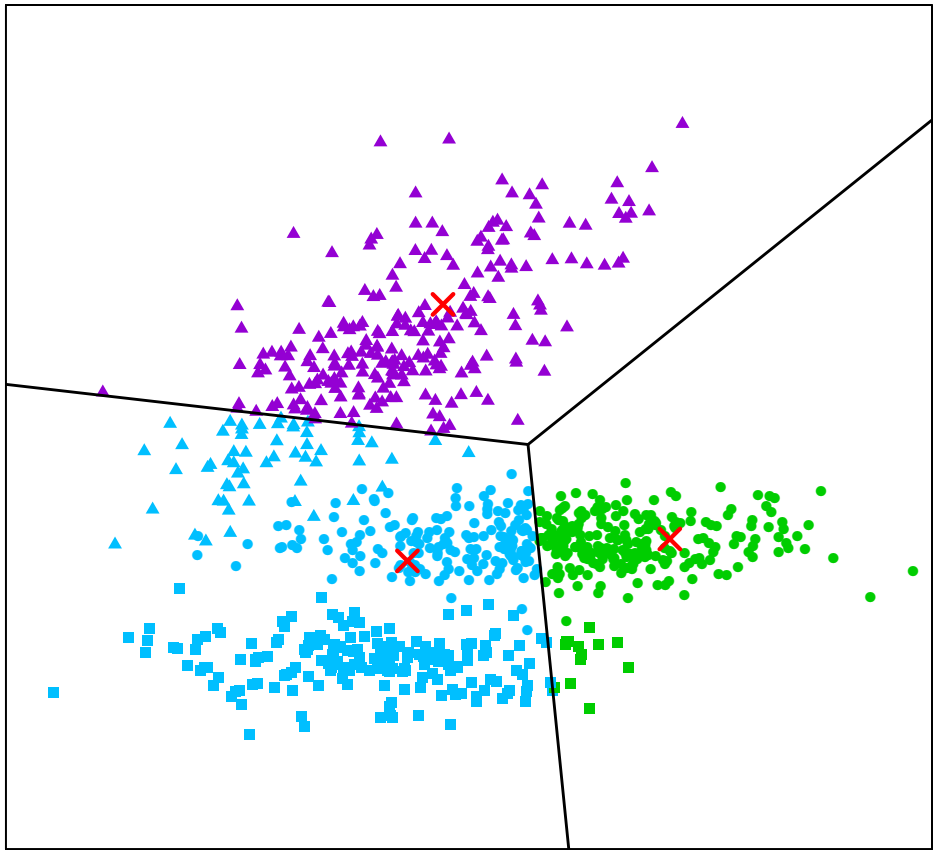
K-Means: prvý výber reprezentantov (náhodný)



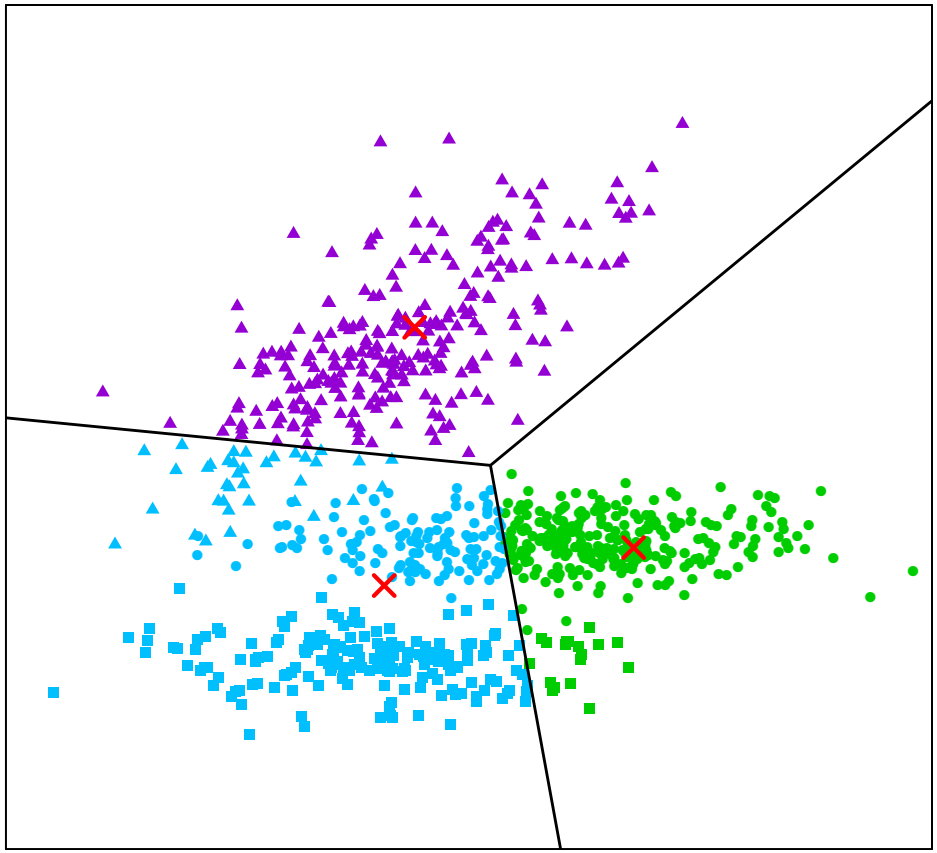
K-Means: priradenie objektov k reprezentantom



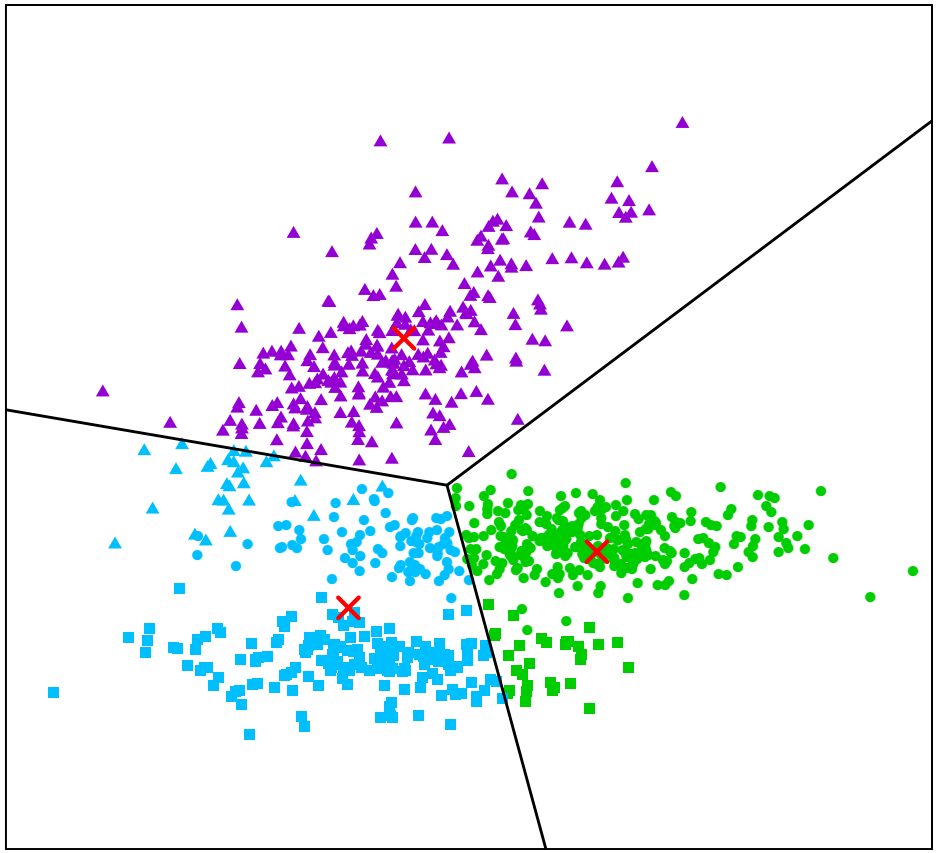
K-Means: druhá iterácia



K-Means: tretia iterácia



K-Means: výsledné klastrovanie



K-Means: výhody a nevýhody

+ jednoduchosť
+ schopnosť určiť klastre, ktoré od seba nie sú jednoznačne oddelené
+ rozklastrovanie všetkých objektov (diskutabilné)

x potreba vopred určiť počet klastrov k
x závislosť výsledného klastrovania na náhodnej voľbe prvých reprezentantov
x nerobustnosť voči odľahlým objektom
x neschopnosť určiť klastre rôznych tvarov

DBSCAN

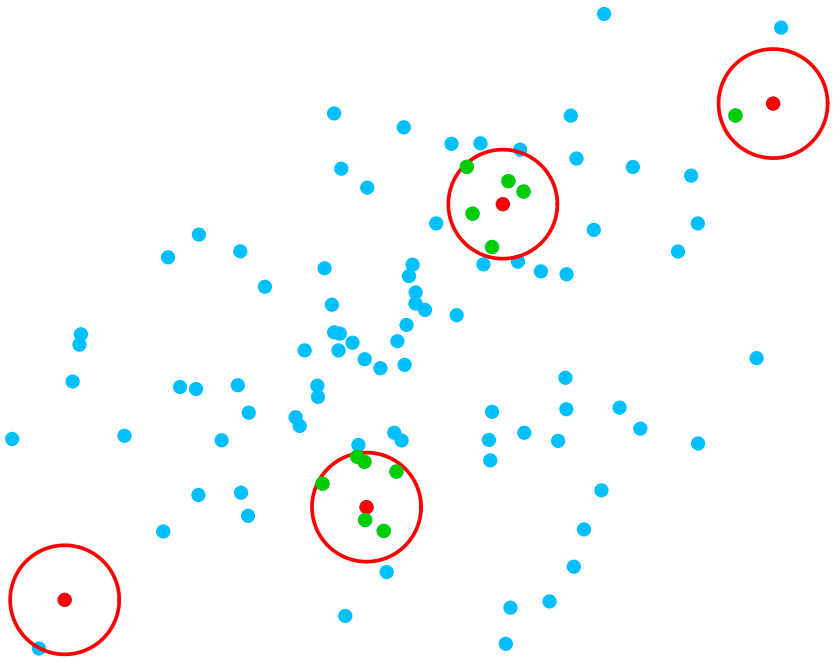
Density-Based Spatial Clustering of Apps. with Noise

- klastrovanie založené na hustote objektov
- nutnosť voľby dvoch parametrov:
  - vzdialenosti ε a
  - minimálneho počtu objektov N_{min}

DBSCAN: ε-okolie objektu

Definícia

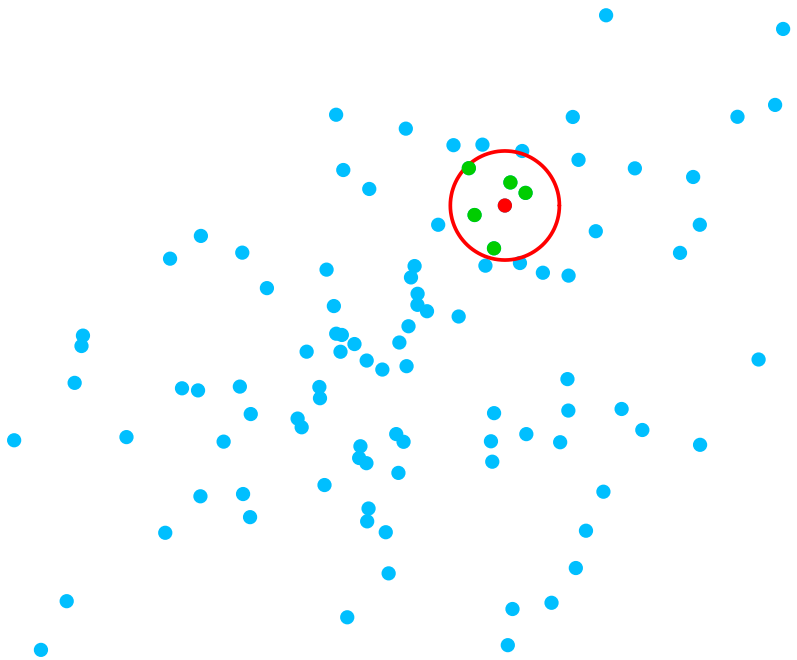
Nech $\mathcal{O}$ je množina objektov. Množina
$N_\varepsilon(O) = \{O_i \in \mathcal{O} : d(O, O_i) \leq \varepsilon\}$ sa nazýva ε-okolie objektu
$O \in \mathcal{O}$, $\forall i = 1, 2, \ldots, n$.



DBSCAN: vnútorný objekt

Definícia

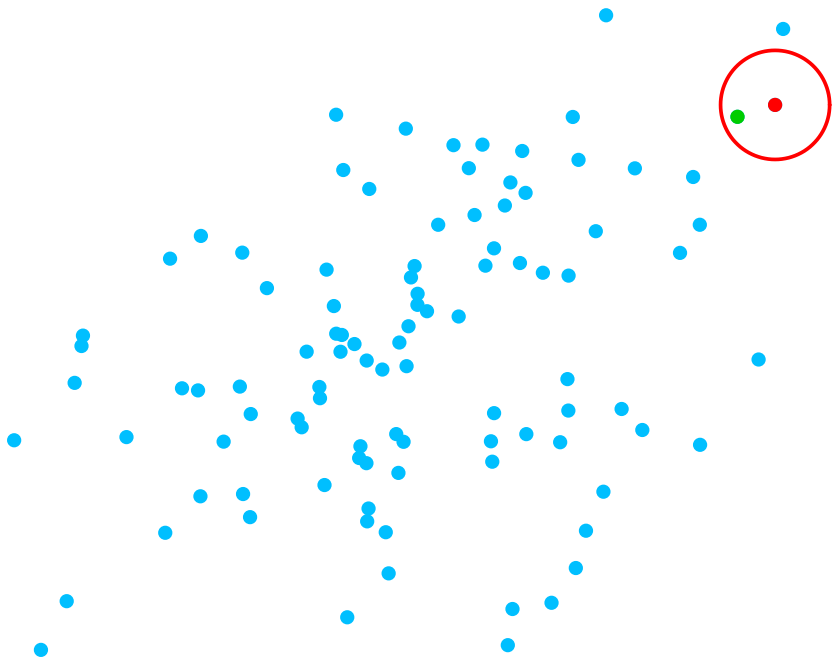
Objekt O sa nazýva vnútorný objekt (core point), ak počet objektov v jeho ε-okolí $N_\varepsilon(O)$ je aspoň N_{min}, teda $|N_\varepsilon(O)| \geq N_{min}$.



DBSCAN: vonkajší objekt

Definícia

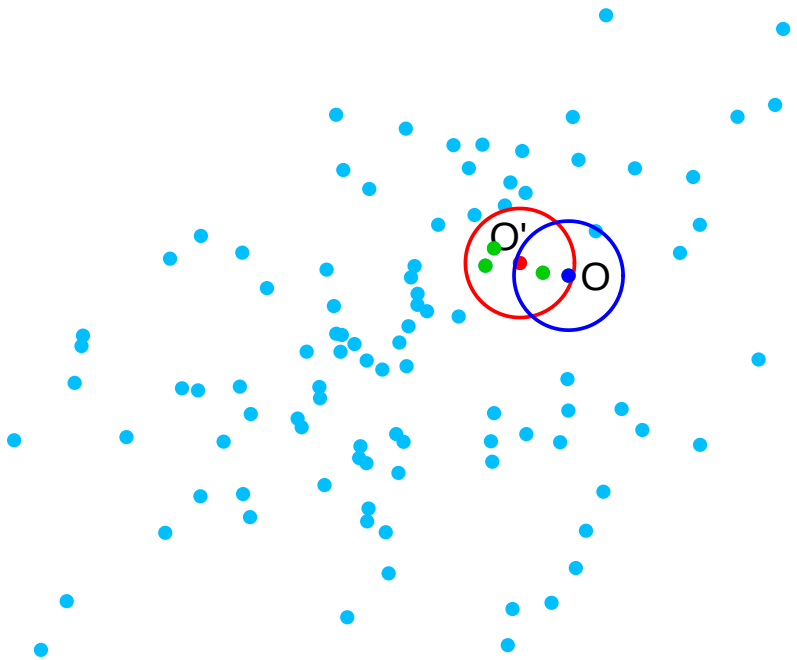
Objekt O sa nazýva vonkajší objekt, ak počet objektov v jeho ε-okolí $N_\varepsilon(O)$ je menej ako N_{min}, teda $|N_\varepsilon(O)| < N_{min}$.



DBSCAN: priamo dostupný objekt

Definícia

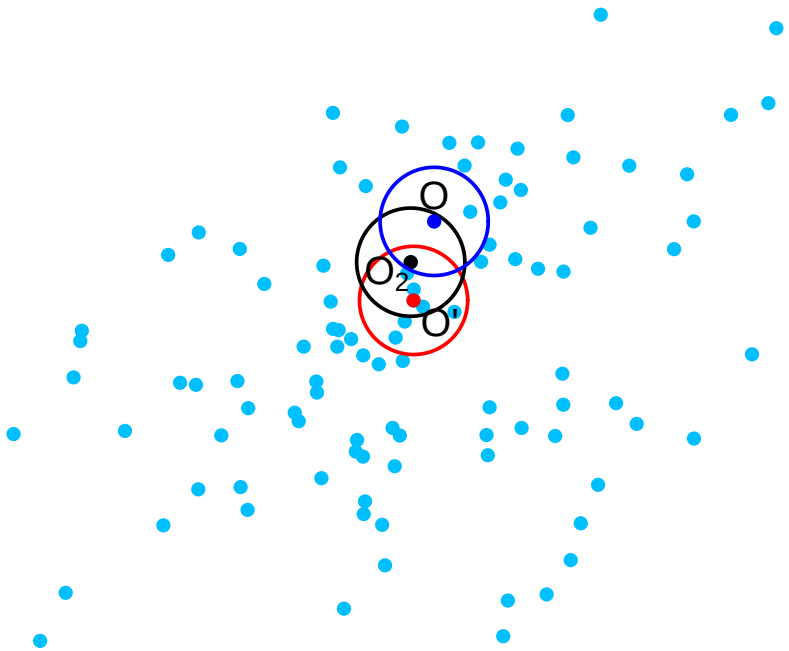
Objekt O je priamo dostupný z objektu O' vzhľadom k danému ε a N_{min}, ak $O \in N_\varepsilon(O')$ a zároveň $|N_\varepsilon(O')| \geq N_{min}$.



DBSCAN: dostupný objekt

Definícia

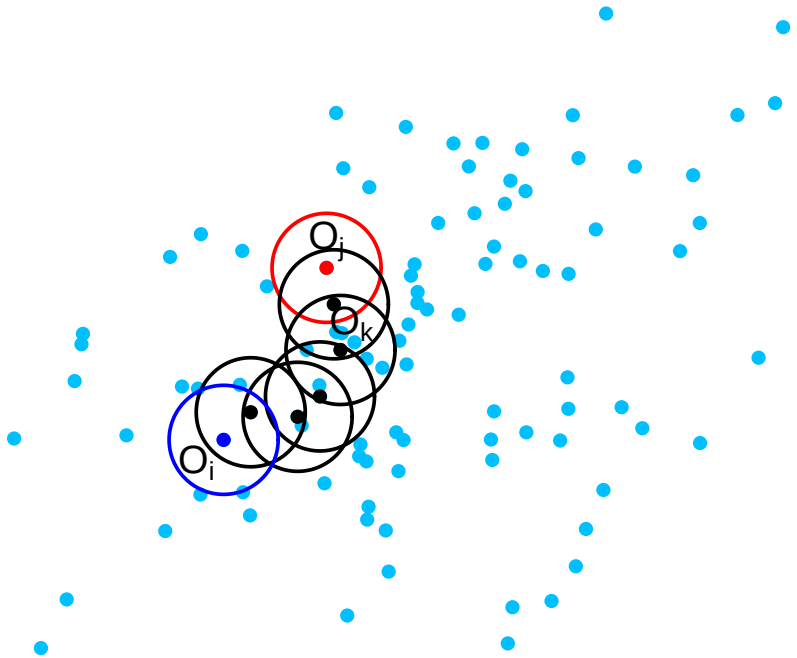
Objekt O je dostupný z objektu O' vzhľadom k danému ε a N_{min}, ak pre danú postupnosť objektov $O_1, \ldots, O_n$, pričom $O_1 = O'$ a $O_n = O$, platí, že objekt O_{k+1} je priamo dostupný z objektu O_k.



DBSCAN: spojené objekty

Definícia

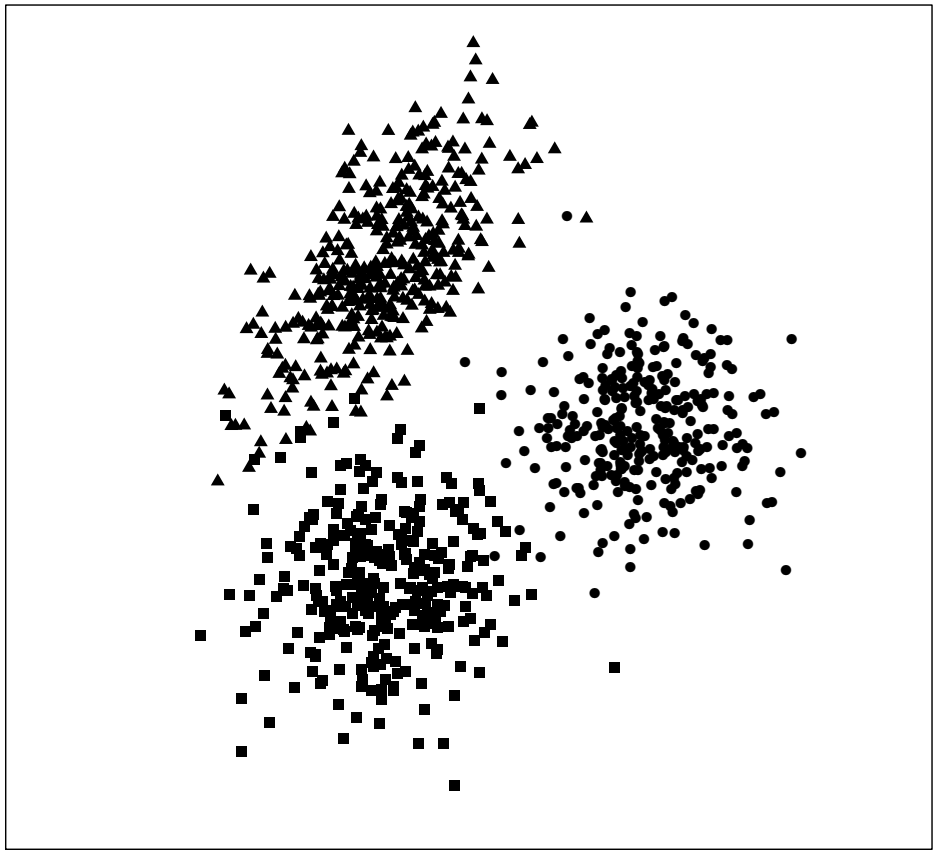
Objekt O_i je spojený s objektom O_j vzhľadom k danému ε a N_{min}, ak existuje objekt O_k taký, že objekty O_i a O_j sú dostupné z O_k vzhľadom k danému ε a N_{min}.



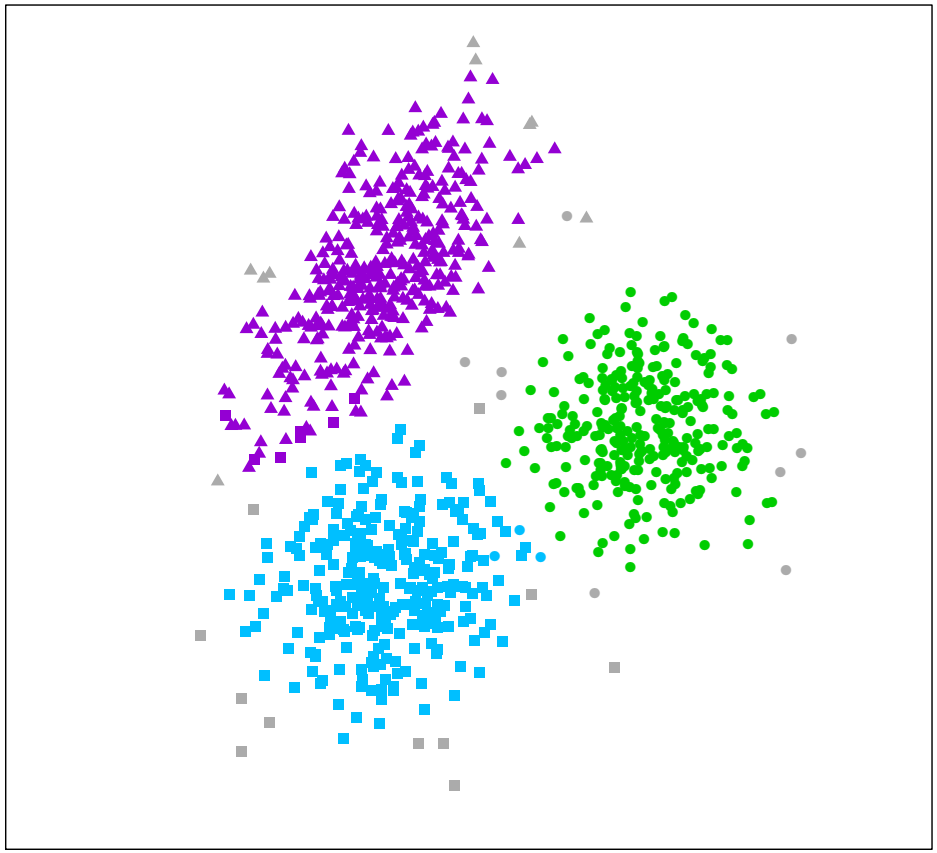
DBSCAN: algoritmus

1. náhodný výber jedného objektu O z množiny nezaradených objektov (na začiatku je to množina všetkých objektov)
2. určenie počtu objektov v ε-okolí vybraného objektu O a
   - ak $|N_\varepsilon(O)| < N_{min}$, tak je daný objekt zaradený do šumu;
   - ak $|N_\varepsilon(O)| \geq N_{min}$, tak je vytvorený nový klaster, do ktorého patria všetky objekty, ktoré sú dostupné z daného objektu;
3. algoritmus končí v prípade, že sú všetky objekty zaradené do niektorého z klastrov alebo do šumu;
   ak ešte ostali nezaradené objekty, algoritmus pokračuje bodom 1.

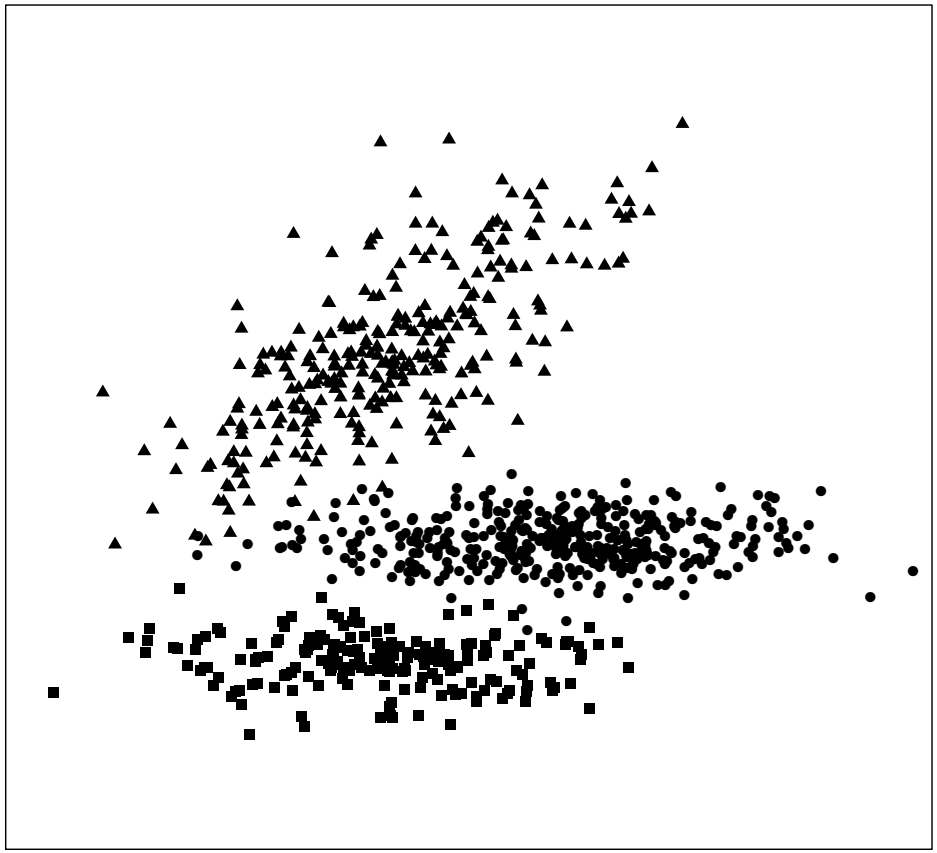
DBSCAN: množina objektov 1



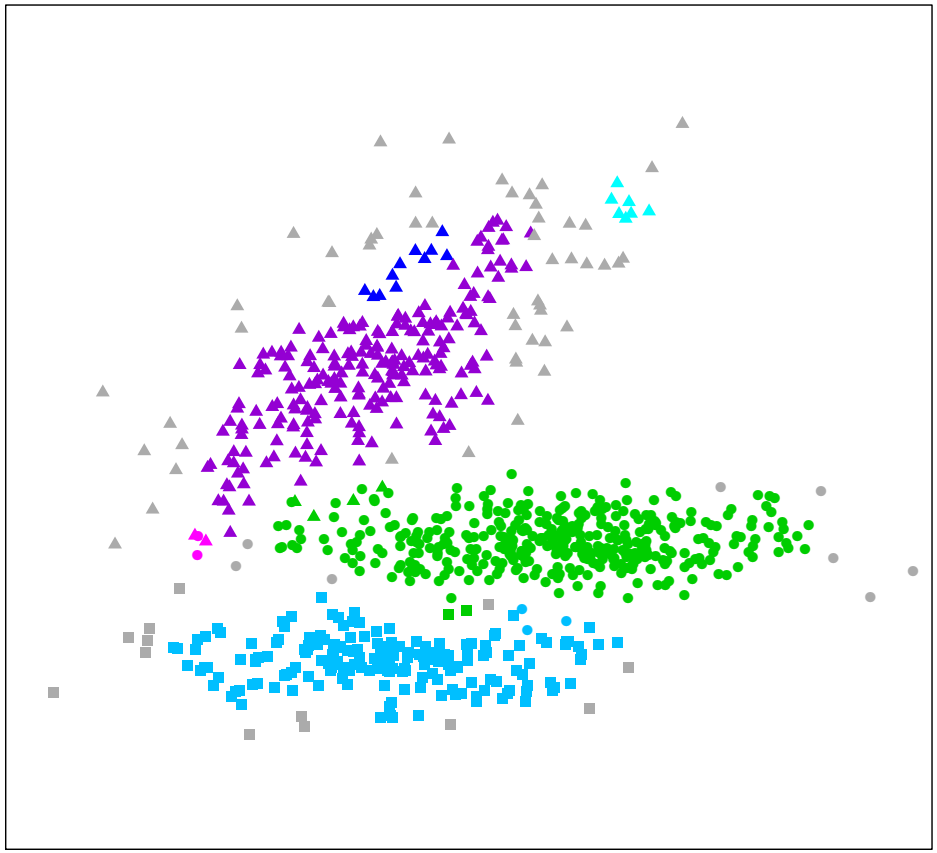
DBSCAN: výsledné klastrovanie



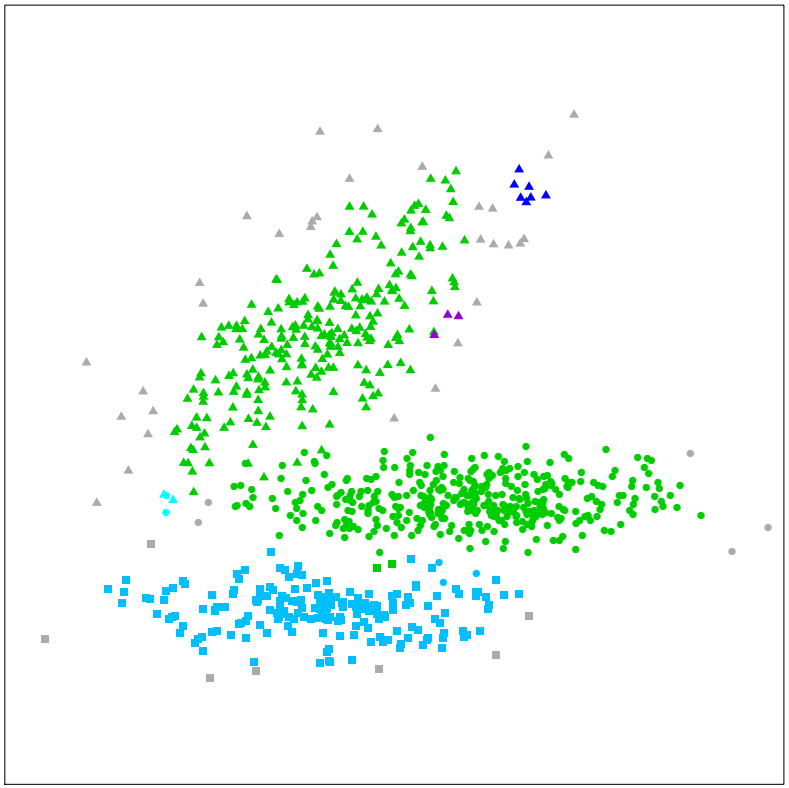
DBSCAN: množina objektov 2



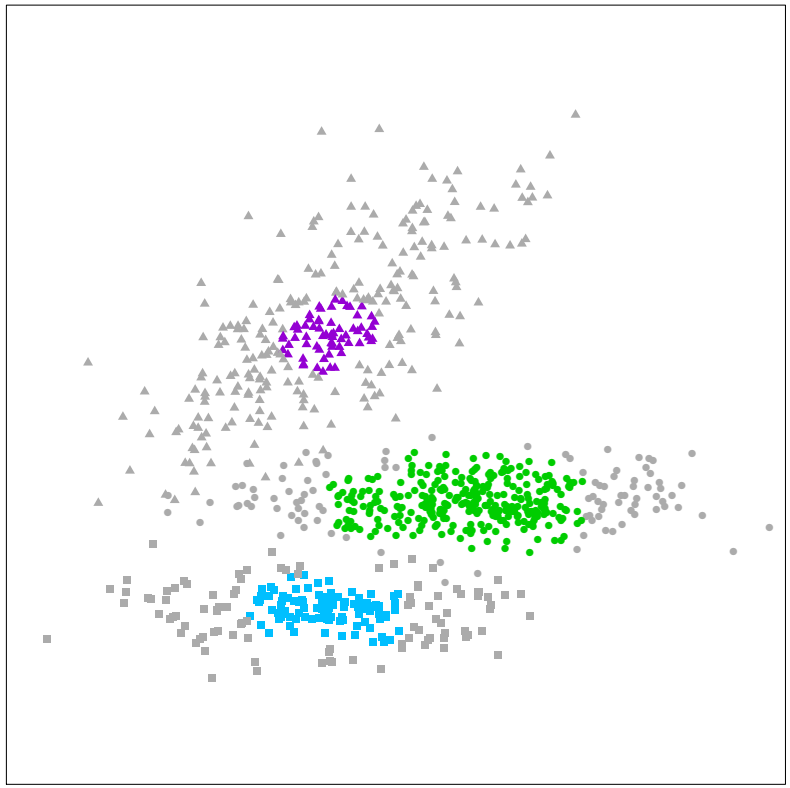
DBSCAN: výsledné klastrovanie



DBSCAN: nevhodne zvolené parametre

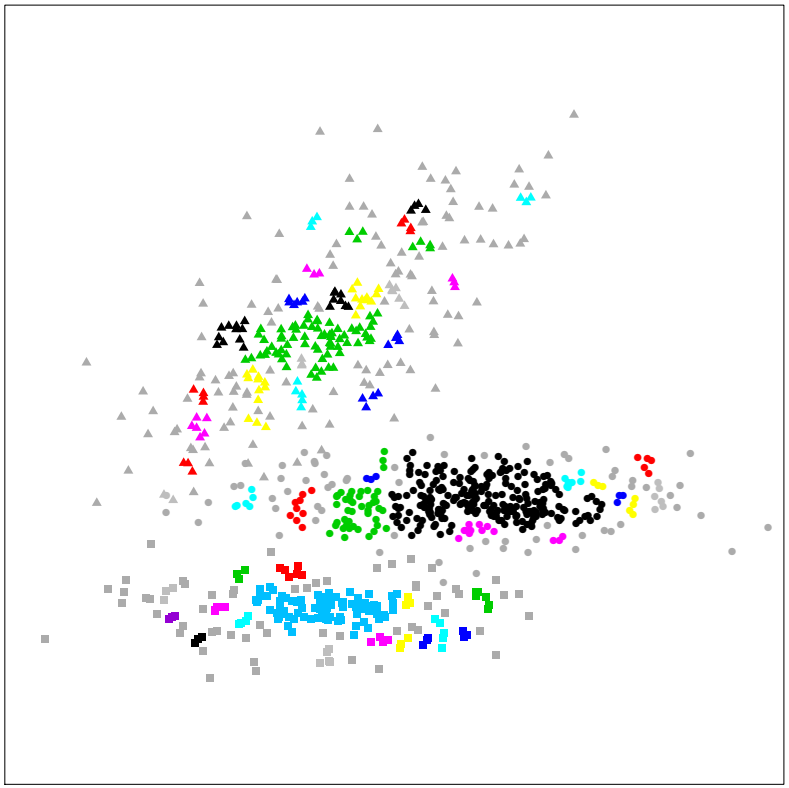


(a) príliš veľké ε a nízke N_{min}

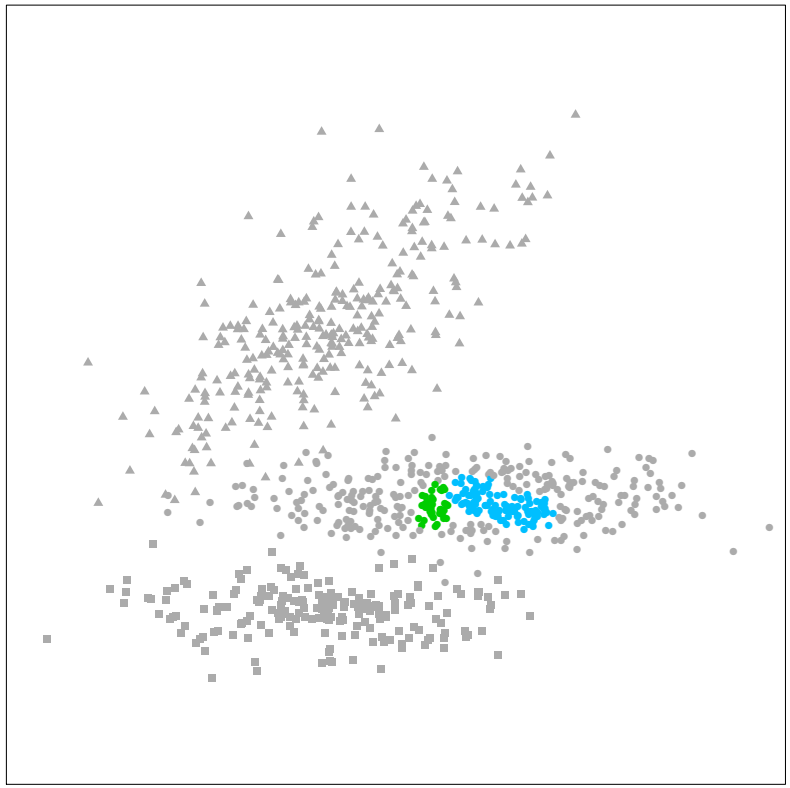


(b) príliš veľké ε a vysoké N_{min}

DBSCAN: nevhodne zvolené parametre



(c) príliš malé ε a nízke N_{min}



(d) príliš malé ε a vysoké N_{min}

DBSCAN - výhody a nevýhody

+ určenie klastrov rôznych tvarov
+ robustnosť voči odľahlým objektom; algoritmus vyčlenení objekty, ktoré nepatria do žiadneho klastra - označí ich za šum
+ nie je potrebné vopred určiť počet klastrov

x voľba parametrov ε a N_{min}
x viazanosť na fixné ε a N_{min}, teda neschopnosť zachytiť klastre rôznej hustoty
x neschopnosť určiť klastre, ktoré od seba nie sú jednoznačne oddelené

OPTICS

Ordering Points To Identify the Clustering Structure

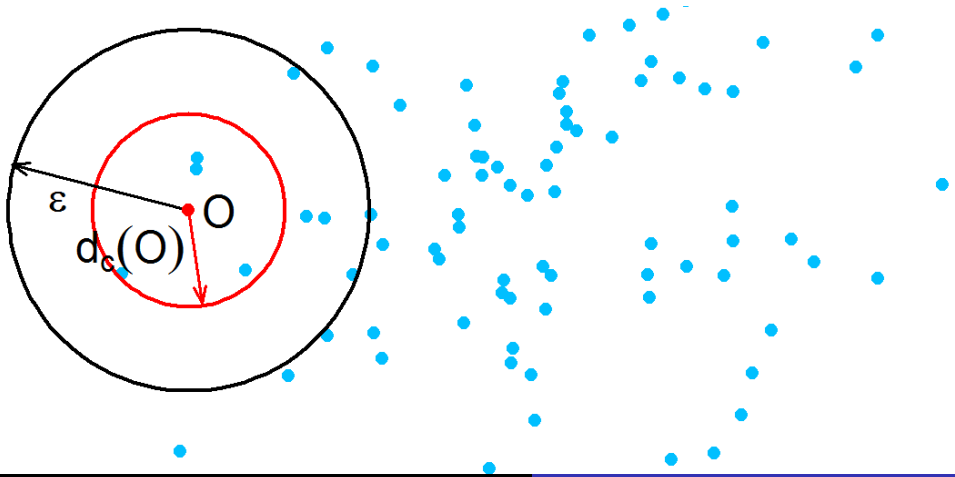
- usporiadanie objektov založené na ich hustote s následnou možnosťou určiť z tohto usporiadania klastrovanie
- nutná voľba maximálneho ε a minimálneho počtu objektov N_{min}

OPTICS: core vzdialenosť objektu

Definícia

Nech N_{min}-vzdialenosť $d_m(O)$ objektu O je vzdialenosť objektu O od jeho N_{min}-tého najbližšieho suseda. Potom core vzdialenosť $d_c(O)$ objektu O je pri danom ε a N_{min} definovaná rovnosťou

$$d_c(O) = \begin{cases} \text{nedefinovaná}, & \text{ak } |N_\varepsilon(O)| < N_{min} \\ d_m(O), & \text{inak.} \end{cases}$$

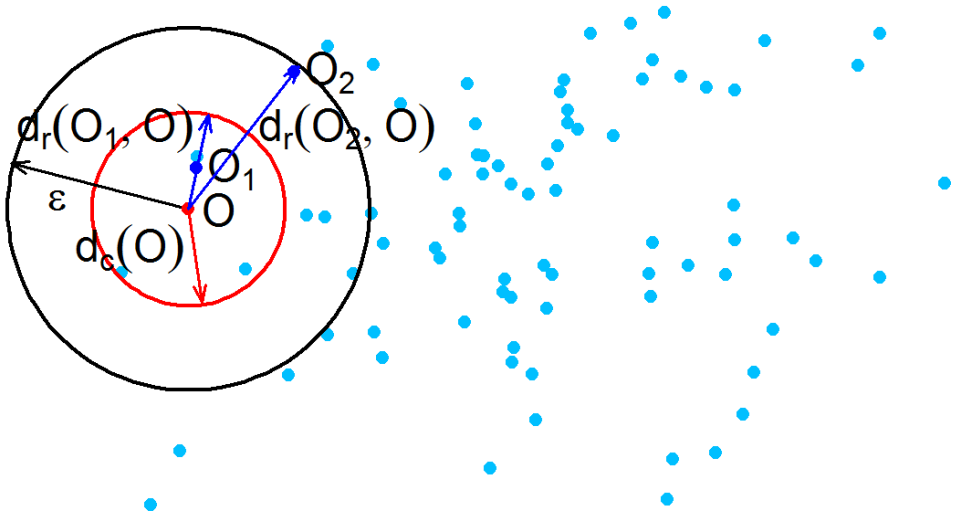


OPTICS: dostupnostná (reachability) vzdialenosť objektu

Definícia

Dostupnostná vzdialenosť $d_r(O_i, O)$ objektu O_i od objektu O je pri danom ε a N_{min} definovaná rovnosťou

$$d_r(O_i, O) = \begin{cases} \text{nedefinovaná}, & \text{ak } |N_\varepsilon(O)| < N_{min} \\ \max(d_c(O), d(O_i, O)), & \text{inak.} \end{cases}$$



OPTICS: algoritmus 1/3

1. výber objektu O z množiny neusporiadaných objektov (na začiatku je to množina všetkých objektov)
2. určenie objektov v ε-okolí objektu O - susedov
3. priradenie core-vzdialenosti a nedefinovanej dostupnostnej vzdialenosti objektu O
4. zápis objektu O do finálneho usporiadania a jeho vylúčenie z množiny neusporiadaných objektov
5. - ak má objekt O nedefinovanú core vzdialenosť, algoritmus pokračuje bodom 1.;
   - ak objekt O má svoju core vzdialenosť, ...

OPTICS: algoritmus 2/3

5.
- ...
- ak objekt O má svoju core-vzdialenosť,

a preusporiadanie susedov

- všetkým susedom určíme dostupnostnú vzdialenosť od objektu O
- novým objektom v množine susedov ju rovno priradíme
- v prípade, že niektoré objekty už boli v množine susedov (teda majú priradenú svoju dostupnostnú vzdialenosť od predtým spracovávaného objektu), priradíme im menšiu z týchto dvoch vzdialeností
- susedov usporiadame podľa ich dostupnostnej vzdialenosti vzostupne

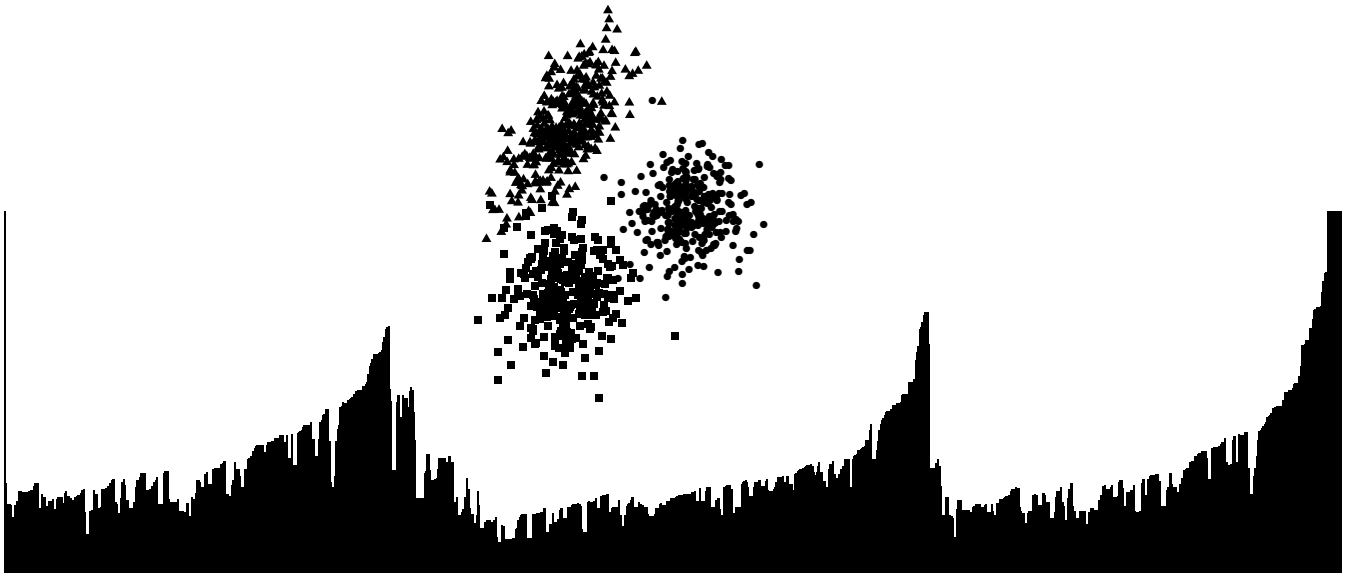
b ...

OPTICS: algoritmus 3/3

5. ▸ ...
   ▸ ak objekt O má svoju core-vzdialenosť,
   a preusporiadanie susedov ...
   b určenie core-vzdialenosti susedovi s najnižšou dostupnostnou vzdialenosťou, teda prvému v poradí
   c zápis spracovávaného suseda do finálneho usporiadania a jeho vylúčenie z množiny susedov a z neusporiadaných objektov
   d určenie objektov v ε-okolí spracovávaného suseda - t.j. nových susedov
   e ak má tento sused svoju core-vzdialenosť, opätovne sa preusporiada množina susedov (bod 5.a)
   f algoritmus pokračuje bodom 5.b až pokým nie sú spracovaní všetci susedia
6. algoritmus končí v prípade, že usporiadal všetky objekty; ak ešte sú neusporiadané objekty, algoritmus pokračuje bodom 1.

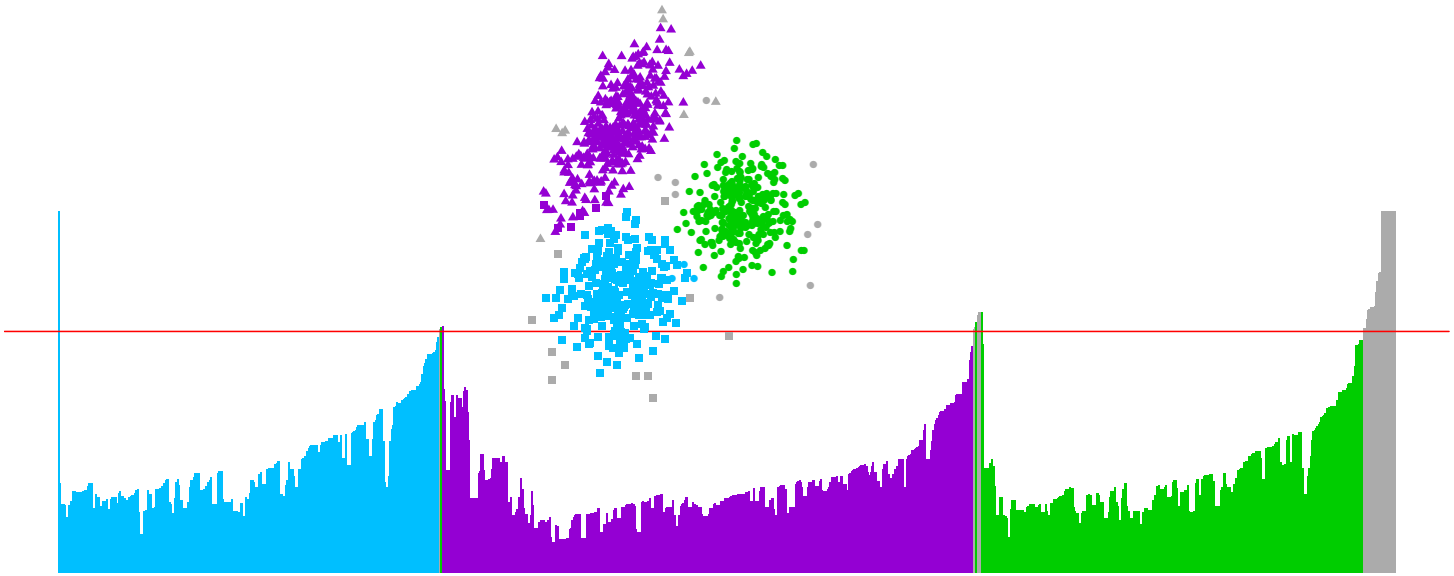
OPTICS: usporiadanie množiny objektov 1

- graf dostupnostnej vzdialenosti objektov, ktoré sú usporiadané algoritmom OPTICS



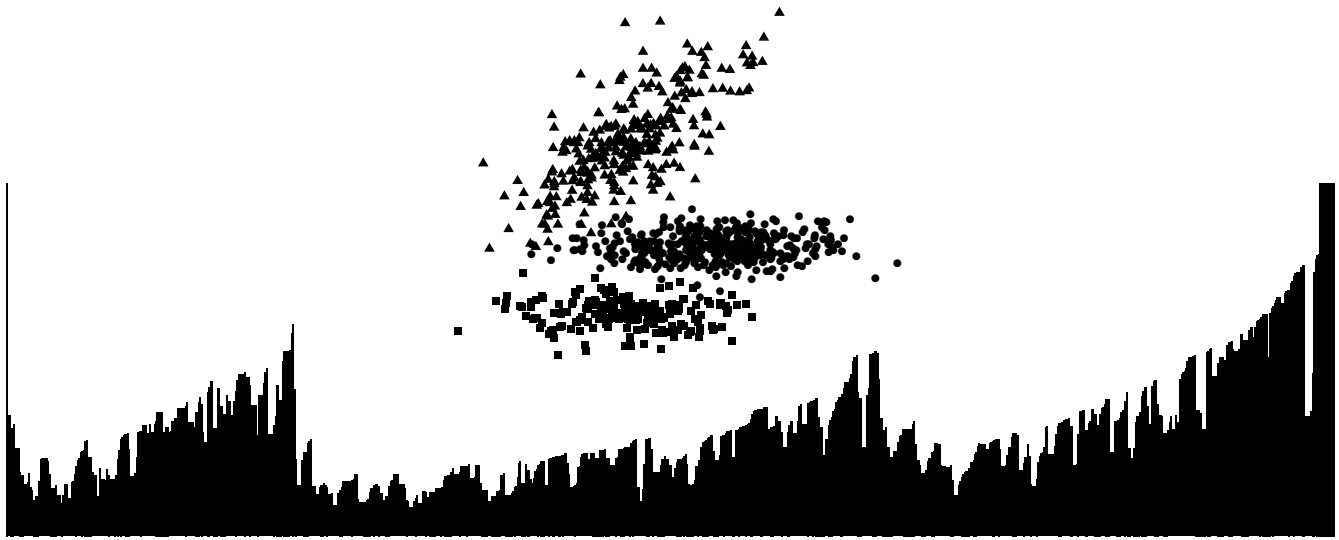
OPTICS: DBSCAN klastrovanie

- preložením čiary vo výške zvoleného ε získavame klastrovanie algoritmom DBSCAN pri danom N_{min}



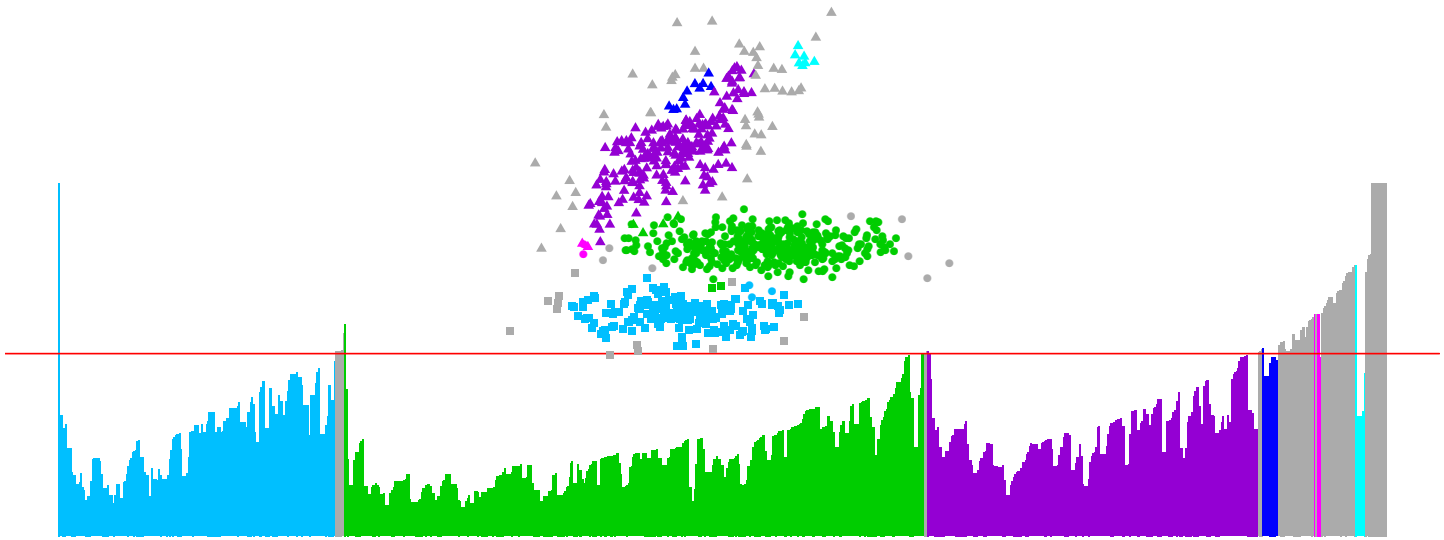
OPTICS: usporiadanie množiny objektov 2

- graf dostupnostnej vzdialenosti objektov, ktoré sú usporiadané algoritmom OPTICS



OPTICS: DBSCAN klastrovanie

- preložením čiary vo výške zvoleného ε získavame klastrovanie algoritmom DBSCAN pri danom N_{min}



OPTICS: výhody a nevýhody

+ rovnaké ako pri DBSCANe plus
+ nie je potrebné voliť fixné, ale iba maximálne ε, čím je možné identifikovať aj klastre rozličnej hustoty
+ poskytuje celkový pohľad na štruktúru dát - je možné vytvoriť hierarchické klastrovanie

x voľba parametrov ε a N_{min}
x neschopnosť určiť klastre, ktoré od seba nie sú jednoznačne oddelené
x otázny spôsob ako z usporiadania určiť najvhodnejšie rozklastrovanie objektov

Klasifikačné algoritmy

Klasifikačné algoritmy

LDA (Lineárna diskriminačná analýza)
- Fisher (1936)

Logistická regresia

Neurónové siete

SVM (Support Vector Machines)
- Cortes, Vapnik (1995)

KNN (k-nearest neighbors; k najbližších susedov)
- Royall (1966)

Klasifikačný strom
- Breiman, Friedman, Olshen, Stone (1984)

Náhodný les
- Breiman (2001)

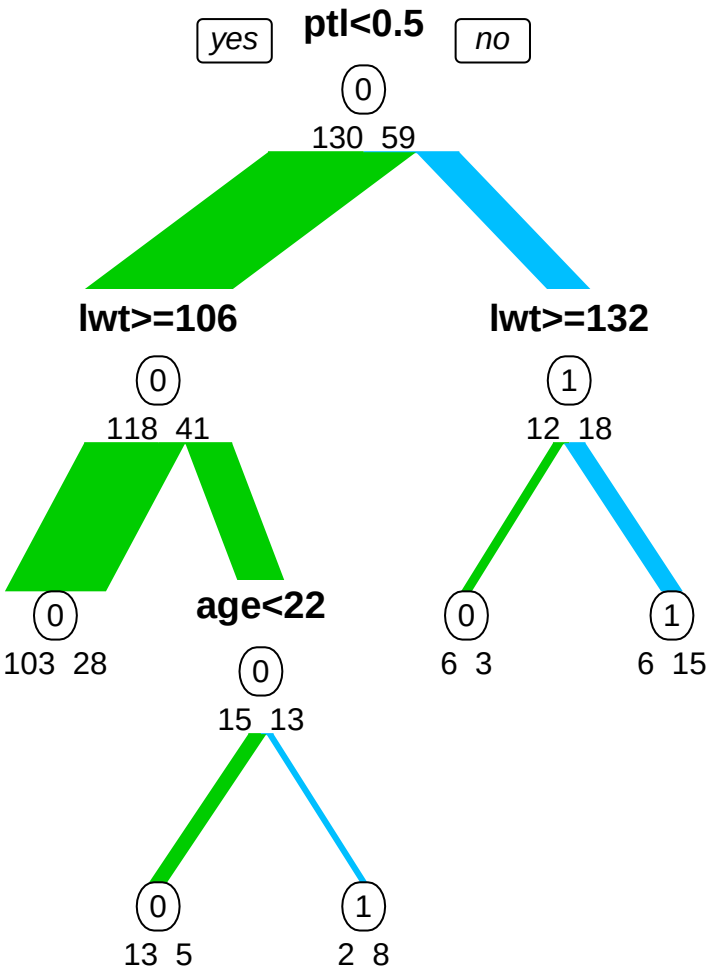
Ilustračný príklad

Dáta `birthwt`, v `R`-kovej knižnici `MASS`.

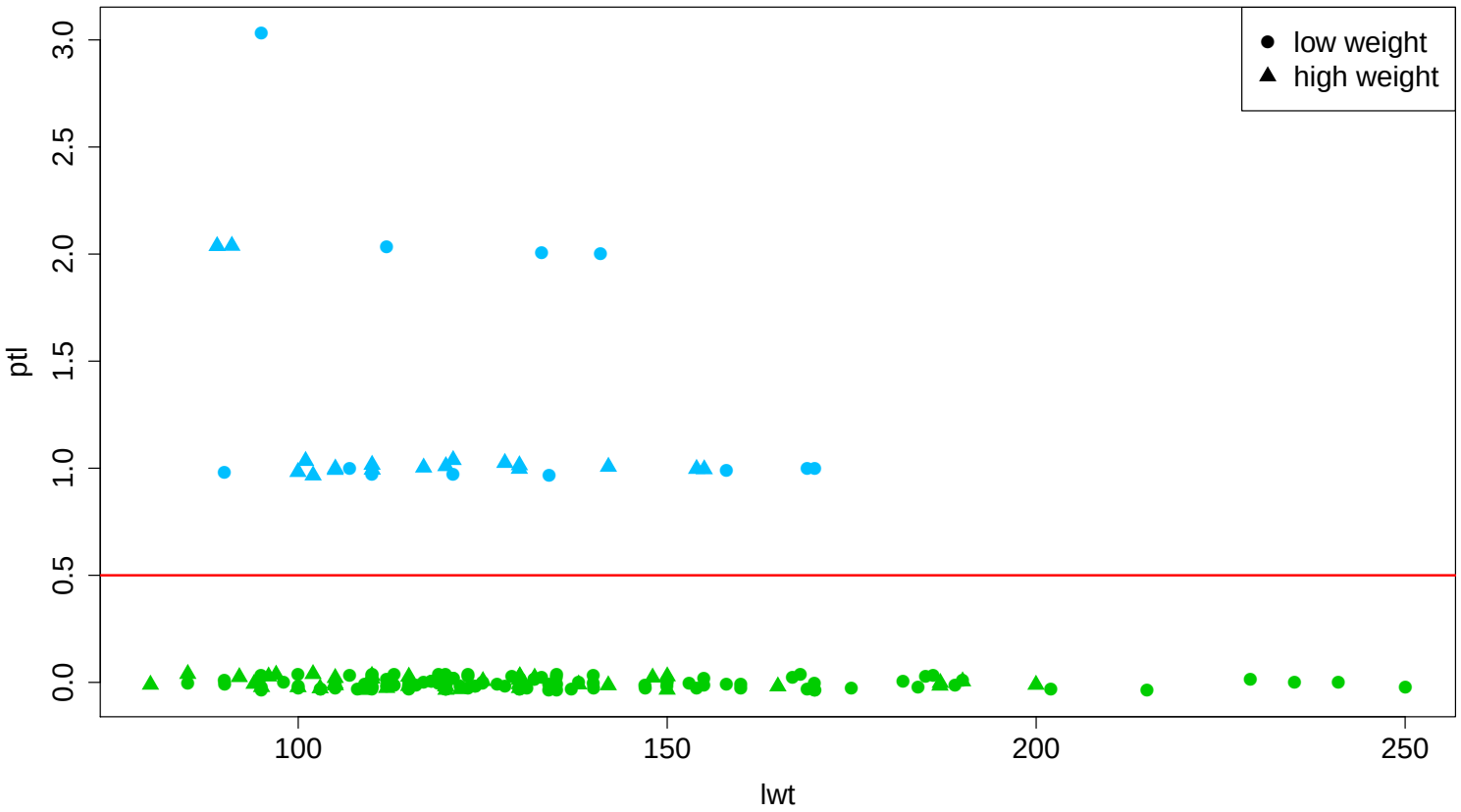
Medicínske dáta. 189 pacientiek - rodičiek. O každej sa vie 8 ukazovateľov (vek, rasa, počet predchádzajúcich predčasných pôrodov, atď.). Vie sa aj to, či sa dotyčnej rodičke narodilo dieťa s nízkou váhou (< 2.5 kg) alebo nie.

Úloha: na základe týchto dát zaradiť budúcu rodičku so známymi hodnotami uvedených ukazovateľov do jednej z tried: riziková (`low`), nerizikova (`high`).

Klasifikačný strom: rekurzívne binárne delenie



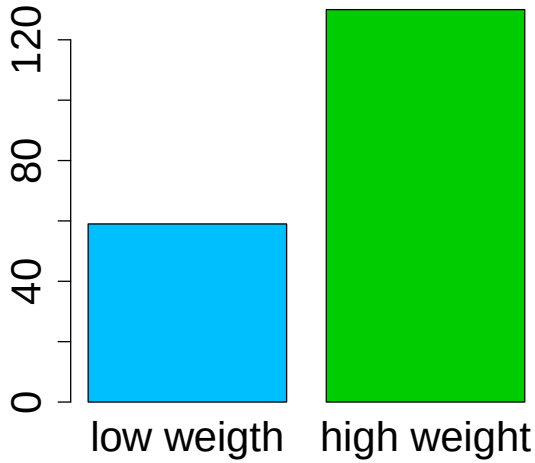
Klasifikačný strom: prvý 'split'



Miera nečistoty: Entropia

Entropia dát:

$$i(\tau) = -\frac{59}{189}\log\frac{59}{189} - \frac{130}{189}\log\frac{130}{189}$$



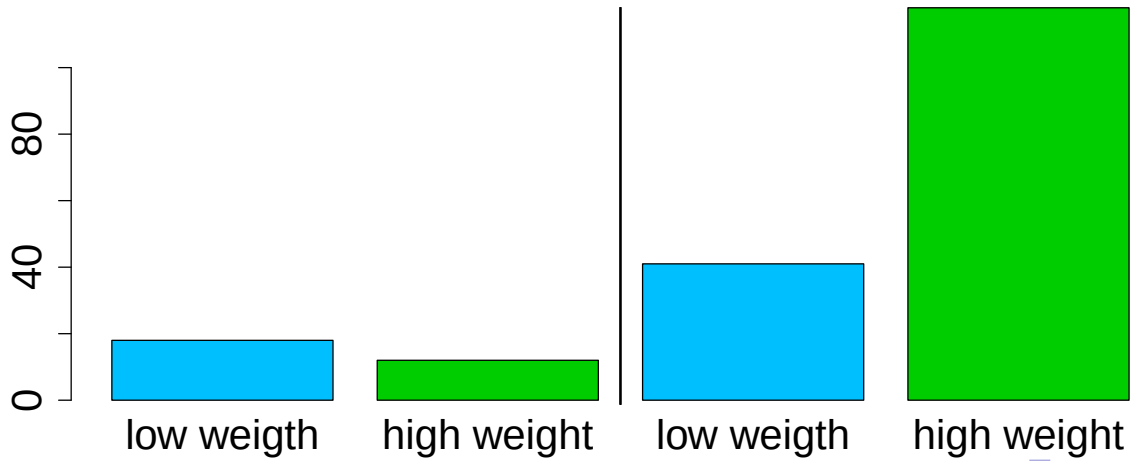
Entropia rozdelených dát (prvý 'split')

Entropia hornej polovice dát:

$$i(\tau_u) = -\frac{18}{30}\log\frac{18}{30} - \frac{12}{30}\log\frac{12}{30}$$

Entropia dolnej polovice dát:

$$i(\tau_l) = -\frac{41}{159}\log\frac{41}{159} - \frac{118}{159}\log\frac{118}{159}$$



Optimalizačné kritérium

Deliaca nadrovina s sa volí tak, aby sa rozdiel medzi nečistotou $i(\tau)$ v materskom uzle τ (t.j. pred delením) a (váženou) sumárnou nečistotou $i(s_u) + i(s_l)$ dcérskych uzlov s_u, s_l (t.j., po delení) bol maximálny:

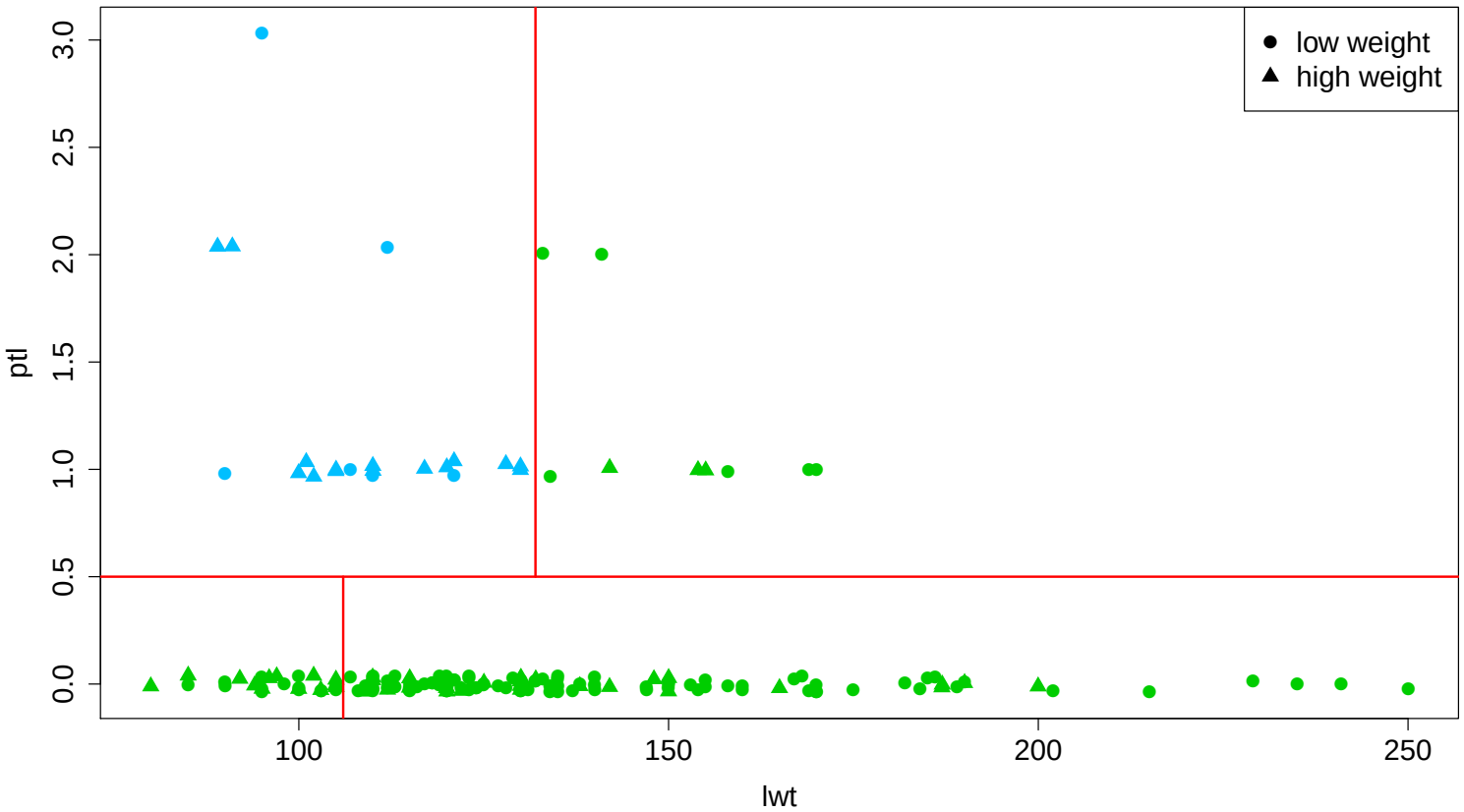
$$\hat{s} = \arg\max_{s} \Delta(s, \tau)$$

kde

$$\Delta(s, \tau) = i(\tau) - (p(s_u)i(s_u) + p(s_l)i(s_l))$$

a $p(\tau_d)$ je relatívna početnosť pozorovaní v dcérskom uzle $d \in \{u, l\}$.

Klasifikačný strom: druhý 'split'



Klasifikačná tabuľka

Klasifikačná tabuľka

	0	1
0	122	8
1	36	23

Klasifikačný strom: +/-

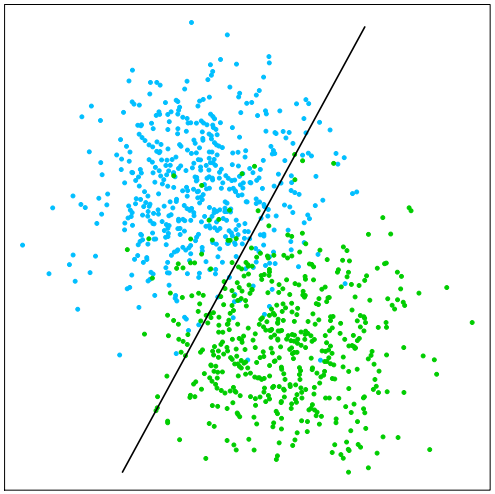
Výhody:

- ľahká interpretovatelnosť,
- ľahko získatelná dôležitosť prediktorov,
- rýchly algoritmus,
- použitelné aj na dáta kde $n \ll p$.

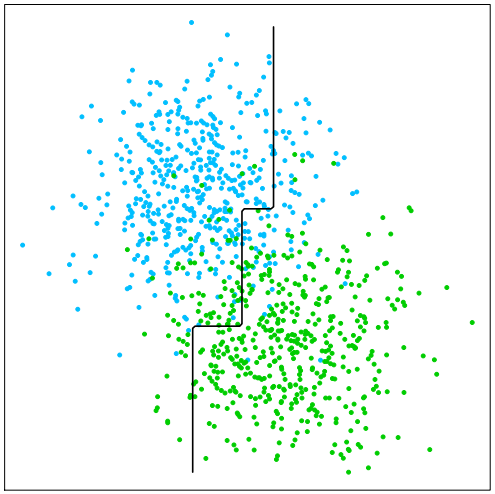
Nevýhody:

- existujú aj lepšie zatrieďovacie algoritmy,
- nestabilita.

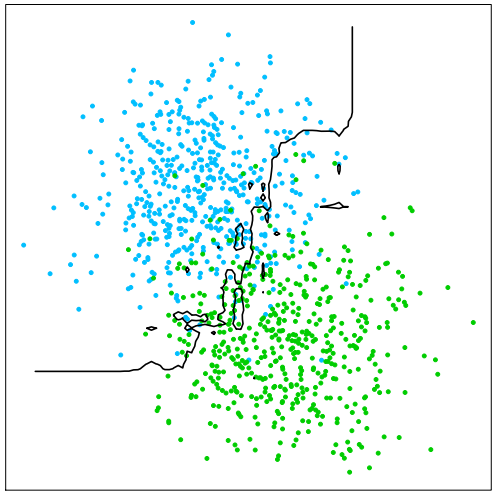
Porovnanie niektorých klasifikačných algoritmov



(a) LDA



(b) klasifikačný strom



(c) náhodný les

Náhodný les: algoritmus

Algoritmus:

1. Zvoľ T, počet stromov v lese.
2. Zvoľ m, počet prediktorov.
3. Nechaj narásť T stromov, a to nasledovným spôsobom:
   - urob bootstrapový výber z dát, na nich nechaj narásť strom;
   - pri raste stromu, v každom uzle náhodne vyber m prediktorov a pre nich nájdi najlepšie delenie;
   - takto nechaj narásť strom až po spodok.
4. Nový objekt je zatriedený do tej triedy, kam ho zaradí väčšina stromov lesa.

Náhodný les: +/-

Výhody:

- prirodzene zvláda dáta zmiešaného typu
- vie si poradiť s chýbajúcimi pozorovaniami
- robustný voči odlahlým pozorovaniam
- invariantný na monotónnu transformáciu prediktorov
- škálovatelný
- vie si poradiť s irelevantnými prediktormi
- vynikajúci výkon

Nevýhody:

- slabá interpretovatelnosť,

Ďakujeme za pozornosť.