Rastislav Kamenický
Michal Jarábek
Michal Piovarči

Záverečná správa

Databáza:

Ako databázu sme použili súbor dát o hríboch. Obsahuje vzorky zodpovedajúce 23 druhom lupeňových hríbov. V databáze máme 8124 hríbov s 23 príznakmi.

Popis príznakov

1. Jedlosť: jedlé=e, jedovaté=p
2. Tvar klobúku: zvon=b, kuželový=c, konvexný=x, plochý=f, uzlový=k, klesnutý=s
3. Povrch klobúku: vlaknitý=f, drážkový=g, šupinatý=y, hladký=s
4. Farba klobúku: hnedá=n, žltohnedá=b, škoricová=c, g=šedá, zelená=r, ružová=p, fialová=u, červená=e, biela=w, žltá=y
5. podliatiny: ano=t, nie=f
6. pach: mandle=a, aníz=l, kreozot=c, rybí=y, hnusný=f, zatuchnutý=m, žiaden=n, ostrý=p, pikantné=s
7. Pripevnenie lupeňov: pripojený=a, klesajúce=d, voľné=f, vrúbkované=n
8. Umiestnenie lupeňov: blízko=c, preplnené=w, vzdialené=d
9. Veľkosť lupeňov: široká = b, úzka = n
10. Farba lupeňov: čierna=k, hnedá=n, žltohnedá=b, čokoláda=h, sivá=g, zelená=r, oranžová=o, ružová=p, fialová=u, červená=e, biela=w, žltá=y
11. Tvar stonky: zväčšujúci=e, zužujúci=t
12. Koreň stonky: baňatý=b, kyj=c, šálka=u, rovný=e, vláknitý=z, korene=r, chýba=?
13. Povrch stonky nad prstencom: vláknitý=f, šupinatý=y, hodvábny=k, hladký=s
14. Povrch stonky pod prstencom: vláknitý=f, šupinatý=y, hodvábny=k, hladký=s
15. Farba stonky nad prstencom: hnedá=n, žltohnedá=b, škoricová=c, g=šedá, oranžová=o, ružová=p, červená=e, biela=w, žltá=y
16. Farba stonky pod prstencom: hnedá=n, žltohnedá=b, škoricová=c, g=šedá, oranžová=o, ružová=p, červená=e, biela=w, žltá=y
17. Typ závoja: čiastočný=p, univerzálny=u
18. Farba závoja: hnedá=n, oranžová=o, biela=w, žltá=y
19. Počet prstencov: žiaden=n, jedna=o, dva=t
20. Typ prstencov: pavučinový=c, pominuteľný=e, rozširujúci=f, veľký=l, nič=n, príveskový=p, opláštovací=s, zóna=z
21. Farba stopy výtrusu: čierna=k, hnedá=n, žltohnedá=b, čokoláda=h, zelená=r, oranžová=o, purpurová=u, w=biela, žltá=y
22. obsadenie: hojné=a, skupinové=c, početné=n, roztrúsené=s, niekoľko=v, osamelé=y
23. nálezisko: trávy=g, lístia=l, lúky=m, cesty=p, mestá=u, odpad=w, lesy=d

Ako je z popisu vidno niektoré huby nemajú zaradený koreň stonky. Keďže naša databáza obsahuje veľké množstvo hríbov sme tento nedostatok vyriešili nezaradením húb do vzorky, s ktorou sme následne pracovali a teda nám ostalo 5936 hríbov.

Zamerali sme sa najme na zistenie jedlosti hríbov keďže v popise databázy bolo spomenuté, že sa jedlosť dá odvodiť s vysokou pravdepodobnosťou dokonca niekoľkými spôsobmi len na základe logických formúl typu farba stopy výtrusu=zelená, ktorá na 99% trafí jedovatý hríb. Ďalší príznak, ktorý sme sa snažili odvodiť bolo nálezisko ktoré nám pripadalo, že sa bude dať tiež dobre odvodiť, keďže prostredie má vplyv na rastliny aj zvieratá

Redukcia príznakov

Na redukciu príznakov sme použili algoritmy PCA a ICA.

PCA

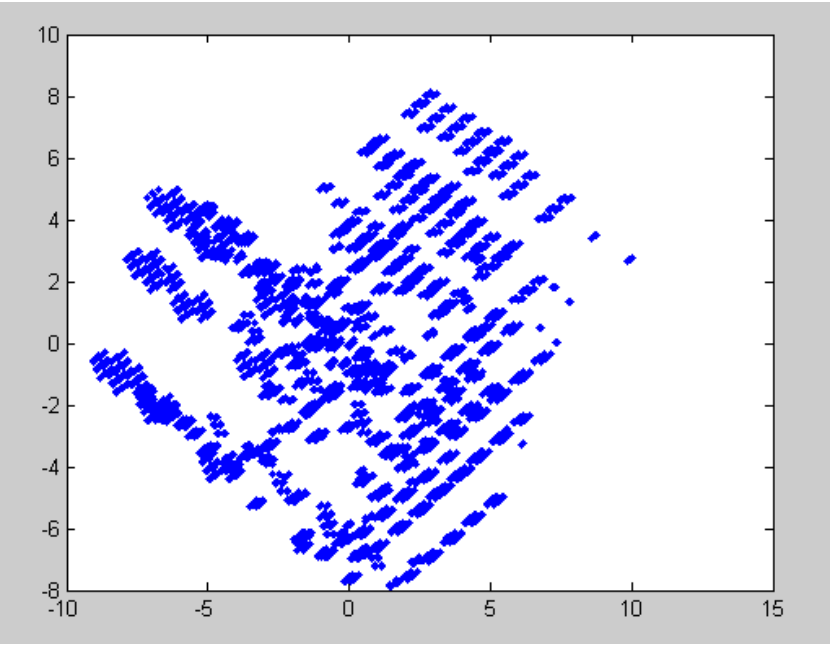
Tento algoritmus vyberá základné prvky na základe variancie pričom sa snaží ponechať kolmosť medzi prvkami. Toto však znamená, že najväčší vplyv budú mať príznaky s najväčším počtom hodnôt. V našom prípade teda budú mat najväčší vplyv príznaky farby rôznych častí hríbu.
V prípadoch, že zredukujeme dáta na málo dimenzii, tak sa môže stať, že nám ostanú hlavne farby, ktoré nemusia pomôcť pri predpovedi požadovaného príznaku.

ICA

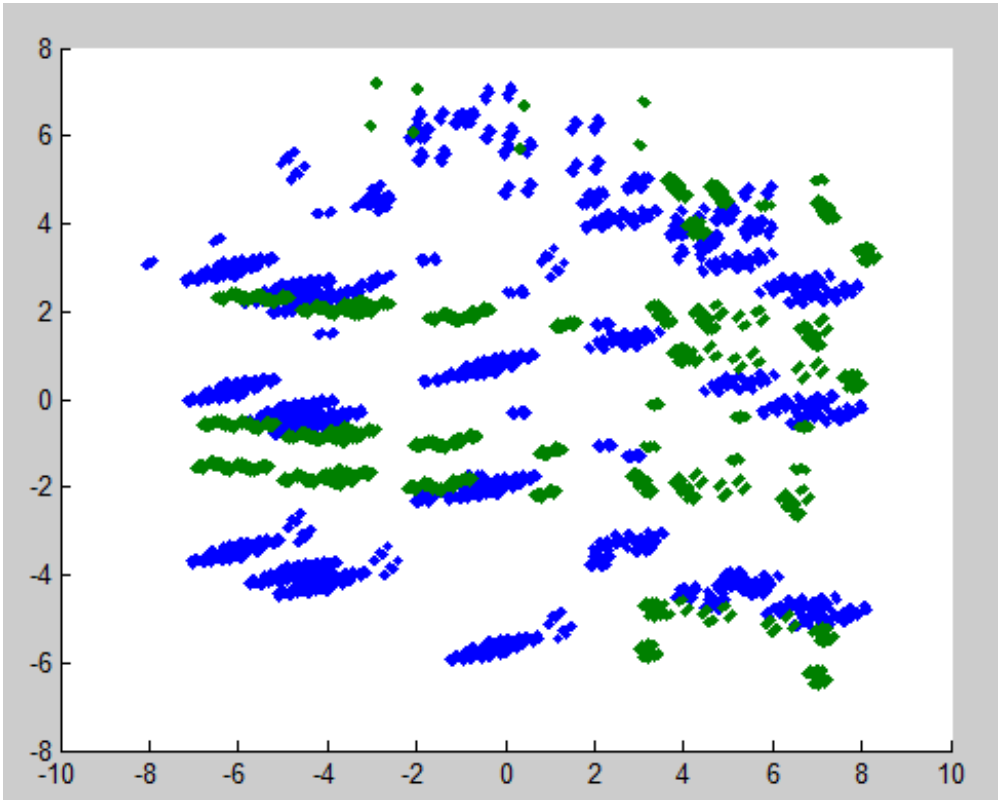
ICA je metóda na rozdeľovanie mnohorozmerného signálu na samostatné komponenty, pričom predpokladá negausovské rozloženie dát a nezávislosť daných dát. Keďže nevieme ako sú naše dáta rozdelené, nevieme dopredu povedať, ktorý z algoritmov bude lepší a môže sa stať že naše dáta nebudú idealné separovateľne a tento algoritmus zlyhá.

Vizualizácia dát - pomocou Multi Dimensional Data Scaling (MDDS)

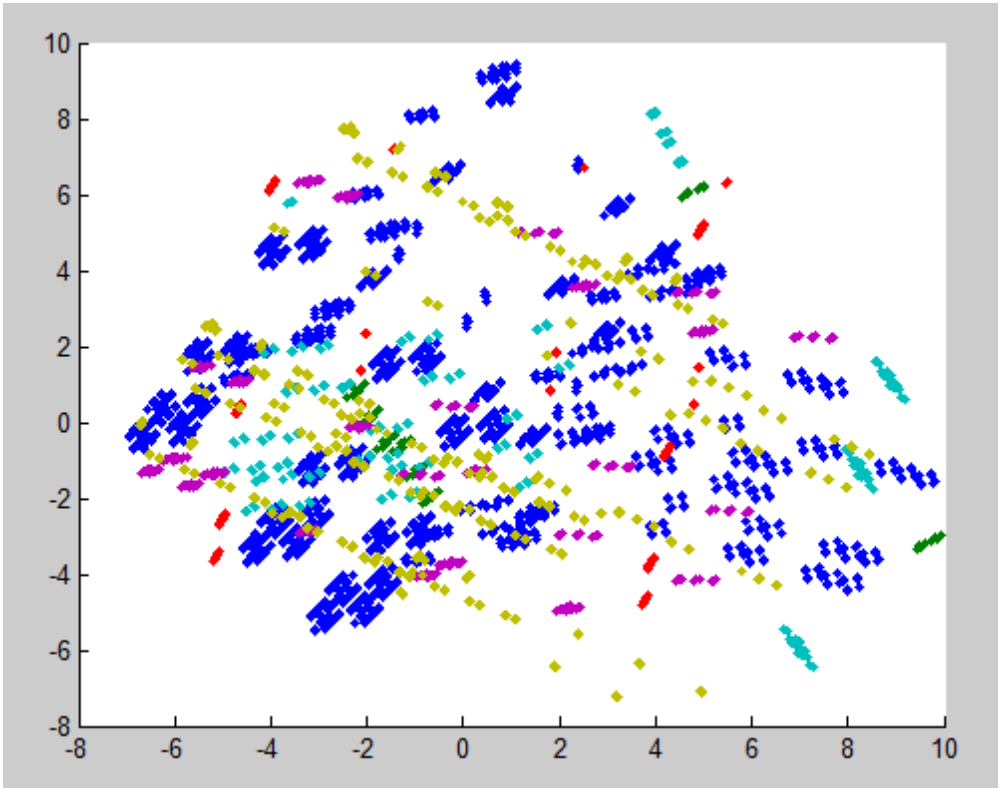
Nezredukované 23 rozmerné dáta:



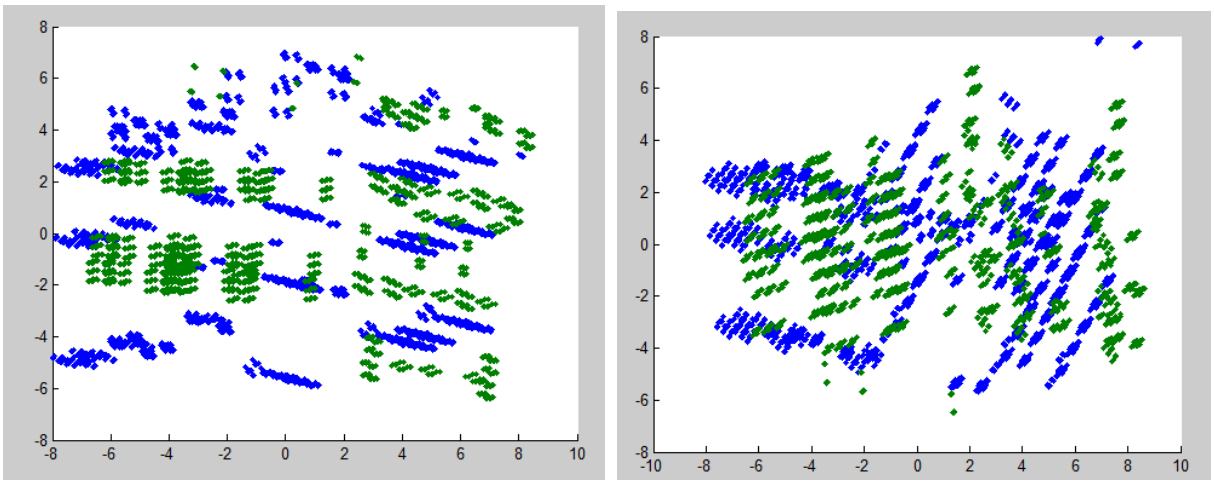
Dáta rozdelené podla jedlosti:



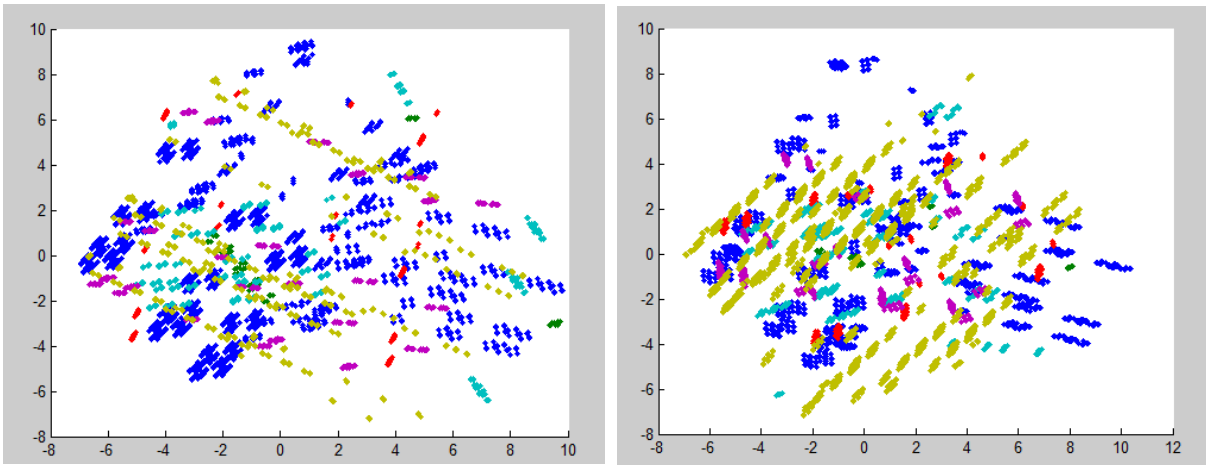
Dáta rozdelené podla náleziska:



Dáta rozdelené podľa jedlosti zredukované na 5 dimenzii (vľavo) 2 dimenzie (vpravo)



Dáta rozdelené podľa výskytu 10 dimenzii (vľavo) 2 dimenzie (vpravo)



Klasifikácia

Dáta sme klasifikovali pomocou lineárneho klasifikátoru, neurónových sietí, SOM a rozhodovacieho stromu. Všetky tieto klasifikátory sme najprv naučili sa rozhodovať na trénovacej množine a následne sme ich nechali rozhodovať na ostatku dát. Ide o vrstvenú 5-násobnú krížovú validáciu.

Lineárny klasifikátor

Rozdeľuje vstupné dáta nadrovinou, tak aby mal jednu triedu na jednej strane a druhú triedu na druhej strane nadroviny.

Predpoklad je, že lineárny klasifikátor nebude veľmi úspešný, lebo dáta by museli byť lineárne separabilné vo veľkých dimenziách, alebo zredukovateľné do veľmi malej dimenzie.

Neurónové siete

Sieť ktorá si na trénovacej množine upraví váhy vstupných dát a prechodov medzi neurónmi šírením chyby, tak aby sme dostávali čo najlepšie výsledky. Pri vhodnom učení dosiahne globálne minimum chyby a preto ak zvolíme vhodnú trénovaciu množinu je pravdepodobné že bude pri klasifikácii správne klasifikovať naše huby.

Predpoklad je, že neurónové siete by mohli byť veľmi úspešné. Náhodné trénovanie ukázalo, že sú schopné trénovaciu množinu pri jedlosti rozdeliť so 100% úspešnosťou a nálezisko s 80% úspešnosťou.

SOM

Samo organizujúce mapy sú podobne ako neurónové siete sústava pospájaných neurónov avšak sú pospájané do mriežky a vstupné dáta sú spojené s každým neurónom. Výsledkom SOM je mapa kde podobné huby budú mapované blízko seba a rozdielne huby oddenelne.

Predpoklad je, že SOM by mohli byť pomerne úspešné, keďže huby s rovnakými vlastnosťami by mohli byť podobné aj v iných aspektoch.

Rozhodovací strom

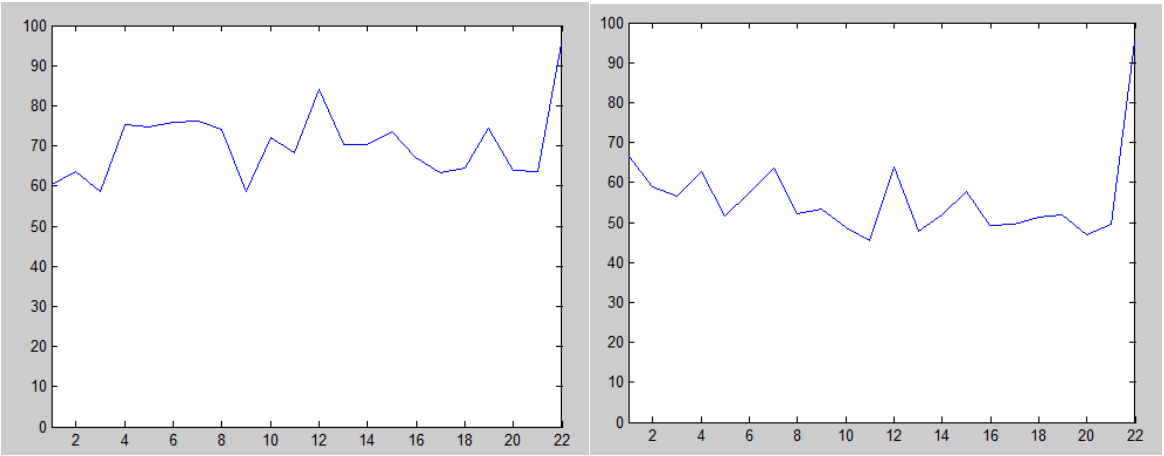
Vytvára strom tak že pre každý uzol vezme príznak s najlepším ohodnotením a rozvetví ho na základe hodnoty tohto príznaku. Dáta rozdeľuje na základe vetvenia stromu na podmnožiny až kým nedostane jednoznačné rozdelenie požadovaného príznaku. Teda každý list obsahuje len hľadaný príznak s jednou hodnotou.

Predpoklad je, že rozhodovací strom by mal byť tiež pomerne úspešný. Vybrali sme ho preto, aby sme overili pravidlá „starých mám“, ktoré hovoria o jedlosti huby podľa niektorých jej znakov tak ako rozhodovací strom.

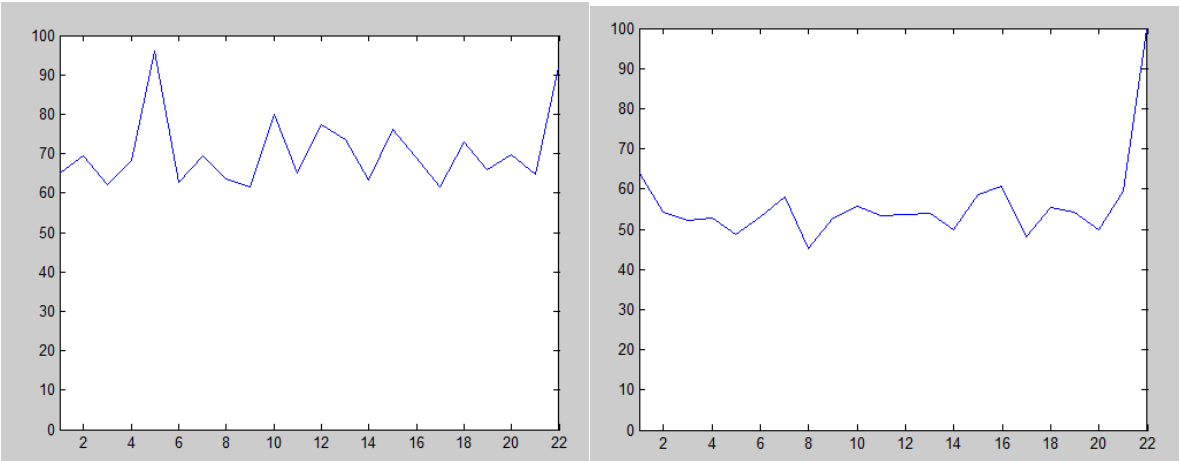
Krivky úspešnosti pracovný názov ROC*

Na validáciu sme použili vrstvenú 5-násobnú krížovú validáciu. Na X-ovej osi ROC* kriviek máme počet parametrov, na Y-ovej osi percentuálnu úspešnosť. Najprv ROC* krivky zobrazujúce úspešnosť zisťovania jedlosti a potom krivky pre zisťovanie náleziska.

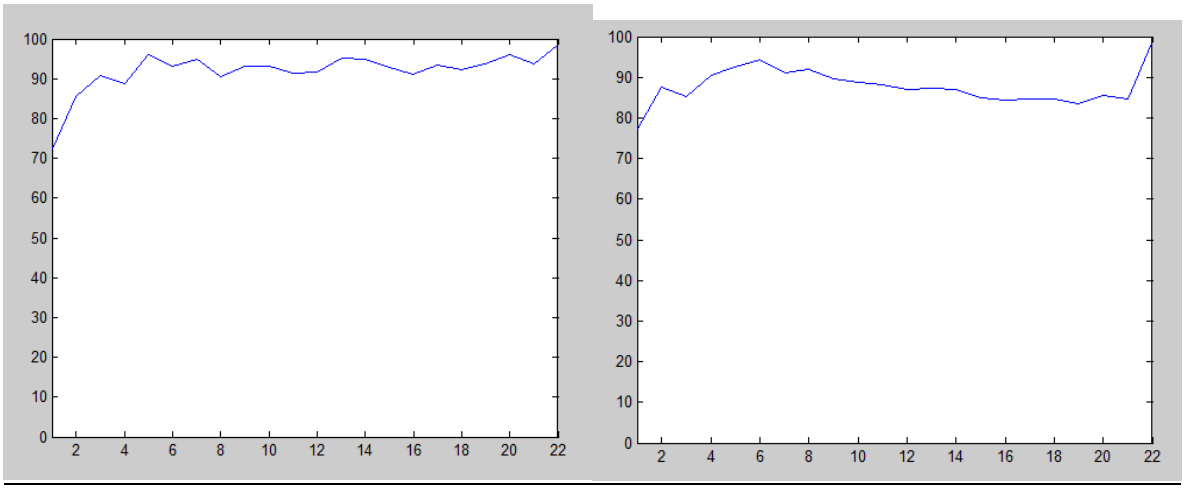
ROC* krivka na zisťovanie **jedlosti** pre **lineárny klasifikátor** naľavo s PCA na pravo ICA



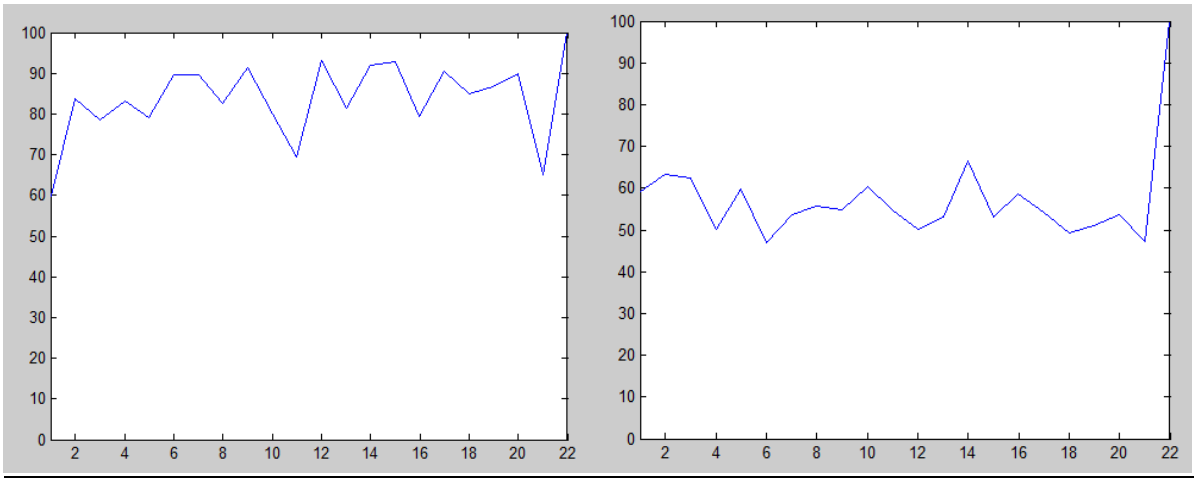
ROC* krivka na zisťovanie **jedlosti** pre **neurónové siete** naľavo s PCA na pravo ICA



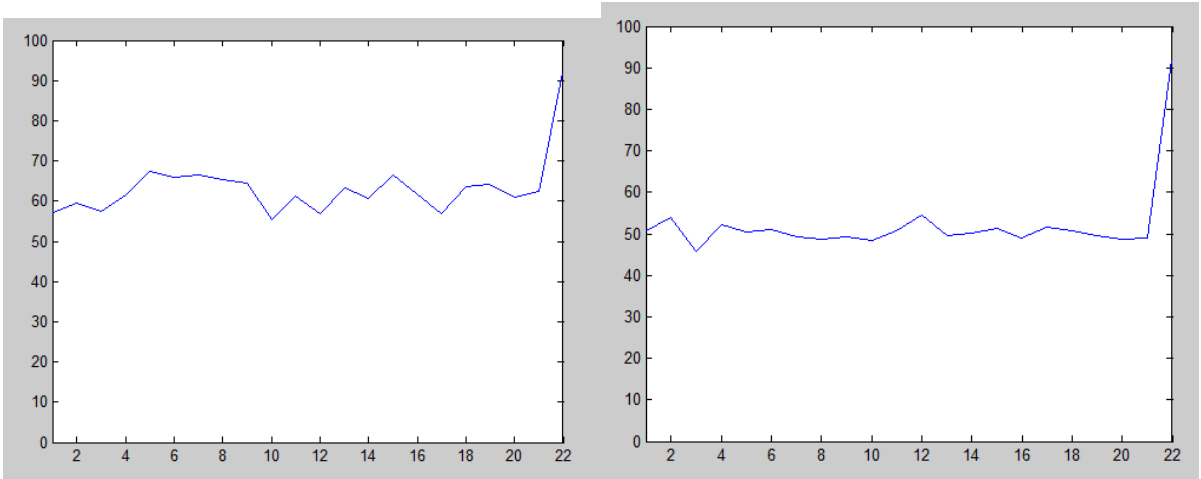
ROC* krivka na zisťovanie **jedlosti** pre **SOM** naľavo s PCA na pravo ICA



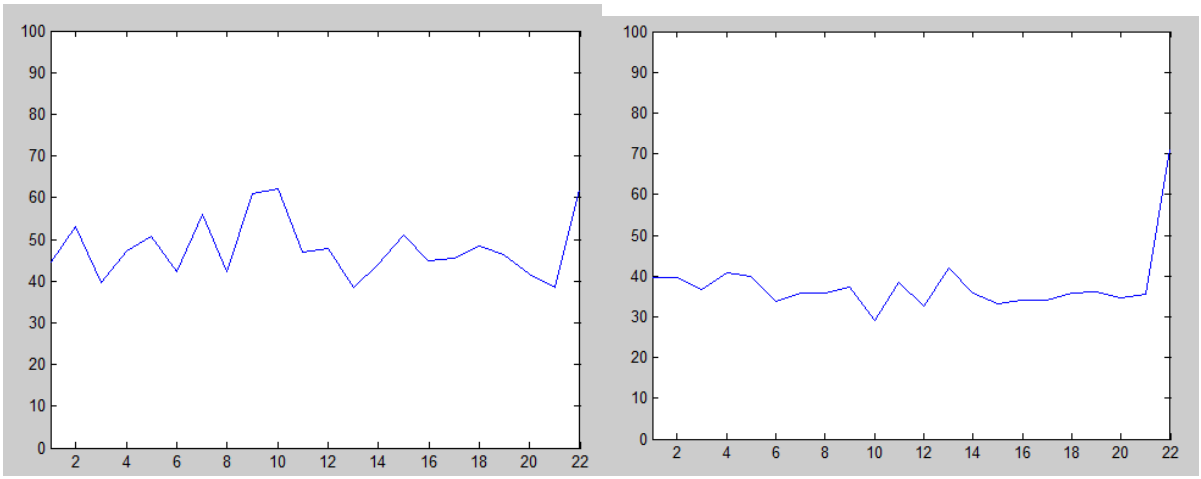
ROC* krivka na zisťovanie **jedlosti** pre **rozhodovací strom** naľavo s PCA na pravo ICA



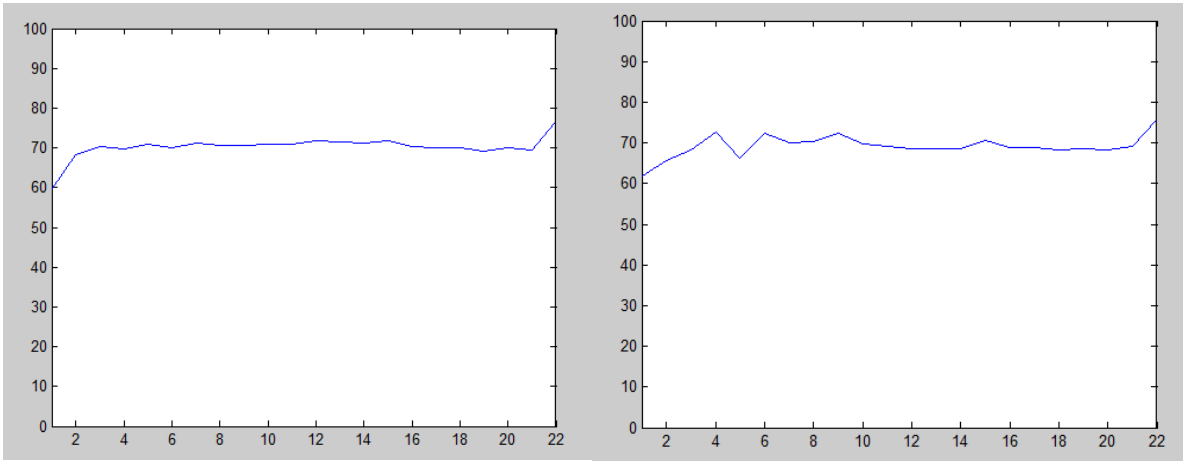
ROC* krivka na zisťovanie **náleziska** pre **lineárny klasifikátor** naľavo s PCA na pravo ICA



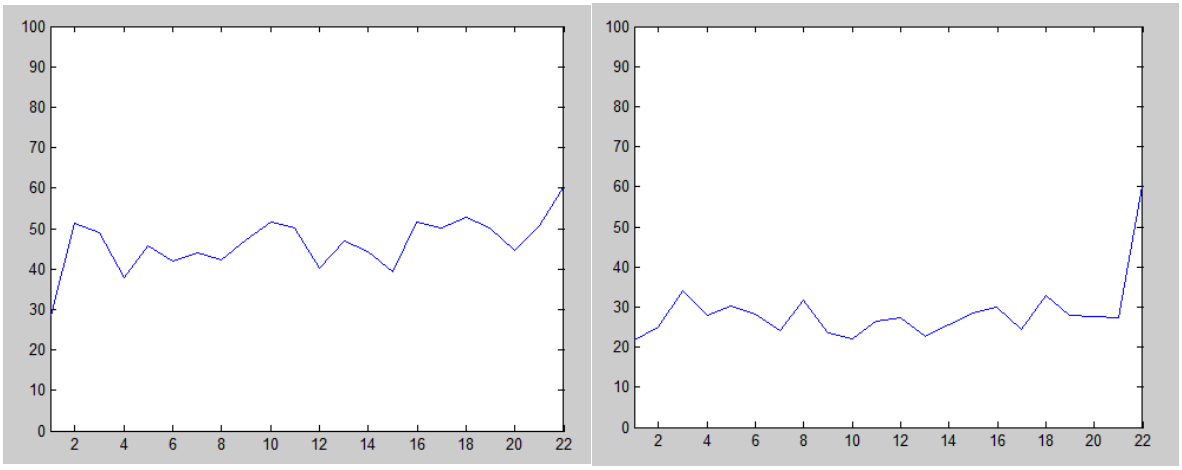
ROC* krivka na zisťovanie **náleziska** pre **neurónové siete** naľavo s PCA na pravo ICA



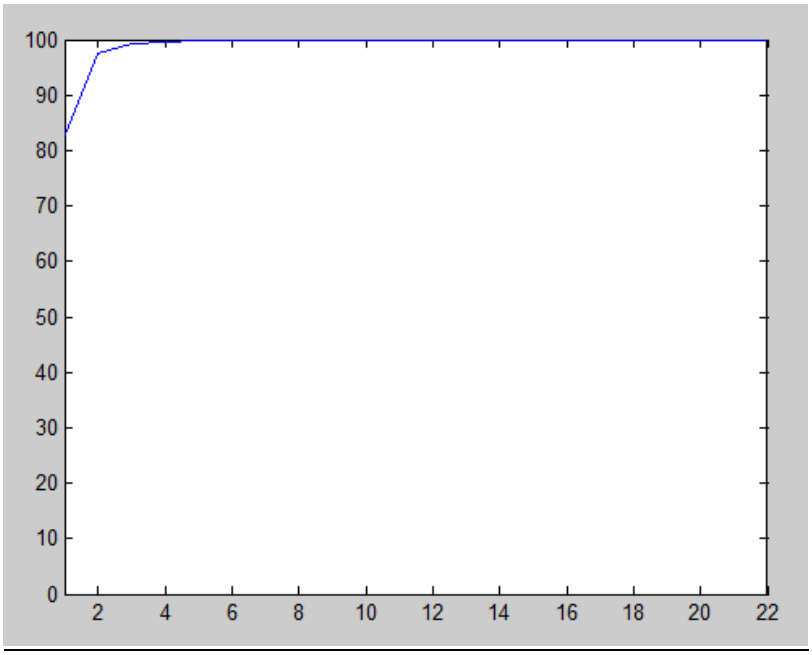
ROC* krivka na zisťovanie **náleziska** pre **SOM** naľavo s PCA na pravo ICA



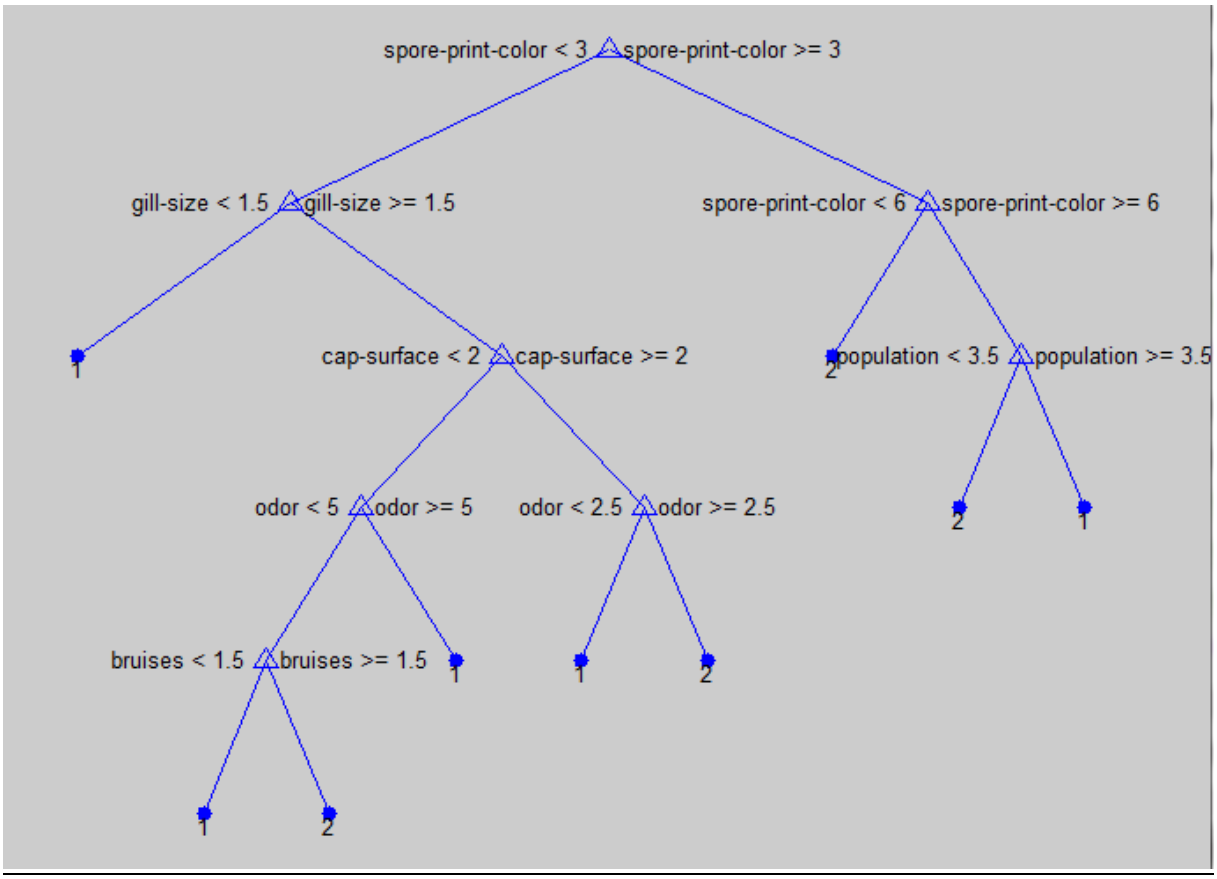
ROC* krivka na zisťovanie **náleziska** pre **rozhodovací strom** naľavo s PCA na pravo ICA



Príklad ROC* krivky pre jedlosť s PCA na pretrénovaných stromoch



Vizualizácia rozhodovacieho stromu pre jedlosť:



Confusion matice

Uvádzame len niektoré confusion matice lebo ostatné boli veľmi podobné (a zabrali by veľa strán ☺).

Confusion matica pre jedlosť na neurónových sieťach s PCA

	Jedlé	Jedovaté
Jedlé	741	2
Jedovaté	12	432

Confusion matica pre jedlosť na SOM s PCA

	Jedlé	Jedovaté
Jedlé	735	18
Jedovaté	18	415

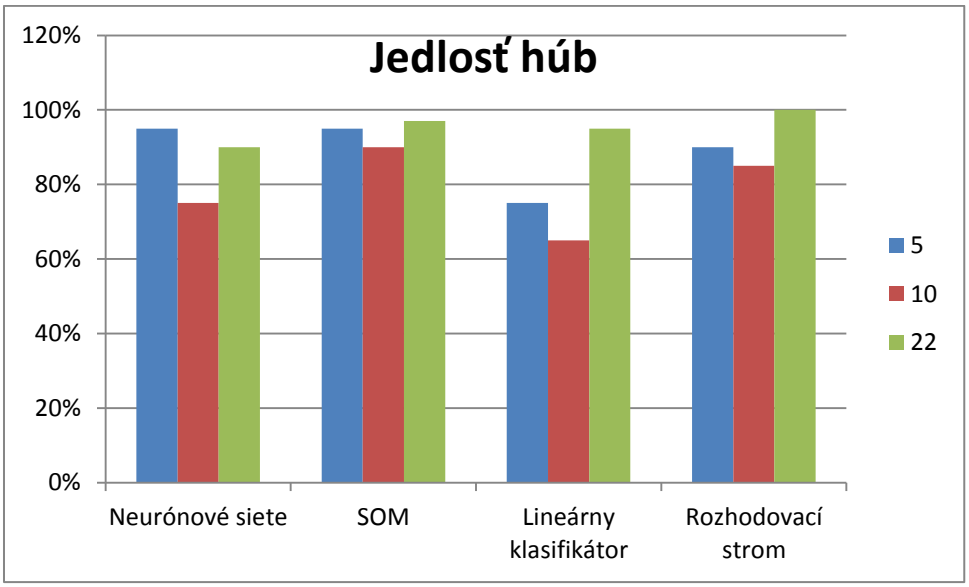
Confusion matica pre náleziská na neurónových sieťach s PCA

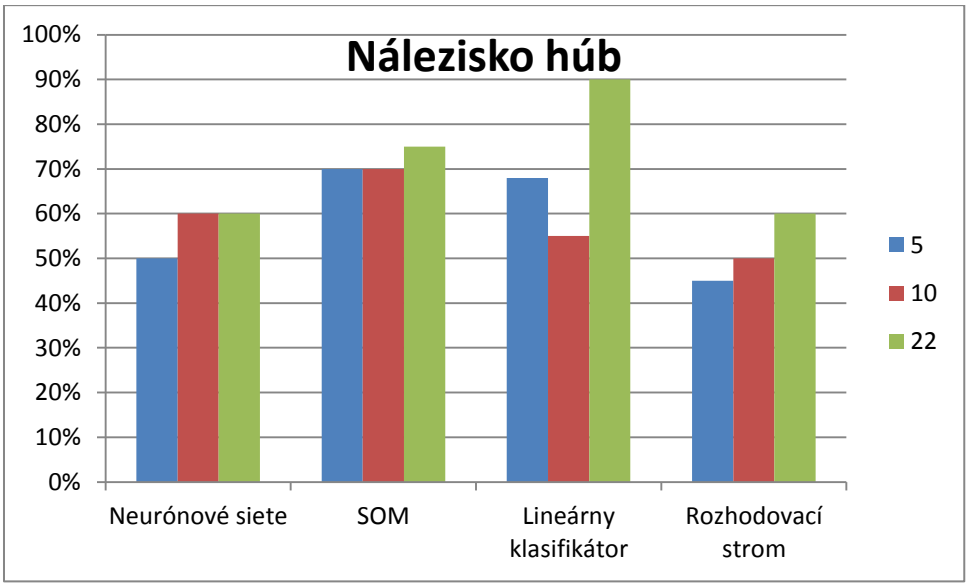
	Tráva	Lístie	Lúky	Cesty	Mestá	Lesy
Tráva	334	0	39	31	69	31
Lístie	0	18	0	0	3	0
Lúky	12	0	19	0	0	0
Cesty	24	0	0	22	0	6
Mestá	0	0	0	0	0	0
Lesy	54	0	0	61	1	463

Confusion matica pre náleziská na SOM s PCA

	Tráva	Lístie	Lúky	Cesty	Mestá	Lesy
Tráva	339	2	41	36	38	43
Lístie	0	16	0	0	0	0
Lúky	5	0	11	0	0	0
Cesty	44	0	0	57	0	29
Mestá	8	0	0	0	35	0
Lesy	27	0	7	20	0	429

Vyhodnotenie





Z ROC* kriviek je ľahko vidno že PCA algoritmus dosahoval všeobecne lepšie výsledky ako ICA algoritmus pričom najväčší rozdiel bol pri rozhodovacích stromoch. Je to preto, lebo dáta boli dobre separovateľné v smere najväčšej variancie dát. V redukovaných dimenziách neboli dáta lineárne separabilné, čo tiež ovlivnilo výsledok ICA.

Pri klasifikácii jedlosti húb sa neurónové siete a lineárny klasifikátor pretrénovali. Separovali trénovaciu množinu vždy na 90%+ ale mali problém správne určiť huby v testovacej množine. Ako najúspešnejší klasifikátor skončili SOM, ktoré mali najlepšie a najkonzistentnejšie výsledky. Z ROC* kriviek je vidno, že narozdiel od ostatných klasifikátorov SOM neboli náchilné na preučenie aj keď im zvyšovanie počtu dimenzii nepomáhalo výrazne zvýšiť úspešnosť klasifikácie.

Rozhodovacie stromy boli tiez pomerne úspešné hlavne pri nezredukovanej dimenzii, čo sme očakávali. Jedlosť húb s celou databázou dokázali rozhodnúť so skoro 100% úspešnosťou, čo

potvrdzuje predpoklad, že by mohli byť dobré kôli pravidlám „starých mám“, ktoré na základe niektorých znakov huby rozlišujú jej jedlosť/nejedlosť.

Neurónové siete mali problém s pretrénovaním. Trenovaciu množinu vždy klasifikovali na 100% v prípade jedlosti a 80% v prípade náleziska. Avšak tieto rozdelenia nesedeli s testovacou množinou húb. Zaujímavé je zisťovanie jedlosti kde pri 5 dimenziách dosiahli neurónové siete cez 95% úspešnosť, lebo neboli ovplivnené preučením.

Lineárný klasifikátor fungoval až nečakane dobre. S plnou databázou nezredukovaných dát dosahoval pre jedlosti húb skoro 95% úspešnosť a teda boli dáta lineárne separabilné. Nálezisko bolo určované lineárnym klasifikátorom po častiach. Teda pri plnej databáze mal v priemere 90% úspešnosť na otázku, či huba rastie/nerastie v jednom prostredí.

Rozhodovacíi strom pre nálezisko