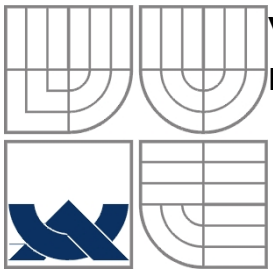
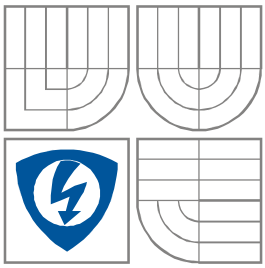


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

APLIKACE SHLUKOVÉ ANALÝZY PŘI ZPRACOVÁNÍ BIOMEDICÍNSKÝCH DAT

APPLICATION OF CLUSTERING METHODS FOR PROCESSING OF BIOMEDICAL DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. Michal Rozinek
AUTHOR

VEDOUCÍ PRÁCE — Ing. Martin Havlíček
SUPERVISOR

BRNO, 2009

LICENČNÍ SMLOUVA
POSKYTOVANÁ K VÝKONU PRÁVA UŽÍT ŠKOLNÍ DÍLO

uzavřená mezi smluvními stranami:

1. Pan/paní

Jméno a příjmení: Michal Rozinek
Bytem: Ještětice 65, Rychnov n. Kn., 516 01
Narozen/a (datum a místo): 30. ledna 1984 v Rychnově n. Kn.

(dále jen „autor“)

a

2. Vysoké učení technické v Brně

Fakulta elektrotechniky a komunikačních technologií
se sídlem Údolní 53, Brno, 602 00
jejímž jménem jedná na základě písemného pověření děkanem fakulty:
prof. Ing. Jiří Jan,CSc, předseda rady oboru Biomedicínské a ekologické inženýrství
(dále jen „nabyvatel“)

Čl. 1

Specifikace školního díla

1. Předmětem této smlouvy je vysokoškolská kvalifikační práce (VŠKP):

   ☐ disertační práce
   ☒ diplomová práce
   ☐ bakalářská práce
   ☐ jiná práce, jejíž druh je specifikován jako
   .....
   (dále jen VŠKP nebo dílo)

   Název VŠKP: Aplikace shlukové analýzy při zpracování biomedicínských dat
   Vedoucí/ školitel VŠKP: Ing. Martin Havlíček
   Ústav: Ústav biomedicínského inženýrství
   Datum obhajoby VŠKP: _____

   VŠKP odevzdal autor nabyvateli*:

   ☒ v tištěné formě – počet exemplářů: 2
   ☒ v elektronické formě – počet exemplářů: 2

2. Autor prohlašuje, že vytvořil samostatnou vlastní tvůrčí činností dílo shora popsané a specifikované. Autor dále prohlašuje, že při zpracovávání díla se sám nedostal do rozporu s autorským zákonem a předpisy souvisejícími a že je dílo dílem původním.

3. Dílo je chráněno jako dílo dle autorského zákona v platném znění.

4. Autor potvrzuje, že listinná a elektronická verze díla je identická.

* hodící se zaškrtněte

Článek 2
Udělení licenčního oprávnění

1. Autor touto smlouvou poskytuje nabyvateli oprávnění (licenci) k výkonu práva uvedené dílo nevýdělečně užít, archivovat a zpřístupnit ke studijním, výukovým a výzkumným účelům včetně pořizovaní výpisů, opisů a rozmnoženin.
2. Licence je poskytována celosvětově, pro celou dobu trvání autorských a majetkových práv k dílu.
3. Autor souhlasí se zveřejněním díla v databázi přístupné v mezinárodní síti

   ☒ ihned po uzavření této smlouvy
   ☐ 1 rok po uzavření této smlouvy
   ☐ 3 roky po uzavření této smlouvy
   ☐ 5 let po uzavření této smlouvy
   ☐ 10 let po uzavření této smlouvy
   (z důvodu utajení v něm obsažených informací)

1. Nevýdělečné zveřejňování díla nabyvatelem v souladu s ustanovením § 47b zákona č. 111/ 1998 Sb., v platném znění, nevyžaduje licenci a nabyvatel je k němu povinen a oprávněn ze zákona.

Článek 3

Závěrečná ustanovení

1. Smlouva je sepsána ve třech vyhotoveních s platností originálu, přičemž po jednom vyhotovení obdrží autor a nabyvatel, další vyhotovení je vloženo do VŠKP.

2. Vztahy mezi smluvními stranami vzniklé a neupravené touto smlouvou se řídí autorským zákonem, občanským zákoníkem, vysokoškolským zákonem, zákonem o archivnictví, v platném znění a popř. dalšími právními předpisy.

3. Licenční smlouva byla uzavřena na základě svobodné a pravé vůle smluvních stran, s plným porozuměním jejímu textu i důsledkům, nikoliv v tísni a za nápadně nevýhodných podmínek.

4. Licenční smlouva nabývá platnosti a účinnosti dnem jejího podpisu oběma smluvními stranami.

V Brně dne: 29. května 2009

…….. ……

Nabyvatel Autor

ABSTRAKT

Cílem této práce je nastudovat postupy pro klasifikaci objektů v medicíně. Zjistit pomocí jakých metod lze takovou klasifikaci provádět. Po vytvoření algoritmů v prostředí MATLAB pro klasifikační metody zjistit jejich funkčnost a spolehlivost na datovém souboru z oblasti medicíny. Formou jednoduchých testů podrobit zkoušce jednotlivé metody a zjistit v čem vynikají a v čem zaostávají.

Velmi důležitou roli zde hraje i volba vstupních datových parametrů, i ty budou podrobeny testu a popsány v závěrech.

ABSTRACT

The goal of this study is to learn about methods in object classification in medicine and find out what are these methods about. Focusing on functionality and reliability of these methods whith datafile from the medicine compartment after making the algorithm in MATLAB. In form of siple tests, put the touch everyone of classification procedure and find out in which they excel and in which they lags.

The choice of input data parametres is very important, this will be tested and noted in conclusions.

KLÍČOVÁ SLOVA:

Klasifikace, biomedicína, shlukování, testování

KEYWORDS:

Classification, biomedicine, clustering, testing

Prohlášení

Prohlašuji, že svou diplomovou práci na téma Aplikace shlukové analýzy při zpracování biomedicínských dat jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 29. května 2009 .....

podpis autora

Poděkování

Děkuji vedoucímu diplomové práce Ing.Martinu Havlíčkovi za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne 29. května 2009 .....

podpis autora

ROZINEK, M. *Aplikace shlukové analýzy při zpracování biomedicínských dat.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2009. 57 s. Vedoucí semestrální práce Ing. Martin Havlíček

Obsah:

1. TEORETICKÝ ÚVOD 9
1.1 klasifikace objektů 9
1.2 názorný příklad klasifikace 9
1.3 metriky 10
2. METODY KLASIFIKACE 14
2.1 Diskriminační analýza 14
2.1.1 Diskriminační pravidlo 14
2.1.2 Fisherova lineární diskriminační funkce (LDA) 15
2.1.3 Kvadratická diskriminační funkce (QDA) 15
2.1.4 Postup klasifikace diskriminační analýzou 16
2.1.5 Závěrem k diskriminační analýze 17
2.2 Analýza shluků 17
2.2.1 Dendrogram 17
2.2.2 Hierarchické postupy 18
2.2.3 Metoda průměrová: 18
2.2.4 Metoda centroidní: 19
2.2.5 Metoda nejbližšího souseda: 19
2.2.6 Metoda nejvzdálenějšího souseda: 20
2.2.7 Metoda mediánová: 21
2.2.8 Wardova metoda: 21
2.3 Nehierarchické shlukovací metody 22
2.3.1 Shlukování metodou nejbližších těžišť (K-Means) 23
2.3.2 Shlukování metodou středů-medoidů 25
2.4 Postup klasifikace analýzou shluků 26
2.5 Fuzzy shlukování 27
3. DALŠÍ KLASIFIKAČNÍ METODY 31
3.1 logistická regrese 31
3.2 Mapování objektů vícerozměrným škálováním 31
3.3 Korespondenční analýza 31
4. FUNKČNOST METOD 33
4.1 realizace metod 34
4.2 data ke klasifikaci 35
4.3 výběr parametrů 35
4.4 porovnávací test č.1 37
4.5 Porovnávací test č.2 40
4.5.1 Volba dat pomocí entropie 40
4.5.2 Výběr dat podle výsledků klasifikací 41
4.5.3 Prověření metody K-means pro 3 a 4 vstupní parametry 45
5. ZÁVĚR 52
6. SEZNAM 53
7. POUŽITÁ LITERATURA 55

1. TEORETICKÝ ÚVOD

1.1 KLASIFIKACE OBJEKTŮ

Vlastnosti objektů okolo nás nabývají různých hodnot, v případě, že chceme objekty rozdělit do několika skupin, vybereme vlastnosti, které mají na rozdělení do skupin vliv a pomocí těchto vlastností (rysů) je do příslušných skupin rozdělíme. Jak efektivně a správně tyto objekty do skupin rozdělíme zaleží např. na zvolené klasifikační metodě. [1]Klasifikační metody jsou postupy, pomocí kterých se jeden objekt zařadí do existující třídy, nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd, či shluků.

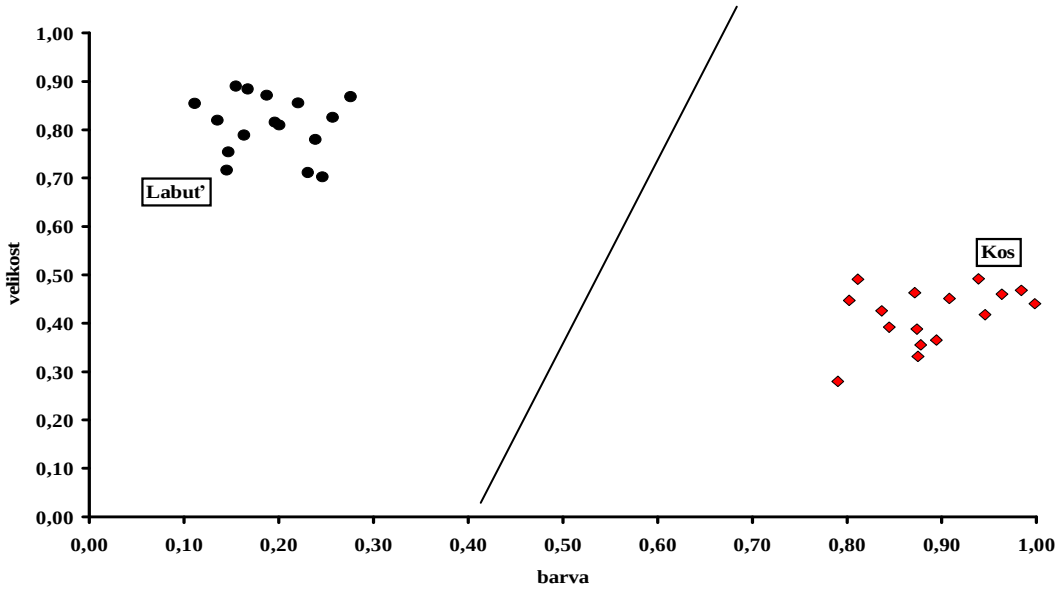
1.2 NÁZORNÝ PŘÍKLAD KLASIFIKACE

Jako příklad, který je jednoduchý si můžeme uvést princip třídění živočichů. Pouhým pohledem člověk pozná, rozdíl mezi dvěma výrazně různými živočichy. V příklad si uvedeme dva z nich. Kos černý (dále jen kos) a labuť velká (dále jen labuť).

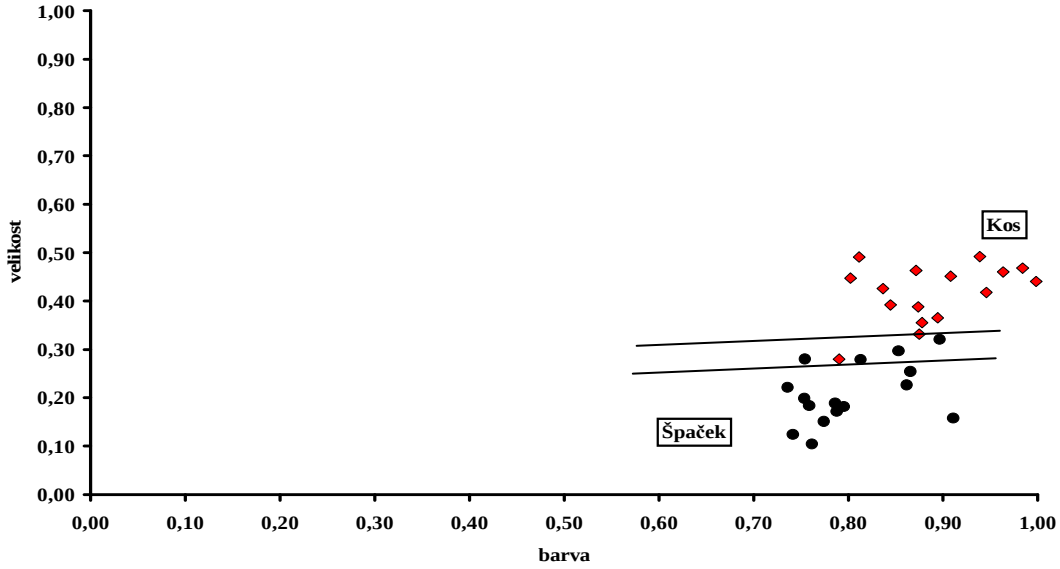
Je na první pohled jasné, že tyto dva objekty se znatelně liší. Kdybychom dostali velké množství těchto „kosů“ a „labutí“, tak je zřejmé, že je dokážeme rozdělit do dvou skupin, na skupinu labutě a na skupinu kosi. Důležité je si uvědomit, podle jakých parametrů těchto objektů jsme je rozdělovali a jakých hodnot tyto parametry nabývali.

Pro posouzení by zcela stačil v tomto případě parametr jeden. Parametrem by mohla být výška ptáka. Podle výšky se téměř bezchybně dají tyto dva druhy rozdělit do dvou skupin. Stejně tak by mohl onen parametr být barva peří, či hmotnost, atd.. Postup klasifikace je komplikovanější tehdy, když jsou si dva objekty podobné a k jejich jednoznačnému rozdělení do skupin je zapotřebí více parametrů. Bylo by tomu tak u dvojice ptáků špaček obecný-kos. Barva již není definitivně rozhodující parametr, spíš je někdy matoucí a kdybychom ho použily jako jediný a výchozí pro klasifikaci, tak bychom zcela jistě ve výsledku ve skupině špačci našli nějaké kosy a naopak.

Dalším parametrem je velikost objektu. Tento parametr je již více vypovídající o tom, do které skupiny lze objekt zařadit, ale přesto není parametr přímo určující příslušnost k určité skupině, proto lze jako klasifikační nástroj použít oba tyto parametry zároveň, což zpřesní výsledek. Vstupní parametry si lze lehce zobrazit v grafu. Jestliže jsou dva, tak se bude jednat o dvojrozměrný graf, kde pro každý objekt je vytvořen bod v prostoru, jehož souřadnice odpovídají velikostem jeho parametrů. Parametry je třeba normovat do intervalu <0;1>, aby měli stejnou váhu. Jako příklad si můžeme zobrazit graf hodnot pro „labuť – kos“ viz. Obrázek 1 a pro „kos – špaček“ viz. Obrázek 2 (od každé třídy je vždy 16 jedinců).



Obrázek 1 - Kos a labuť (dva parametry)



Obrázek 2 – Kos a špaček (dva parametry)

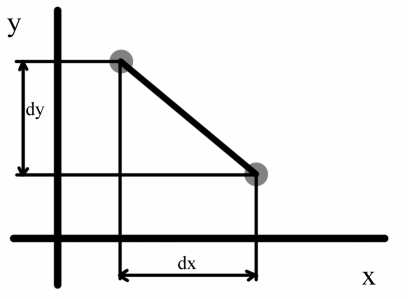
Na druhém obrázku je zřejmé, že nelze jednoduše oddělit od sebe rovnou čarou příslušníky obou skupin, ale lze je rozdělit do skupin tří. V jedné budou špačci, ve druhé budou kosi a třetí skupina bude směsí obou dvou.

1.3 METRIKY

Základem pro klasifikaci je volba způsobu, kterým bude počítána vzdálenost mezi dvěma objekty, nejvhodnější metoda je taková, která je schopna výpočtu vzdálenosti v n-rozměrném prostoru. Jelikož je známá povaha dat (jsou standardizované), tak může být použita některá ze známých metod. Odlišnost metod spočívá v jejich přístupu k vzdálenějším bodů. Výsledkem některé z metod může být odpověď ve formě vzdálenosti, jenž lineárně odpovídá skutečné vzdálenosti, u jiné

může docházet k potlačování vzdálenějších objektů. Volba metody výpočtu vzdálenosti má určitý vliv na celkový výsledek analýzy.

Eukleidovská vzdálenost:



$$D = \sqrt{dx^2 + dy^2}$$

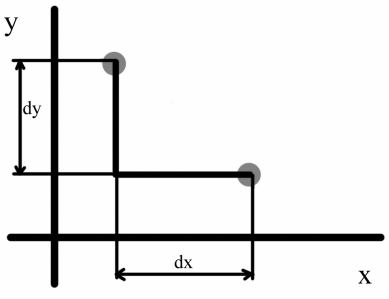
Obrázek 3 – Eukleidovská vzdálenost v 2D

Jedna z nejpoužívanějších. Jde o geometrickou metriku, je v ní aplikovaná Pythagorova věta. Míru představuje velikost přepony pravoúhlého trojúhelníka. Je to nejjednodušší metrika, vzdálenost je nejkratší přímá vzdálenost mezi dvěma objekty. Pro dva body *(x, y)* v n-rozměrném prostoru platí vzorec pro výpočet jejich vzájemné vzdálenosti,

$$D_{EUKLID}(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}; \qquad (1)$$

vedle eukleidovské vzdálenosti se také používá čtverec eukleidovské vzdálenosti, jenž je potom základem Wardovy metody shlukování.

Hammingova vzdálenost:



$$D = |dy| + |dx|$$

Obrázek 4 – Hammingova vzdálenost v 2D[1]

Také často používaná, jinak také nazývaná *manhattanská vzdálenost*, nebo podle jejího principu také *vzdálenost městských bloků*.

$$D_{HAMM}(x,y) = \sum_{j=1}^{n} \left| x_j - y_j \right|; \tag{2}$$

Minkovského metrika:

Zobecněný typ vzdálenosti, figuruje zde koeficient $z \in N$ jenž při vzrůstající hodnotě zvýrazňuje rozdíl mezi vzdálenými objekty. Z rovnice je zřejmé, že pro z =1 jde o Hammingovu vzdálenost a pro z = 2 o Eukleidovskou vzdálenost.

$$D_{MINK}(x,y) = \sqrt[z]{\sum_{j=1}^{n} \left| x_j - y_j^{\,z} \right|}; \tag{3}$$

Chebyshevova vzdálenost:

Chebyshevova vzdálenost je metrika definovaná v prostoru kde vzdálenost mezi dvěma vektory je největší z rozdílů podél jednoho rozměru. Chebyshevova vzdálenost mezi dvěma vektory *x* a *y* činí:

$$D_{CHEBY}(x,y) = \lim_{k \to \infty} \sqrt[k]{\sum_{j=1}^{n} \left| x_j - y_j \right|^k} \tag{4}$$

Mahalanobisova vzdálenost:

Jako jediná z uvedených uvažuje závislost mezi znaky, zahrneme-li do vztahu pro výpočet vzdálenosti i vazby mezi znaky, vyjádřené kovarianční maticí C, dostaneme míru,

$$D_{MAHALA}(x,y) = \sqrt{(x-y)^T C^{-1} (x-y)} \tag{5}$$

, kde jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být ortogonální.

Tětivová vzdálenost:

Je to přímá vzdálenost dvou bodů na povrchu koule s jednotkovým poloměrem a počátkem v těžišti.

$$D_{TĚTIV}(x,y) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}\right)} \qquad (6)$$

Nutno poznamenat, že vzdálenost přestane nabývat imaginárních hodnot teprve když $x, y \geq 1$, tudíž pro běžně standardizované hodnoty v intervalech $x \in <0;1>$ a $y \in <0;1>$ tato vzdálenost nemá smysl.

2. METODY KLASIFIKACE

Klasifikační metody se obecně dělí na dva typy. První typ klasifikační metody je klasifikace, při které známe vstupní data, víme odkud jsou získány a k čemu by měli inklinovat výsledky taková klasifikace je *parametrická*. U druhého typu neznáme původ dat a klasifikační metoda je tedy *neparametrická*.

Klasifikovat lze několika způsoby.

2.1 DISKRIMINAČNÍ ANALÝZA

Diskriminační analýza umožňuje hodnocení rozdílů mezi dvěma nebo více skupinami objektů charakterizovaných více znaky. Obyčejně se dále dělí na techniky, které interpretují rozdíly mezi předem stanovenými skupinami objektů, a techniky, kde je cílem klasifikace objektů do skupin.

Při klasifikaci dochází k porovnávání znaků objektu se znaky ostatních objektů. Na základě rozdílů (či podobností) se objekty rozdělí do skupin, buď subjektivně, nebo na základě zkušeností (kdy sami víme, že když má objekt takové parametry, přesto ho řadíme do jiné skupiny).

2.1.1 Diskriminační pravidlo

Při diskriminační analýze se snažíme vyčíslit hodnotu *diskriminační funkce*, která nám usnadní zařazení do primární třídy. Takto vyčíslené hodnoty funkce používáme také ke třídění nezařazených objektů do předem známých primárních tříd a to na základě diskriminátorů.[2]

Každá primární třída je charakterizována svou funkcí hustoty pravděpodobnosti $f_j(x)$, kde

$$x^T = [x_1, x_2, ..., x_p] \tag{7}$$

Zobecnění diskriminačního pravidla:

Po zobecnění diskriminačního pravidla dostaneme vztah pro zařazení vektoru x do třídy y_1, když y_1 je třída objektů s normálním rozdělením a střední hodnotou μ_1 a y_2 je obdobná třída se střední hodnotou μ_2 a za předpokladu, že kovarianční matice obou tříd jsou shodné a označíme je C.

$$(\mu_1 - \mu_2)C^{-1}\left(x - \frac{\mu_1 + \mu_2}{2}\right) > 0; \tag{8}$$

2.1.2 Fisherova lineární diskriminační funkce (LDA)

$$z_i = a_{i1}x_1 + a_{i2}x_2 + ... + a_{ip}x_p \tag{9}$$

p … je počet proměnných primárních tříd, čili počet diskriminátorů
x … jsou standardizované hodnoty těchto proměnných
z …standardizované klasifikační koeficienty Fischerovi diskriminační funkce
Jedná se o prakticky nejvíce využívanou

2.1.3 Kvadratická diskriminační funkce (QDA)

Jsou-li střední hodnoty dvou souborů stejné, ale kovarianční matice jsou různé, tak již nelze dále používat LDA. Pravidlo pro zařazení do první skupiny $f_1(x)\,\pi_1 > f_1(x)\,\pi_2$ vede ke kvadratické nerovnici:

$$x^T G x + h^T x + C > 0 \tag{10}$$

kde matice

$$G = 0{,}5(C_2^{-1} - C_1^{-1}) \tag{11.1}$$

Vektor

$$h^T = \mu_1^T C_1^{-1} - \mu_2^T C_2^{-1} \tag{11.2}$$

a konstanta

$$C = 0{,}5 \ln \frac{\det C_2}{\det C_1} - \left(0{,}5\mu_1^T C_1^{-1}\mu_1 - \mu_2^T C_2^{-1}\mu_2\right) - \ln\left(\frac{\pi_2}{\pi_1}\right) \tag{11.3}$$

Logistická diskriminace

Fisherova lineární diskriminace je optimální, když dva soubory mají vícerozměrné normální rozdělení se stejnými kovariančními maticemi. Tato diskriminační funkce se jeví také dostatečně robustní na odchylky od normality. Existuje však řada případů silné nenormality, např. přítomnost binárních proměnných[2]. V této chvíli se použije logistické diskriminace.

2.1.4 Postup klasifikace diskriminační analýzou

1. **Bodové odhady parametrů polohy a rozptýlení všech diskriminátorů:**
   Vyčíslí se :
   - Aritmetické průměry ve třídách
   - Směrodatné odchylky
   - Celková korelační a kovarianční matice všech diskriminátorů
   - Mezitřídní korelace a kovariance za použití průměru místo hodnot objektů
   - Vnitrotřídní korelace a kovariance za použití dat ve kterých byli odečteny průměry tříd a provede se shodnocení dosažených výsledků
2. **Vyšetření vlivu jednotlivých diskriminátorů:**
   Vliv jednotlivých diskriminátorů na výsledky diskriminační analýzy se sleduje pomocí statistik při odstranění odpovídajícího diskriminátoru
3. **Odhady neznámých parametrů b_0, b_1, … , b_p lineární diskriminační funkce pro každou třídu:**
   Odhady neznámých parametrů b_0, b_1, … , b_p jsou mezivýpočtem k vyčíslení diskriminačního skóre
4. **Odhady regresních parametrů b_0, b_1, … , b_p lineárního regresního modelu pro každou třídu:**
   Těmito regresními parametry predikované hodnoty budou ležet mezi nulou a jedničkou. Zařazení se provede na základě třídy s nejvyšším skóre, blízkým jedničce
5. **Klasifikace objektů diskriminační funkcí(dělení do tříd):**
   Provede se:
   - Vyčíslení klasifikačních počtů objektů v jednotlivých třídách po diskriminaci do tříd
   - Přehled chybně klasifikovaných objektů tak, že vedle skutečné třídy je predikována třída a procento pravděpodobnosti výskytu objektu v predikované třídě
   - Přehled klasifikovaných objektů – skutečná primární třída predikovaná třída všech objektů a procento pravděpodobnosti výskytu objektu v dané třídě
6. **Kanonická korelační analýza:**
   - Analýza kanonických proměnných (první soubor obsahuje diskriminátorů a druhý soubor třídní proměnné)
   - Odhady parametrů u kanonických proměnných
   - Kanonické proměnné u třídních průměrů
   - Standardizované kanonické koeficienty slouží k výpočtu kanonického skóre, což jsou vážené průměry objektů
   - Korelace původních a kanonických proměnných představuje korelace původních proměnných na kanonické proměnné. Tím se usnadní vysvětlení dotyčné kanonické proměnné.
7. **Lineární diskriminační skóre všech objektů:**
   Jsou vyčísleny hodnoty předikovaných skóre lineárních diskriminačních proměnných pro všechny objekty.
8. **Regresní skóre všech objektů:**
   Hodnoty předikovaných skóre regresních proměnných pro všechny objekty jsou založeny na regresních koeficientech.

9. **Kanonické skóre:**
   Hodnoty předikovaných skóre kanonických proměnných pro všechny objekty jsou založeny na kanonických koeficientech
10. **Volba proměnných:**
    Z velké palety diskriminátorů se vybírají pouze ty, které jsou dostatečně účinné, maximálně 8 proměnných. Výběr se provádí krokově: k nejlepšímu diskriminátorů se nalezne druhý nejlepší tak, že se prověří,zda diskriminace bude tak dokonalá, jako když byl jeden diskriminátor odebrán. U nové proměnné se ověřuje zda její ***F***má hodnotu pravděpodobnosti menší než $\alpha = 0{,}05$.
11. **Výklad grafů:**
    Výsledkem diskriminační analýzy je grafické zařazení do tříd. Zobrazení se provede na třech grafech:
    - Zobrazení lineárních diskriminačních skóre
    - Zobrazení regresního skóre
    - Zobrazení kanonického skóre

2.1.5 Závěrem k diskriminační analýze

Důležité je si uvědomit, k čemu slouží. DA stanový postup pro rozřazení objektů do tříd na základě jejich získaného skóre. Jejím úkolem je stanovit, v čem se objekty liší. Jejím přínosem je diskriminační funkce, která dokáže nově příchozí objekty zařadit podle pravidla přímo do jedné z výstupních skupin. Problém u této analýzy nastane, když více objektům chybí řada jejich proměnných, potom na výsledek má hlavně vliv jen sada objektů, které mají definovanou většinu proměnných.

[1]Diskriminační analýza je značně citlivá na poměr velikosti výběru a počtu diskriminátorů. Existuje řada studií, v nichž je navrhován poměr 20 objektů na jeden diskriminátor. I když je v experimentální praxi obtížné tento poměr dodržet, musíme si uvědomit, že výsledky diskriminační analýzy budou nestabilní, když bude poměr velikosti výběru a počtu diskriminátorů malý.

2.2 ANALÝZA SHLUKŮ

Analýza shluků (Cluster analysis, CLU) patří mezi metody, které se zabývají vyšetřováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili shluků. Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Podle způsobu shlukování se postupy dělí na hierarchické a nehierarchické. Hierarchické se dělí dále na aglomerativní a divizní.

2.2.1 Dendrogram

V našem případě se používá dendrogram pro zobrazení postupného vývoje výsledného shluku. Zachycuje, jak se jednotlivé objekty mezi sebou propojili a jaká vzdálenost mezi objekty danému propojení odpovídala.

Je to vývojový strom, diagramatické znázornění vývojových vztahů taxonů organismů formou větvícího se stromu. Termín je odvozen od řeckého slova dendron což znamená strom. U obvyklého typu dendrogramu znázorňuje osa y absolutní nebo relativní čas, v našem případě vzdálenost, jednotlivé větve znázorňují samostatné vývojové linie, jejich větvení zobrazuje štěpení vývojových linu, společný kmen představuje vývoj (linií) společného předka. Dendrogram může být podobný i fenogram, který však ukazuje pouze stupeň matematicky hodnocené fenotypické podobnosti. [4]

2.2.2 Hierarchické postupy

Jsou založeny na postupném spojování objektů a jejich shluků do dalších, větších shluků. Nejprve se vypočte základní matice vzdáleností mezi objekty. U *aglomerativního shlukování* se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako celek. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. *Inverzní postup* je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů.

Výhodou hierarchických metod je nepotřebnost informace o optimálním počtu shluků v procesu shlukování; tento počet se určuje až dodatečně. Při shlukování vznikají dva základní problémy:

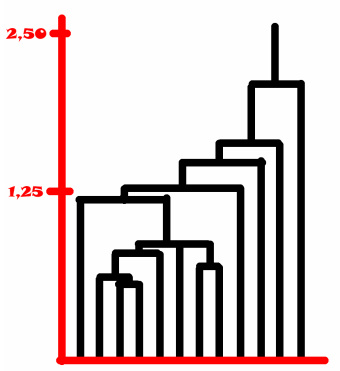
1. *Způsob měření vzdáleností* mezi objekty. I když existuje celá řada měr vzdáleností (vícerozměrných metrik), nejčastěji se užívá euklidovská metrika, která je přirozeným zobecnením běžného pojmu vzdálenosti.
2. Volba vhodné *shlukovací procedury* dle zvoleného způsobu metriky.

Shlukovací procedury:

2.2.3 Metoda průměrová:

Vzdálenost dvou shluků se počítá jako průměr z mezishlukových vzdáleností dvou objektů, kdy se mezishluková vzdálenost objektů je vzdálenost dvou objektů, z nichž každý patří do jiného shluku. Nejbližší jsou shluky, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jednoho a všemi objekty druhého shluku.

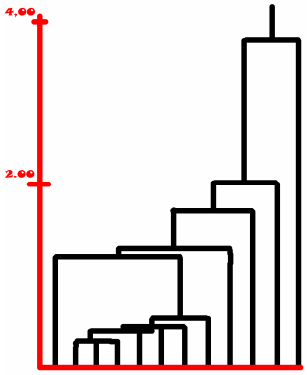
Výsledek je podobný metodě *nejvzdálenějšího souseda*, pouze spojení je provedeno při obvykle vyšších vzdálenostech.



Obrázek 5 – typický dendrogram průměrové metody

2.2.4 Metoda centroidní:

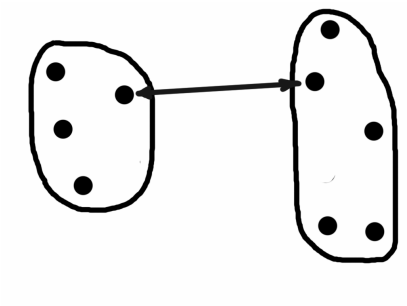
Vzdálenost shluků se počítá jako euklidovská vzdálenost jejich těžišť. Nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi těžišti.



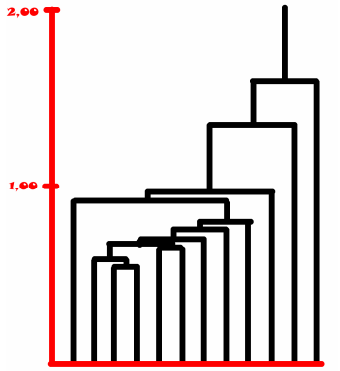
Obrázek 6 – typický dendrogram centroidní metody

2.2.5 Metoda nejbližšího souseda:

Kritériem pro vytváření shluků je minimum z možných mezishlukových vzdáleností objektů. Metoda tvoří nový shluk na základě nejkratší vzdálenosti mezi shluky (či objekty) a neumí proto rozlišit špatně separované shluky. Je zde silná tendence ke tvorbě řetězců. Jsou-li objekty na opačných koncích řetězce zcela nepodobné, může řetězování vést k jejich spojení. Metoda se tak může dostat k zcela milným závěrům. Na druhé straně je to jedna z mála metod, která umí roztřídit a rozlišit i neeliptické shluky.(Obrázek 3)



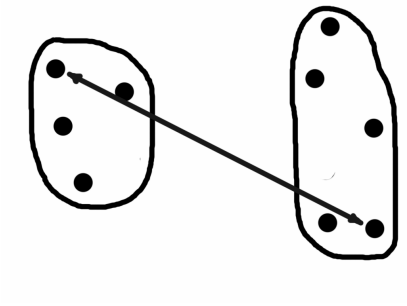
Obrázek 7 – Nejbližší soused



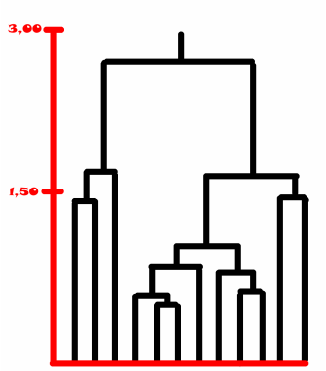
Obrázek 8 – typický dendrogram metody NN

2.2.6 Metoda nejvzdálenějšího souseda:

Počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů. Vzdálenost (či nepodob-nost) mezi shluky je určována vzdáleností (či nepodobností) mezi dvěma nejvzdálenějšími objekty, každý přitom je z jiného shluku. Proto všechny objekty ve shluku jsou klasifikovány na základě maximální vzdálenosti či minimální podobnosti vůči objektům ve druhém shluku.



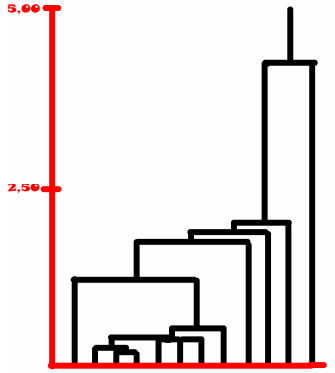
Obrázek 9 – nejvzdálenější soused



Obrázek 10 – typický dendrogram metody FN

2.2.7 Metoda mediánová:

Jde o jisté vylepšení centroidní metody, neboť se snaží odstranit rozdílné váhy, které centroidní metoda dává různě velkým shlukům.

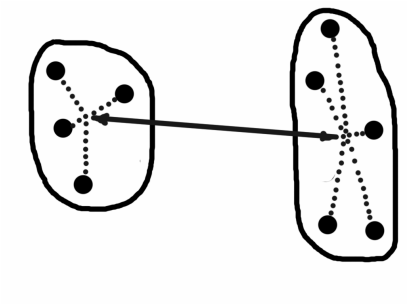


Obrázek 11 – Typický dendrogram mediánové metody

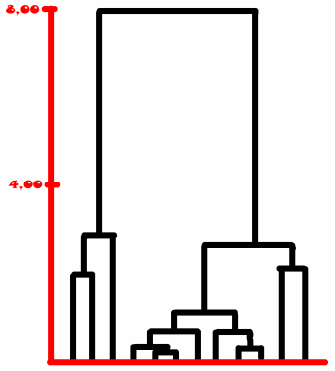
2.2.8 Wardova metoda:

Založena na minimalizaci ztráty informace při spojení dvou tříd. V každém kroku je uvažován takový možný pár objektů (či shluků), aby suma čtverců odchylek od střední hodnoty VSS (Rovnice 1) dosáhla při vzniku shluku svého minima.

$$VSS = \sum_{j=1}^{m} \sum_{i=1}^{k} \left(x_{ij} - \overline{x_j} \right)^2 ; \qquad (11)$$



Obrázek 12 – Wardova metoda



Obrázek 13 – typický dendrogram Wardovy metody

2.3 NEHIERARCHICKÉ SHLUKOVACÍ METODY

U Metody „Seed“ uživatel na základě svých věcných znalostí určí, které objekty mají být “typickými” představiteli nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů. V nehierarchických shlukovacích metodách je počet shluků obvykle předem dán, i když se v průběhu výpočtu může změnit. Zůstává-li počet shluků zachován, hovoříme o *nehierarchických metodách s konstantním počtem shluků*, v opačném případě o *nehierarchických metodách s optimalizovaným počtem shluků*. Nehierarchické metody zahrnují dvě základní varianty - optimalizační metody a analýzu módů, medoidů. Optimalizační nehierarchické metody hledají optimální
rozklad přeřazováním objektů ze shluku do shluku s cílem minimalizovat nebo maximalizovat nějakou charakteristiku rozkladu. Metody, označované jako analýza módů, medoidů, představují hledání rozkladu do shluků, kde shluky jsou chápány jako místa se zvýšenou koncentrací objektů v m-rozměrném prostoru proměnných. Místo výchozí matice vzdáleností může být v některých případech ke shlukování použita i korelační matice.

Těmto nehierarchickým metodám se také říká *K-means shlukování*. Pro takové shlukování se je použitelný jeden za tři postupů.

Sekvenční práh

Nejdříve se zvolí zárodek 1. shluku a všechny objekty ve specifikované vzdálenosti okolo se zahrnou do shluku. Když je hotovo, zvolí se zárodek 2. shluku a opět se objekty ve specifikované vzdálenosti zahrnou to druhého shluku. Pak je vybrán 3. a postup se opět opakuje. Jakmile je objekt shluknut se zárodkem, již se s ním dál nepočítá.

Poznámka: Tento postup je vhodný pro velké datové soubory. Uživatel může specifikovat nové zárodky výpočtem po každém přiřazení objektu.

Paralelní práh

Metoda se liší od původní tím, že původní zárodky jsou vybrány současně v jedné chvíli a pak jsou objekty uvnitř prahové vzdálenosti zahrnovány do shluku k nejbližšímu zárodku. V průběhu vývoje procesu lze nastavit prahovou vzdálenost tak, aby pohlcovala do shluku více či méně objektů. Může se stát, že některé objekty ležící mimo dosah prahové vzdálenosti nebudou do shluků vůbec zahrnuty.

Optimalizace

Optimalizační metoda se liší od předchozích tím, že najde-li mechanismus objekt který má nejkratší vzdálenost od shluku A, ale tento objekt se již nachází ve shluku B, je ze shluku B vytržen a zařazen do shluku A.

2.3.1 Shlukování metodou nejbližších těžišť (K-Means)

Při vytváření malého počtu shluků z velkého počtu objektů se jeví nejužitečnější shlukovací metodou. Vyžaduje spojité proměnné a především bez odlehlých hodnot. Diskrétní data mohou být rovněž analyzována, ale mohou způsobit problémy. Princip metody spočívá v rozdělení **n** objektů o **m** proměnných do **k** shluků tak, že mezishluková suma čtverců je přitom minimalizována. Jelikož počet možných uspořádání je enormě veliký, není praktické očekávat vždy nejlepší řešení. Algoritmus nalezne spíše optimum lokální než globální. Je to takové uspořádání shluků, kdy již přemístění objektu z jednoho shluku do druhého nezpůsobí snížení sumy čtverců. Algoritmus pracuje opakovaně, startuje vždy z jiného počátečního uspořádání. Nakonec vybere optimální řešení ze všech možných dosažených uspořádání shluků. Uživatel zadává počet shluků, jež mají být nalezeny. Pak jsou vytvořeny prostorové shluky nalezením souboru středů shluků tak, že každý objekt je přiřazen do jednoho shluku, načež jsou určeny nové shluky a celý proces se opakuje.

Předpokládejme n objektů rozdělených do k shluků. Pak k-tý shluk obsahuje **nK** objektů. Každý objekt je v jednom řádku popsán m proměnnými. Chybějící hodnota i-té proměnné v j-tém řádku u k-tého shluku je označena δ_{ijk}. Data **x** i **j** jsou předem standardizována a označena z_{ij}. Počáteční přiblížení ovlivňuje konečné uspořádání shluků. Proto algoritmus pro každý pokus zcela náhodně přiřazuje každý objekt jednomu shluku. Toto uspořádání je pak optimalizováno. Pokus nastartovat proces z rozličných náhodných uspořádání vysoko zvýší pravděpodobnost nalezení globálního optima počtu shluků.[2]

Kritérium věrohodnosti: jde o kritérium těsnosti proložení, které je založeno na srovnání rozličných konfigurací shluků a vychází z mezishlukové sumy čtverců WSS_K definované vztahem

$$WSS = \frac{nm}{nm-m}\sum_{l=1}^{k}\sum_{i=1}^{m}\sum_{j=1}^{n_k}\left(1-\delta_{ijl}\right)\left(y_{ij}-c_{il}\right)^2 \qquad (12)$$

kde *c* je střední hodnota (průměr) *i*-té proměnné v *k*-tém shluku. Procento variace PV_K (proměnlivosti) je definováno vztahem:

$$PV_K = 100\%\frac{WSS_K}{WSS_1} \qquad (13)$$

K-means algoritmus slouží z zařazení bodů v n-rozměrném prostoru k příslušným shlukům, neboli rozdělení bodů do K tříd. V algoritmu se volí středy tak, aby se minimalizovala vzdálenost změřená nějakou metrikou, obvykle euklidovskou. Vzdáleností je myšlena vzdálenost bodu v prostoru od středu. Nutné je znám předem počet tříd, do kterých se budou body klasterizovat. K-means probíhá iterativně. Na počátku jsou středy rozmístěny v prostoru (obvykle náhodně) a je určena příslušnost trénovacích bodů k těmto středům. Ze souřadnic bodů příslušných ke stejné třídě je pro každou dimenzi spočítán aritmetický průměr. Bod který dostaneme je jakousi analogií k těžišti dané třídy. Přesunutím středu do tohoto těžiště se minimalizuje metrika konkrétní třídy (pro žádné jiné umístění středu nemůže být součet vzdáleností jednotlivých bodů třídy od tohoto středu menší). Tak končí první iterace, v každé další iteraci je opět řešena příslušnost trénovacích bodů k nově posunutým třídám a středy tříd jsou podle bodů znovu opraveny. Algoritmus končí, pakliže se během předešlé iterace neposunul střed žádné třídy, což znamená, že v příští iteraci se nezmění třída žádného trénovacího bodu.

Algoritmus skončí, protože existuje jen konečný počet přiřazení bodů ke středům. Jelikož každá další iterace musí výsledek vylepšit, nebude se žádná konfigurace opakovat. Musí tedy dojít k ustálení na nějaké konfiguraci v konečném čase. V praxi se často podmínka ukončení algoritmu modifikuje, čímž se dospěje rychleji k výsledku, sice různě nepřesnému. Nepřesností se ovšem, pokud pracujeme s čísly v plovoucí řádové čárce, musíme v počítačové implementaci stejně dopustit. Jednou variantou může být ukončení algoritmu v případě, že se během předchozí iterace žádný ze středů neposunul o více, než o dané e > 0. Zadáním e určujeme požadovaný řád chyby. Další variantou je ukončení algoritmu přesně po provedení daného počtu iterací. V tomto případě nemáme zaručený řád chyby, ale místo toho známe dobu provádění algoritmu. Konečně můžeme oba přístupy zkombinovat a ukončit při dosažení požadované přesnosti ale nejpozději po daném počtu iterací. Přestože K-means algoritmus musí skončit, nemusí dojít ke globálnímu optimu, naopak často končí pouze u lokálního optima. Toto je zapříčiněno požadavkem na počáteční rozmístění středů tříd. Najít takové rozmístění středů , které povede ke globálnímu optimu, je zpravidla obtížnější než algoritmus K-means samotný, proto se obvykle volí iniciální souřadnice středů náhodně. Pokud spustíme algoritmus opakovaně, přičemž pokaždé se bude volit jiné náhodné počáteční rozmístění středů, můžeme pokaždé dospět k jinému výsledku. Jednou metodou pro zpřesnění

výstupu z K-means je proto opakované spuštění algoritmu a následné vybírání konfigurace s nejlepším výsledkem, např. nejnižším součtem odchylek všech bodů od středů jejich příslušných tříd. Při náhodném výběru souřadnic středů se také nabízí několik možností. Pokud známe horní a dolní meze pro souřadnice ve všech dimenzích, je možné každou souřadnici zvolit jako náhodné číslo z příslušného intervalu. Pokud však trénovací body nejsou rozmístěny příliš rovnoměrně, může se stát, že některý navrhovaný střed padne do takového místa prostoru, že nebude pro žádný trénovací bod nejbližší (bude se tedy jednat o prázdnou třídu). Druhou variantou hledání náhodných středů je vybrat pro každý střed náhodně jeden bod z trénovací množiny a umístit ho na jeho souřadnice. Tím zřejmě nevzniknou prázdné třídy, ale navrhované středy se budou pro změnu kumulovat v místech s vysokou hustotou bodů . Pro implementaci v rámci této práce byla zvolena druhá varianta, která navíc
nepotřebuje znát meze jednotlivých souřadnic. V popisu algoritmu K-means není definováno, jak se má postupovat v případě prázdných tříd. Pakliže ponecháme střed prázdné třídy na svém místě (protože není ze kterých bodů vypočítat aritmetický průměr souřadnic), zůstane třída prázdná nejspíš až do konce algoritmu. To ale zhorší výslednou hodnotu metriky, kterou K-means minimalizuje.

2.3.2 Shlukování metodou středů-medoidů

Medoid, čili střed shluku, je střední objekt, pro který platí, že průměrná vzdálenost k ostatním objektům v tomto shluku je minimální. Je-li požadováno *k* shluků, bude existovat také *k* medoidů. Po nalezení medoidů jsou data klasifikována do shluků vždy okolo nejbližšího medoidu. Medoidy a shluky se vytvářejí na základě *vzdáleností* čili *nepodobností* (dissimilarities).

Proměnné shluku

Druhů proměnných je celá řada: *Intervalové* jsou spojité kladné či záporné, v lineární škále, např. výška, hmotnost, cena, teplota, čas atd. *Ordinální* jsou pořadová čísla stupnice, hodnotící nějakou vlastnost, např. silný nesouhlas (5), nesouhlas (4), neutrální (3), souhlas (2) a silný souhlas (1). *Poměrové* jsou kladné hodnoty, kdy vzdálenost mezi dvěma čísly je stejná, když i jejich poměr je stejný, např. mezi 3 a 30 je stejná jako mezi 30 a 300, chemická koncentrace, intenzita záření, absorbance atd. *Nominální* jsou proměnné, které vyjadřují pouze kvalitu a nikoliv kvantitu, např. PSČ, rasa, barva, název města atd. Symetrické binární: mají dvě možnosti, obvykle ano (1), ne (0). Asymetrické binární: přítomnost či nepřítomnost zřídka se vyskytující události, kdy nepřítomnost není tak důležitá, např. osoba má jizvu na tváři, a tím je lépe identifikovatelná.

Späthova metoda

Metoda minimalizuje účelovou funkci přemísťováním objektů z jednoho shluku do druhého. Začíná u počátečního uspořádání shluků, algoritmus pak najde lokální minimum inteligentním přesouváním objektů ze shluku do shluku. Jakmile se nepřemístí už žádný objekt, metoda terminuje. Lokální minimum však nemusí být globálním. Aby program překonal toto omezení, zopakuje se několikrát hledání vždy

z jiného startovacího uspořádání a nejlepší uspořádání shluků je nakonec bráno za výsledné. Za účelovou funkci se bere celková vzdálenost mezi všemi objekty ve shluku podle vzorce $D = \sum_{l=1}^{k} \sum_{i \in c_k} \sum_{i \in c_k} d_{ij}$, kde *K* je celkový počet shluků, d_{ij} je vzdálenost mezi *i*-tým a *j*-tým objektem a c_K je soubor všech objektů ve shluku *k*.

Metoda PAM (Partition Around Medoids)

Algoritmus se opět pokouší minimalizovat celkovou vzdálenost *D* ve dvou krocích:

1. Nalezne se reprezentativní soubor *k* objektů. První objekt má nejkratší vzdálenost ke všem ostatním objektům, čili představuje *střed*. Pak se *k-1* objektů nachází tak, že hodnota *D* je co možná nejmenší.

2. Možné alternativy polohy *k* objektů jsou vybírány iteračním způsobem. Algoritmus vyhledává dosud nezařazené objekty a přemísťuje je tak, aby se hodnota *D* snižovala. Iterace skončí, jakmile změny nezpůsobí další snížení hodnoty *D*.

Silueta

poskytuje klíčovou informaci o dobrém a špatném shluku. Hodnota siluety *s* se vypočte postupem:

1. Objekt *i* je ve shluku A a má průměrnou vzdálenost *a* ke všem objektům ve svém shluku. Je-li ve shluku A jediný objekt, je *a* = 0.

2. Sousední shluk B obsahuje objekty, které jsou nejbližší k objektu *i* ve shluku A a *b* je průměrná vzdálenost mezi objektem *i* a všemi objekty ve shluku B.

3. *Silueta s* objektu *i* se vyčíslí následovně: když shluk A obsahuje pouze jeden objekt, je *s* = 0. Když *a* < *b, je s = 1 - a/b.* Když *a*> *b*, je *s = b/a - 1.* Když *a* = *b*, je *s* = 0. Silueta se vyčíslí pro každý objekt. Hodnota siluety se mění od -1 do +1 a je mírou úspěšné klasifikace do shluků při porovnání vzdáleností uvnitř shluku A se všemi vzdálenostmi objektů nejbližšího souseda B dle pravidla:

1. Je-li *s* blízko +1, objekt *i* je dobře klasifikován do shluku A, protože jeho vzdálenosti k ostatním objektům v tomto shluku jsou podstatně kratší než vzdálenosti k objektům nejbližšího souseda B.

2. Je-li *s* blízko nule, objekt *i* se nachází kdesi uprostřed mezi shluky A a B, a čistě náhodou byl přiřazen do shluku A.

3. Je-li *s* blízko -1, objekt *i* je špatně klasifikován. Vzdálenosti k ostatním objektů ve svém shluku jsou mnohem větší než vzdálenosti k objektům nejbližšího souseda B. Otázkou pak je, proč byl vlastně zařazen do shluku A.[2]

2.4 POSTUP KLASIFIKACE ANALÝZOU SHLUKŮ

1. **Volba databáze:**
   Zadává se typ dat:
   - Proměnných sloupců analyzovaných objektů(řádků)
   - Sloupců matice vzdáleností
   - Sloupců korelační matice
2. **Volba druhu veličin:**
   Zadává se typ užitých veličin v datech která mohou být

- Intervalová
- Ordinální
- Nominální
- Symetrická binární
- Asymetrická binární
- poměrová

3. **Název objektů:**
   Zadání pojmenování jednotlivých objektů umístěných v řádcích, které se mohou objevit v dendrogramu místo indexů(pořadových číslel objektů
4. **Typ shlukovací techniky:**
   Volby metody z možností:
   - Průměrová
   - Centroidní
   - Nejbližšího souseda
   - Nejvzdálenějšího souseda
   - Mediánová
   - Wardova
   - Flexibilní, atd..
5. **Druh užité vzdálenosti:**
   - Euklidova
   - Hammingova
   - Minkovského
   - Mahalanobisova, atd..
6. **Postup linkování a řazení do shluků:**
   Tabelární výpočet vzdáleností (podobností) mezi objekty a shluky a postupné vytváření dendrogramu
   - Metoda hierarchického shlukování
   - Metoda nejbližších středů
   - Metodou středů (medoidů)
   - Fuzzy shlukování
7. **Výpočet skutečných vzdáleností v dendrogramu:**
   Jsou porovnány skutečné vzdálenosti mezi objekty a vypočtené vzdálenosti v dendrogramu jejich rozdíl a konečně i procentuální vyjádření tohoto rozdílu
8. **Zhodnocení úspěšnosti metody:**
   Porovnání výsledku s estimovaným předpokladem, kontrolní procedury, jenž zjišťují, zda byla zvolena správné metoda přístupu využívající výstupního dendrogramu

2.5 FUZZY SHLUKOVÁNÍ

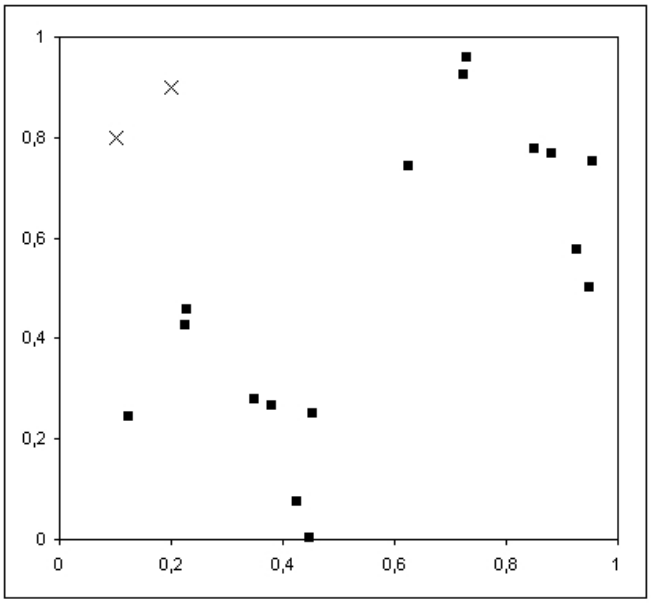
Fuzzy shlukování zobecňuje všechny shlukovací metody tím, že umožňuje shlukování jednoho objektu do více než jednoho shluku, zatímco v běžném shlukování je každý objekt členem pouze jednoho shluku. Předpokládejme, že máme K shluků a budeme definovat soubor proměnných $m_{i1}, m_{i2}, ..., m_{iK}$, které představují pravděpodobnost, že objekt i je klasifikován do k-tého shluku. V běžném shlukovacím algoritmu je jedna z těchto proměnných rovna jedné a zbytek roven

nule. To představuje skutečnost, že takový algoritmus klasifikuje každý objekt do jednoho a právě jednoho shluku.

Ve fuzzy shlukování je “účast objektu”, čili přítomnost objektu, rozdělena do všech shluků. Proměnná *m* může zde být rovna 1 nebo 0 a suma těchto hodnot musí být rovna 1. Nazveme tento proces *fuzzifikací shlukové konfigurace.* Proces má výhodu, že nenutí objekt aby byl zařazen pouze do jediného specifického shluku. Nevýhodou však je, že se zde objevuje mnohem více informací, které musí být vysvětleny.

Pomocí iteračních metod se najdou středy předem daného počtu shluků, tedy pokud je cílem analýzy obdržet například 3 shluky, tak počáteční iterací jsou 3 body v n-rozměrné prostoru, které se iterací, v závislosti na zvolené metodě přibližují centrům shluků. Za centra se považují místa vypočtená například mediánem. Po „dostatečném“ přiblížení, což je tehdy, kdy rozdíl mezi iterací (n) a iterací (n+1), nepřesahuje předem stanovené hodnoty v podobě minimální chyby. Iterací rozumíme souřadnice nových center.

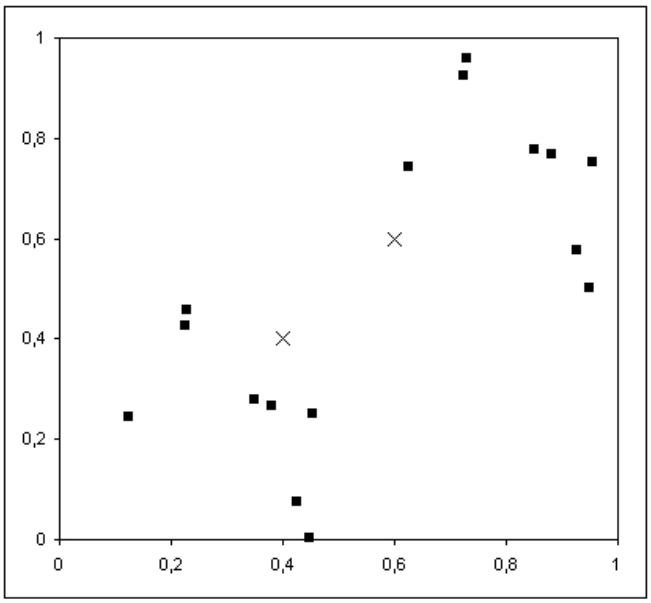
Postup který je zde vysvětlen lze pro názornost ukázat na příkladu.



Obrázek 14 – Počáteční iterace (nultá)

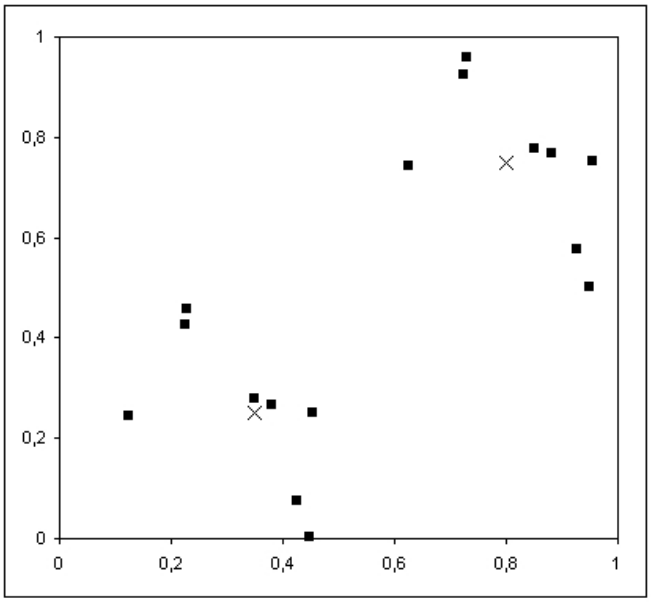
V grafu (Obrázek 14) jsou ve dvojrozměrném prostoru zobrazeny body (čtverečky) které představují jednotlivé jedince a pro demonstraci je již okem patrné, že tvoří dva shluky. Jako počáteční iterace jsou dva body (označené křížkem), což znamená, že cíle je separovat jedince do dvou výsledných shluků.

Na počáteční iteraci tolik nezáleží, nemá zásadní vliv na počet výsledných iterací potřebných k nalezení center.



Obrázek 15 – 2. iterace

V dalších iteracích se body přibližují centrům jednotlivých shluků.

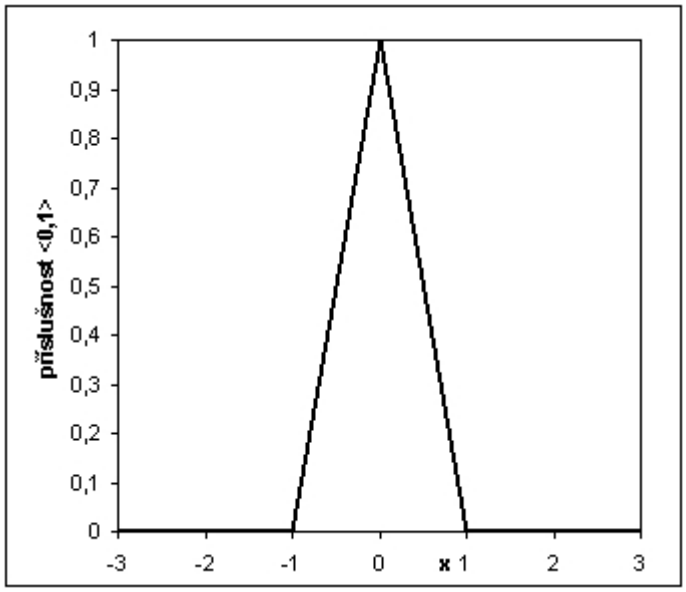


Obrázek 16 – n-tá iterace

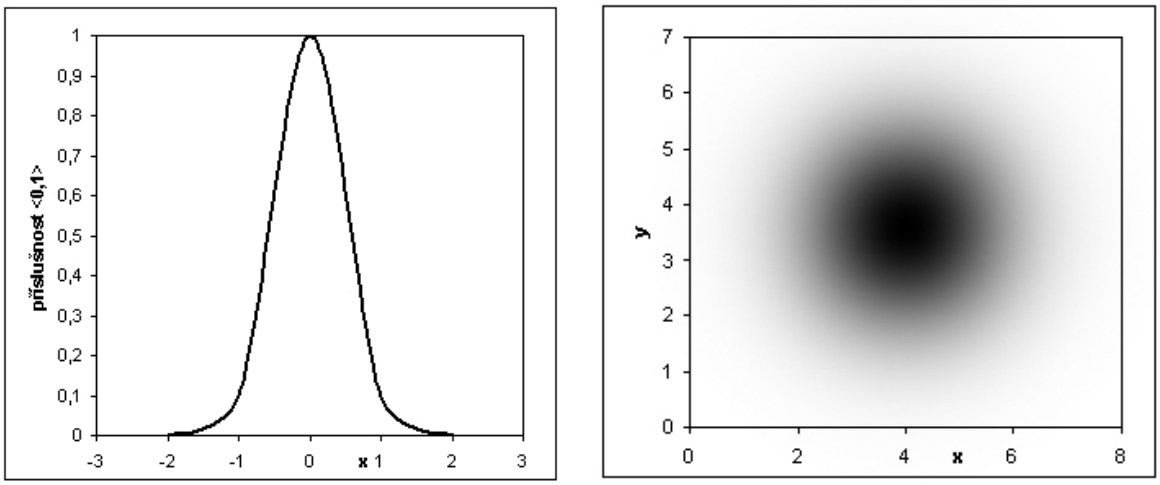
V n-té iteraci, kdy se „křížky“ posunuly do center dvou shluků s předepsanou přesností, další iterace už nemají zásadní vliv na výsledek.

Samotná konstanta, jenž určuje minimální přesnost výpočtu se volí v závislosti na povaze dat. To znamená, že je vhodné aby rozestupy (vzdálenosti) jednotlivých jedinců ve shluku byli o nejméně řád větší, nežli samotná konstanta. Tím se zamezí tomu, aby na základě nepřesnosti jedinci spadali do jiných shluků, nežli by náleželi při větší přesnosti.

Poté, co jsme obdrželi dvě centra shluků, zbývá jen přidělit jednotlivým jedincům příslušnost k nim.



Obrázek 17 – Funkce příslušnosti



Obrázek 18 – Funkce příslušnosti

(vlevo) pro 2D prostor zobrazena příslušnost ve stupních šedi

Ve fuzzy množinách existuje nepřeberné množství různých funkcí příslušnosti k dané množině. Některé jsou v podobě gassova klobouku, jiné mohou mít nespojitý průběh. V samé podstatě závisí na uživateli, jaký průběh funkce si zvolí pro daný problém.

3. DALŠÍ KLASIFIKAČNÍ METODY

3.1 LOGISTICKÁ REGRESE

Logistická regrese byla navržena v 60tých letech jako alternativní postup k metodě nejmenších čtverců pro případ, že vysvětlovaná proměnná je binární. Vhodná pro výpočty rizika vzniku chorob, přičemž jako vstupní data slouží věk, životospráva, krevní tlak, teplota a mnoho dalších parametrů.

Analýza logistickou regresí se podobá DA. Lze ji požít v případě jak spojitých , tak diskrétních proměnných. Výsledný model může být jako u předchozích metod použit k přímému klasifikování.

Dělí se na:

1. binární logistickou regresi
2. nominální logistickou regresi
3. ordinální logistickou regresi

3.2 MAPOVÁNÍ OBJEKTŮ VÍCEROZMĚRNÝM ŠKÁLOVÁNÍM

Vícerozměrné škálování (**M**ulti**D**imensional **S**caling, MDS) je technika vytvoření diagramu relativního umístění objektů v rovině dvojrozměrného grafu na základě dat vzdáleností mezi objekty, tzv. *matice proximity* (blízkosti). Diagram může obsahovat jeden, dva, tři a zřídka i více rozměrů, dimenzí. Technika vyčíslí metrické klasické (CMDS) nebo nemetrické (NNMDS) řešení a vychází buď přímo z experimentálních hodnot ***X***, z korelační matice ***R*** nebo z matice podobností ***S*** či vzdáleností ***D***. Vzdálenost mezi oběma objekty je Eukleidovská, počítaná na základě Pythagorovy věty, $d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2}$, kde *m* je počet proměnných a x_{ik} jsou data *i*-tého řádku a *k*-tého sloupce. I když vynášíme vzdálenosti do dvojrozměrného grafu, může být d_{ij} vyčísleno na základě většího počtu proměnných $m \geq 2$. Matice vzdáleností je potom trojúhelníková a zajímá nás jenom její horní část. S růstem objektů však roste i počet dimenzí, takže pro *tři* objekty je to *dvoj*-rozměrná rovina, pro *čtyři* objekty pak *troj*-rozměrný prostor atd.[1]

3.3 KORESPONDENČNÍ ANALÝZA

Je grafická metoda k zobrazení vnitřní závislosti, asociace v tabulce přesností. Je soustředěná na dvojrozměrnou tabulku četností zvanou kontingenční tabulka, která obsahuje *n* řádků a *m* sloupců. Diagram korespondenční analýzy – subjektivní mapa – obsahuje dvě skupiny bodů: skupinu *n* bodů odpovídajících řádků a skupinu *m* bodů odpovídajících sloupců.[1]

Výhody:

1. Subjektivní mapa zobrazující vícerozměrných kategorických proměnných.

2. Zobrazení vztahů mezi kategoriemi řádků, nebo sloupců. Jsou-li například sloupci znaky v těsné blízkosti, budou mít u všech objektů vesměs podobné profily.

3. Poskytnutí společného obrazu řádkových a sloupcových kategorií ve stejném počtu dimenzí[1]

4. FUNKČNOST METOD

Základem pro testování metod jsou soubory pokusných dat, kde již předem je vytyčeno, který jedinec patří to které skupiny. Například v diagnóze choroby mohou takové skupiny být ve své podstatě dvě (pozitivní, negativní), nebo tři (pozitivní, negativní, neurčitá). Samozřejmě jich může být i více. Cílem tohoto testování bude zjistit, do jaké míry koresponduje ověřená diagnóza (nejspíše skutečná) s výsledkem té které klasifikace.

Jako dobrý ukazatel spolehlivosti klasifikačních metod je testování senzitivity, specificity, předpovědní hodnoty. Tento ukazatel je ale vhodný pro klasifikace, kde jsou výsledkem pouze 2 třídy (v medicíně tedy pozitivní a negativní).

Tento test si zde ovšem můžeme na okraj dovolit i přestože výsledkem analýzy jsou 3 třídy. Je to z důvodu, že pacientů, jenž nemají diagnózu buď pozitivní, nebo negativní je pouze osm, což je z celkových 154 osob téměř zanedbatelné.

Senzitivita, specificita, předpovědníhodnota a správnost diagnostickémetody

Falešně pozitivní diagnóza (FP)
U osoby, jenž nemá danou chorobu je tato choroba diagnostikována.

Falešně negativní diagnóza (FN)
U osoby, jenž má danou chorobu tato choroba není diagnostikována

Správně pozitivní diagnóza (SP)
U osoby, jenž má danou chorobu je tato choroba diagnostikována

Správně negativní diagnóza (SN)
U osoby, jenž nemá danou chorobu tato choroba není diagnostikována

Senzitivita (S+)
Vyjadřuje procento osob majících danou chorobu a tato choroba je zároveň diagnostikována.

$$S+ = \frac{SP}{SP + FN} \cdot 100 \tag{14}$$

Specificita (S-)

Vyjadřuje procento osob zdravých, jenž mají zároveň negativní test.

$$S- = \frac{SN}{SN + FP} \cdot 100 \tag{15}$$

Pozitivní předpovědní hodnota (PPH)

Procentuální pravděpodobnost přítomnosti choroby při pozitivním testu.

$$PPH = \frac{SP}{SP + FP} \cdot 100 \tag{16}$$

Negativní předpovědní hodnota (NPH)

Procentuální pravděpodobnost nepřítomnosti choroby při negativním testu.

$$NPH = \frac{SN}{SN + FN} \cdot 100 \tag{17}$$

Test shody

Test shody je obyčejný porovnávací test, který nám řekne v kolika procentech jsou data shodná. Důležité je si uvědomit, že výsledná data - matice kde řádky znamenají jednotlivé pacienty a sloupce jejich přiřazení ke třídě (viz Tabulka 3) – mohou být v různém pořadí, protože jak již bylo zmíněno, tak klasifikace rozdělí data do tříd, ale jednotlivé třídy nenesou informaci o tom, zda korespondují s třídou *pozitivní*, nebo s třídou *negativní*.

Test bude mít vypovídající hodnotu tehdy, když na toto bude brán zřetel.

4.1 REALIZACE METOD

Jako výhodné testovací prostředí byl zvolen MATLAB. Je to nástroj, jenž výborně pracuje s maticemi, což je zde časté a zároveň má některé z matematických funkcí a metod již implementované v sobě ve formě skriptů (toolboxů).

- První zásadou realizace funkční klasifikující metody je výběr správné sady parametrů pro jednotlivé objekty. Je totiž možné, že některé z parametrů mají minimální, nebo až žádný vliv na dělení objektů do tříd. Takové parametry jsou „šumem“, který může zcela zvrátit výsledek klasifikace.

- Druhá zásada při klasifikaci s dat u nichž je předem zřejmé do kolika tříd se budou separovat, je nastavit mechanizmy jednotlivých metod na tento počet. U metod nehierarchických je to snadné v tom, že vycházejí právě z daného počtu výsledných tříd. U metod hierarchických je nutné si údaje příslušnosti ke třídám vyzvednout dříve, nežli metoda vytvoří jeden kompaktní shluk.

- Důležitý předpoklad správné analýzy je ten, že veškeré obdržené data (viz Tabulka 1) ze shlukovací metody musí být ve stejném formátu, aby

je bylo možné kvalitativně porovnávat jak mezi sebou, tak o ostatními výsledky

- Vy výsledném údaji o jednotlivých třídách a jejích členech není obsažena informace o identitě třídy. Pokud tedy pro jednoduchost máme třídy dvě, máme třídu A a třídu B, v nichž jsou rozděleny objekty, tak nelze z výsledku analýzy bez znalostí povahy problému jednoznačně určit, která ze dvou tříd má charakter třídy A, a která má charakter třídy B. Zjednodušeně to lze vysvětlit na příkladu pacienta, který je pozitivní na nějakou chorobu, či negativní. Analýza ho přiřadí (více, či méně úspěšně) do skupiny pacientů se stejnou diagnózou,ale nebude zřejmé, která skupina má kterou diagnózu

4.2 DATA KE KLASIFIKACI

Pro analýzu byla požita data v určitém formátu. Datový soubor musí obsahovat řádky, jenž představují jednotlivé objekty a sloupce, které představují jednotlivé parametry objektů. Tyto data musejí být standardizována (to již bylo vysvětleno na začátku). Data obsahují parametry, o kterých ani nemusí být zřejmé, jak jsou pro klasifikaci důležité.

V případě datového souboru z medicínského prostředí je nutné znát skutečné diagnózy jednotlivých pacientů, na jejich základě se poté určuje přesnost klasifikační metody.

	xcom_G'	Fourier_Dist'	deriv&diff_mean'	xcom_C'	laplace&diff_ind90'	polar_EfValue'	xcom_m2'	ratio_mean'	sharpen&diff_mean'	diff_mean'	polar_max'	laplace&diff_mean'
objekt 1	0,529	1,000	0,142	0,317	0,234	0,333	0,162	0,297	0,308	0,307	0,348	0,143
objekt 2	0,338	0,459	0,186	0,172	0,208	0,184	0,069	0,016	0,014	0,016	0,181	0,062
objekt 3	0,717	1,000	0,235	0,500	0,686	0,502	0,309	0,549	0,526	0,525	0,479	0,222
objekt 4	0,406	1,000	0,247	0,286	0,250	0,375	0,195	0,214	0,223	0,222	0,558	0,094
objekt 5	0,227	0,353	0,197	0,102	0,236	0,108	0,034	0,066	0,068	0,068	0,192	0,006
objekt 6	1,000	1,000	0,437	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,620	0,355
objekt 7	0,290	0,509	0,021	0,156	0,059	0,186	0,070	0,010	0,009	0,011	0,277	0,048
objekt 8	0,362	0,509	0,299	0,196	0,147	0,222	0,089	0,006	0,009	0,006	0,733	0,111
objekt 9	0,339	1,000	0,253	0,179	0,147	0,205	0,080	0,178	0,189	0,186	0,322	0,166
objekt 10	0,306	1,000	0,149	0,173	0,073	0,212	0,083	0,131	0,141	0,140	0,402	0,072
objekt 11												

Tabulka 1 – příklad datového souboru

4.3 VÝBĚR PARAMETRŮ

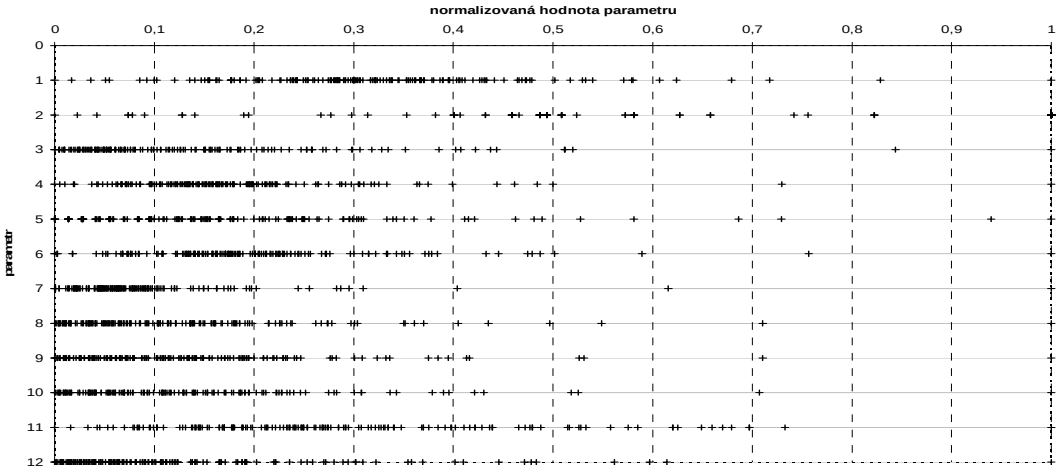
Jak již bylo zmíněno (kap. 4.1), tak jedním z důležitých zásad správné klasifikace je volba správných parametrů u objektů.

Docílíme tím toho, že klasterizace do jednotlivých tříd nebude provázena šumem a zároveň při snižování počtu parametrů snižujeme i dimenzi prostoru, kde

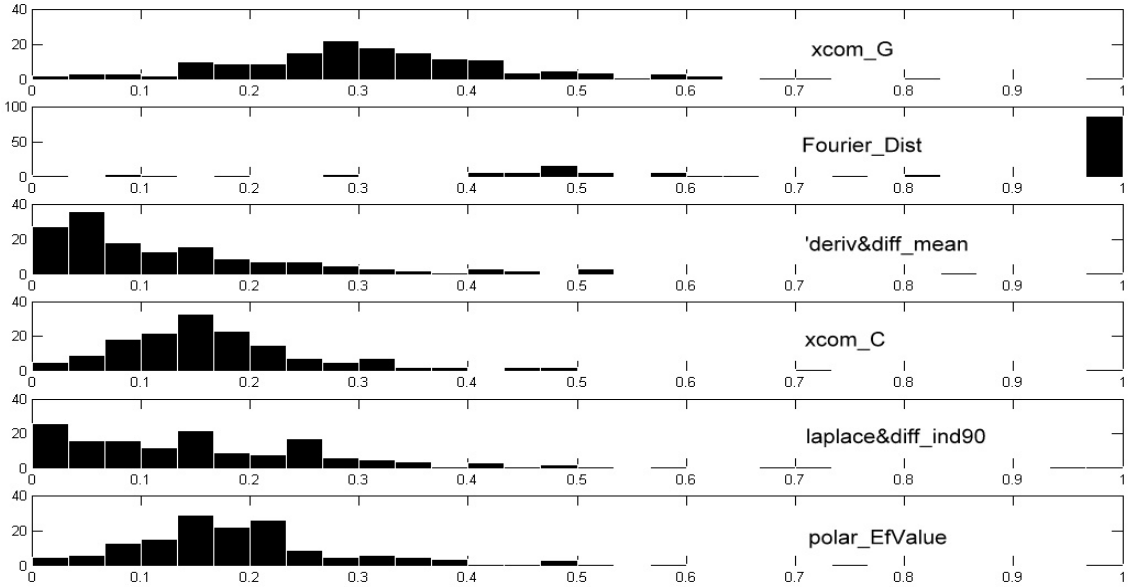
jsou objekty umístěny. V důsledku snížení dimenze se urychlí průběh klasifikace, protože některé výpočty budou snadnější.

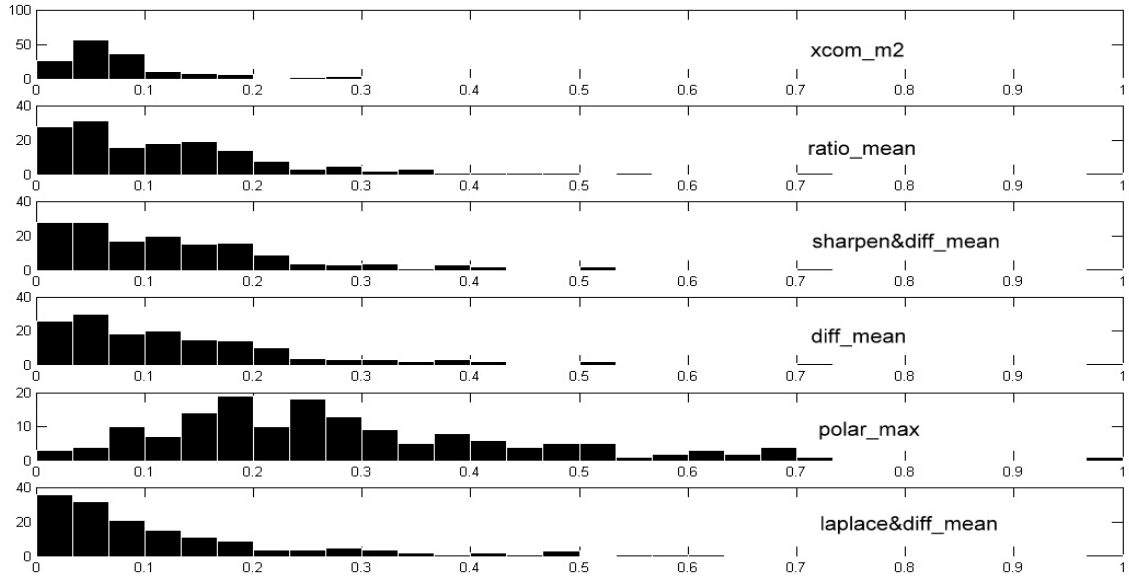
Výběr parametrů je možné udělat v základě dvěma způsoby.

- Jedna možnost je udělat si na data určitý náhled a na základě zkušenosti, nebo odhadu z nich pak vybrat ta validní a vnést je do klasifikace. V našem případě dat je možný náhled průmětu souřadnic objektů do jednotlivých os 12ti-rozměrného prostoru (Obrázek 22), nebo si lze pro každý tento průmět vytvořit histogram(Obrázek 23). Nebo ideální případ je v momentě kdy jsme si jisti, který parametr u pacienta změřený bude mít pro klasifikaci přínos.



Obrázek 19 – průmět parametrů(1-12)





Obrázek 20 – histogramy průmětu parametrů (1-12)

- Druhou alternativou je tento výběr přenechat pouze nějakému výpočtu, bez lidského zásahu. Pro takovou alternativu jsem zvolil výpočet entropie jednotlivých průmětů (parametrů) a na základně velikosti entropie pak vybral vhodné parametry ke klasifikaci.

  Čím menší totiž entropie u dat bude, tím větší energie daných dat a to znamená, že histogram dat bude mít jistá lokální maxima a minima a bude tedy pravděpodobnější, že data budou ke správné klasifikaci přispívat, nežli histogram, který značí rovnoměrně rozmístěné parametry.

Entropie:

Je obecně veličina udávající míru neuspořádanosti zkoumaného systému nebo také míru neurčitosti daného procesu.

$$S = -\sum_{i=1}^{N} (p_i \log p_i). \tag{18}$$

Zde suma probíhá přes všechny možné stavy, p_i jsou přitom pravděpodobnosti daných stavů[5].

4.4 POROVNÁVACÍ TEST Č.1

V porovnávacím testu na jednom datovém souboru jsou použity 4 shlukovací **metody:**
Metoda nejbližšího souseda
Metoda nejvzdálenějšího souseda
Metoda K-means
Metoda C-means

Datový soubor obsahuje 154 pacientů, každý se 12ti příznaky a třemi diagnózami. Test se provádí pomocí senzitivity, specificity, PPH a NPH výsledných

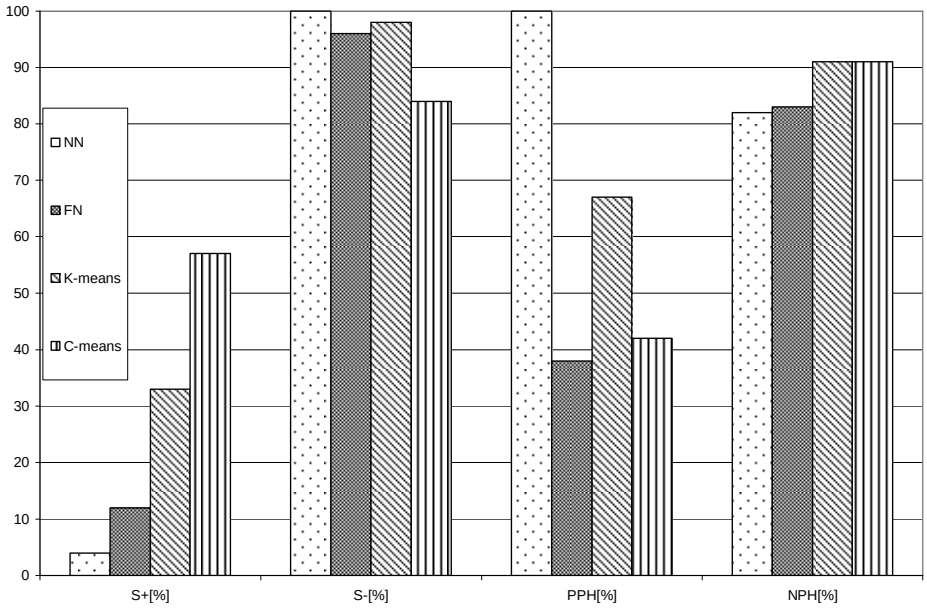
dat. Jelikož jsou mezi výslednými daty diagnostikovány v klasifikaci jedinci i do 3.třídy (diagnóza *neurčitá*), jsou z tohoto testu vynecháni, ale v poslední řádek tabulky (Tabulka 2) nese informaci, z kolika osobami test proběhl.

Též je důležité říci, že pro většinu testů je potřebné výsledek „fuzzy klasifikace“ převést do tvaru, kde výsledek každému objektu bude náležet právě jedné třídě a to té, ke které má největší míru příslušnosti.

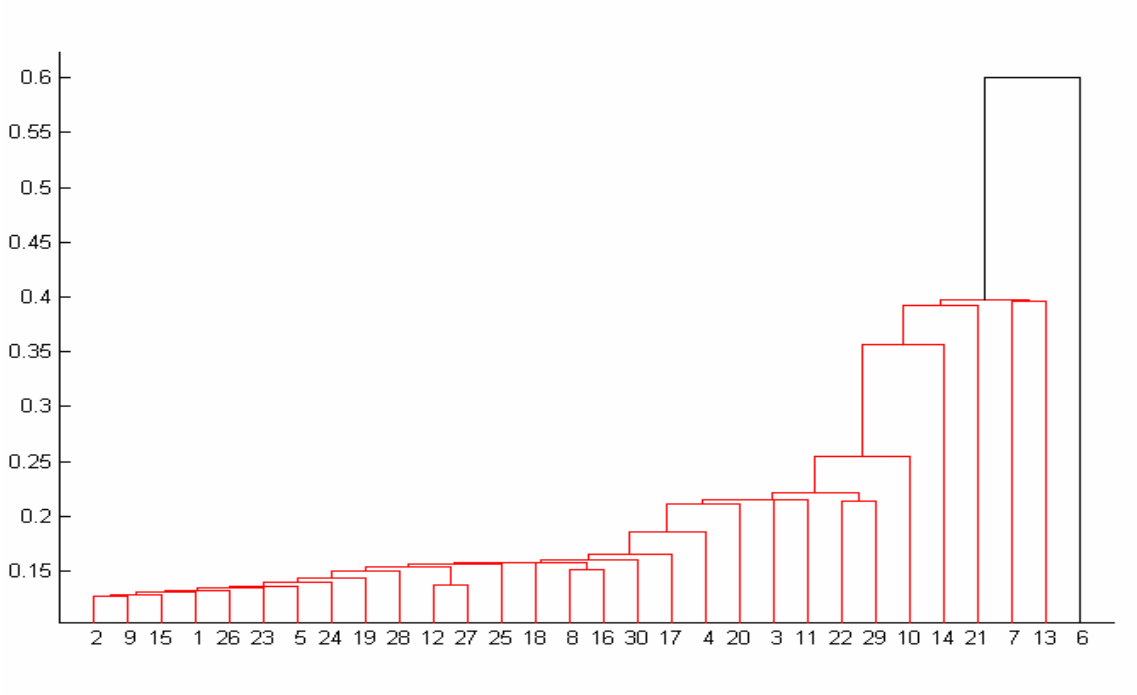
Metody se zdají být citlivé na odhalování pacientů, kteří jsou zdraví s negativním testem a jako nejúspěšnější se jeví metoda nejbližšího souseda, což bude zřejmě výjimka, jelikož tato metoda má tendenci vytvářet jeden shluk s většinovým počtem jedinců a zbývající s pouze zanedbatelným počtem. Tato vlastnost je patrná na dendrogramu (Obrázek 20).

	Nejbližší soused	Nejvzdálenější soused	K-means	C-means
S+[%]	4	12	33	57
S-[%]	100	96	98	84
PPH[%]	100	38	67	42
NPH[%]	82	83	91	91
osob[%]	93,51	92,86	61,04	54,55

Tabulka 2 – Úspěšnost metod při pěti parametrech

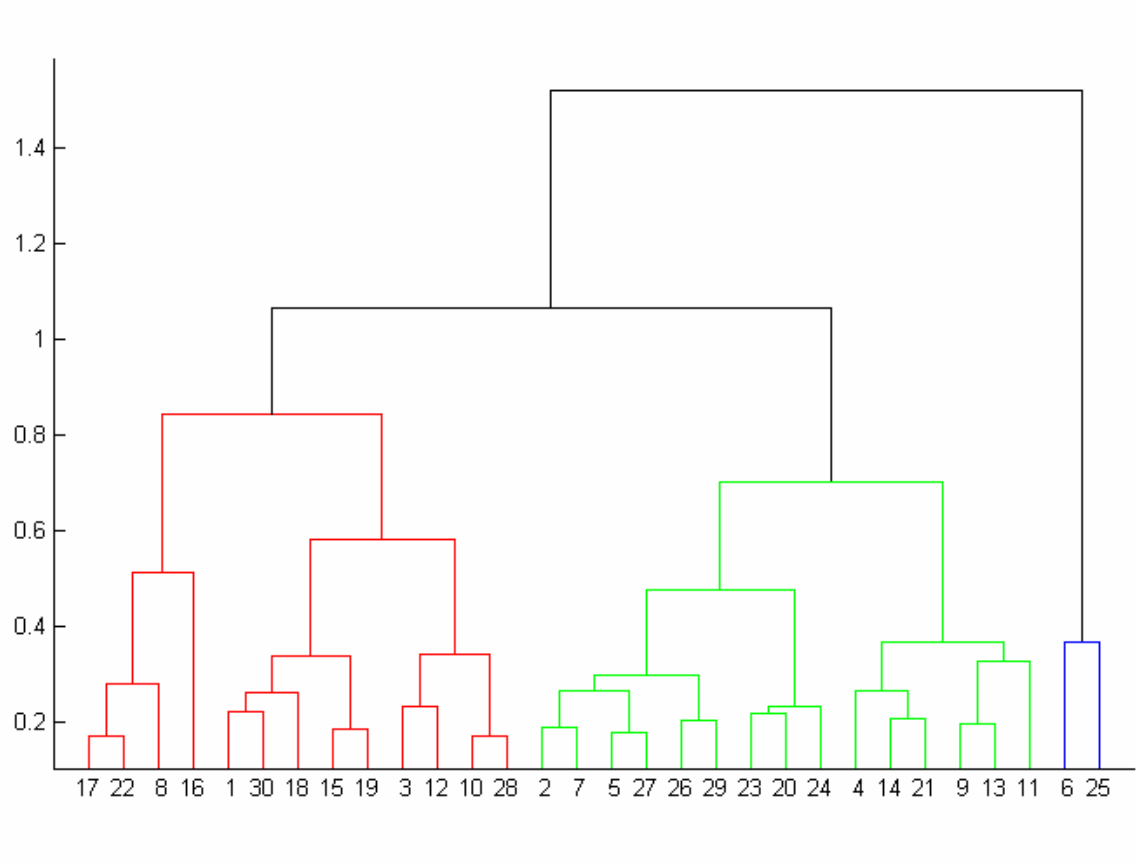


Obrázek 21 – grafické vyjádření (Tabulka 2)



Obrázek 22 – Nejbližší soused – dendrogram

Oproti tomu by se zdála vhodnější metoda nejvzdálenějšího souseda, Jejíž způsob klasterizace dat je mnohem sofistikovanější (viz Obrázek 21).



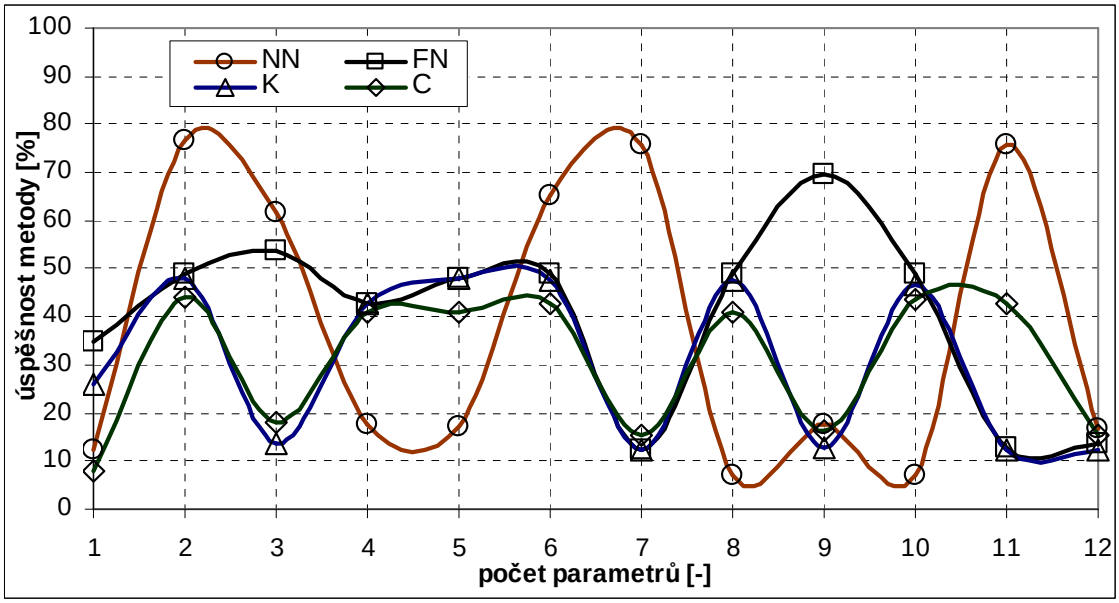
Obrázek 23 – Nejvzdálenější soused – dendrogram

4.5 POROVNÁVACÍ TEST Č.2

Tento test bere v úvahu pouze přímou shodu dat a vyjadřuje v procentech shodnost dat získaných klasifikací, s daty skutečnými.

4.5.1 Volba dat pomocí entropie

Jelikož nemusí být známa povaha jednotlivých měřených parametrů u osob, je nutné výběr vhodných parametrů ponechat programu, nebo za pomocí histogramů tyto parametry vybrat ručně. Tento test zjistí, zda lze data vybírat čistě na základě jejich entropie, čili upřednostňovat ty parametry, jejichž datový vektor má entropii menší. Pomocí entropie byli postupně voleny „vhodné“ parametry v počtu od 1 až do výběru všech parametrů. Tyto parametry vstoupily do klasifikace pomocí již uvedených 4 metod. Výsledek je vynesen do grafu (Obrázek 24). Graf ukazuje závislost počtu parametrů, jenž byli zvoleny entropickým kritériem na úspěšnosti metody s takto zvolenými parametry.



Obrázek 24 – test výběru dat entropií

Zkratky: *NN – Nejbližší soused; FN - Nejvzdálenější soused; K – K-means; C – C-means*

Výsledek nenese jednoznačnou informaci o tom, že by entropie byla správným kritériem pro výběr vhodných parametrů, ale z grafu (Obrázek 24) je patrné, že úspěšnost všech metod je při jednom vybraném parametru a též při 12ti parametrech (tedy všech dostupných) nepřesahuje hodnotu 40%. To je jen potvrzení skutečnosti, že klasifikaci je potřeba více parametrů, ale pokud jsou zde parametry „šumové“, tak je třeba je z klasifikace vypustit. Zde je možné, že entropie dobře vypouštěla poruchový parametr, jenž se nakonec projevil na poslední funkční hodnotě v grafu, kdy ani jedna z metod již nedosáhla úspěšnosti větší nežli 20%.

O který parametr se jedná je zřejmé z tabulky (Tabulka 3). Mimo jiné je v tabulce uvedeno postupné zavádění parametrů do klasifikace.

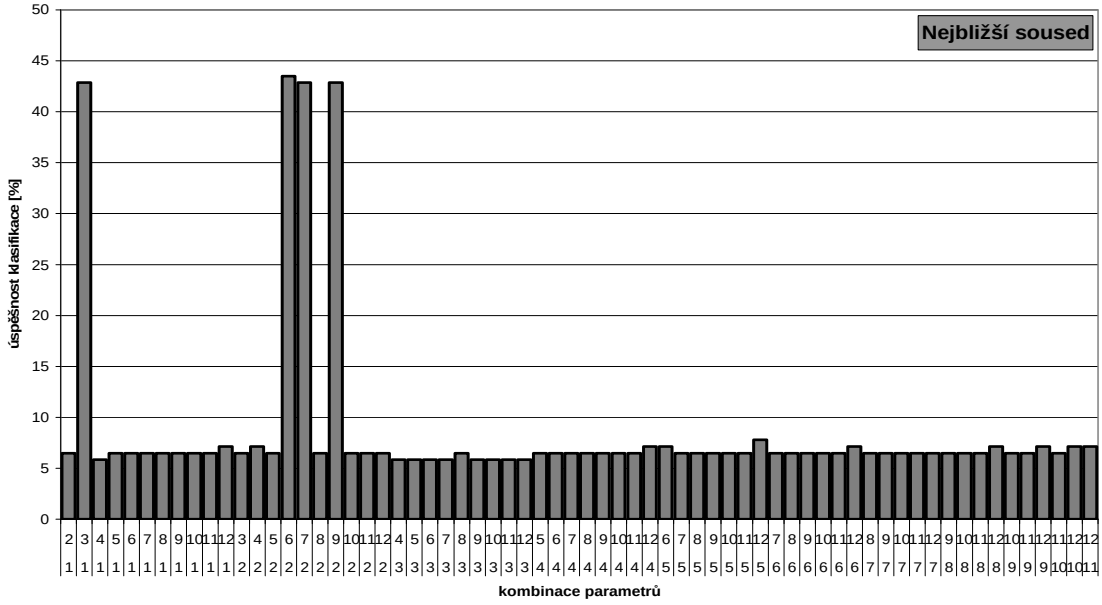
počet parametrů	Použité parametry											
	'Fourier_Dist'	'xcom_m2'	'laplace&diff_ind90'	'laplace&diff_mean'	'ratio_mean'	'diff_mean'	'xcom_C'	'deriv&diff_mean'	'polar_EValue'	'sharpen&diff_mean'	'xcom_G	'polar_max'
1.	2											
2.	2	7										
3.	2	7	5									
4.	2	7	5	12								
5.	2	7	5	12	8							
6.	2	7	5	12	8	10						
7.	2	7	5	12	8	10	4					
8.	2	7	5	12	8	10	4	3				
9.	2	7	5	12	8	10	4	3	6			
10.	2	7	5	12	8	10	4	3	6	9		
11.	2	7	5	12	8	10	4	3	6	9	1	
12.	2	7	5	12	8	10	4	3	6	9	1	11

Tabulka 3 – postupné vybírání parametrů

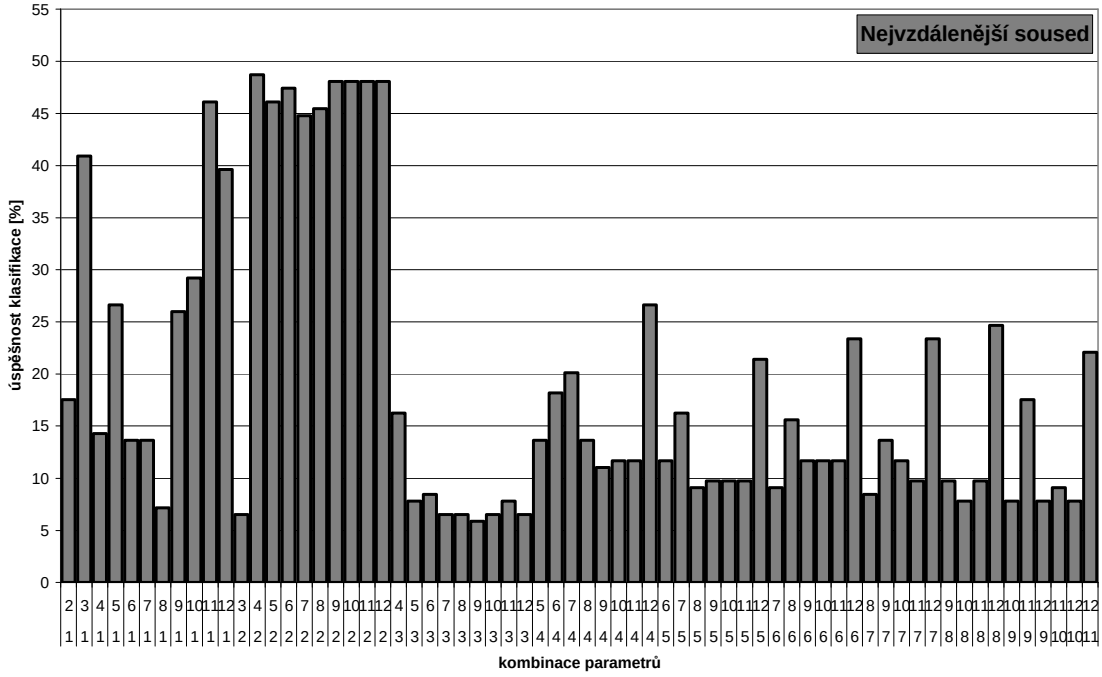
4.5.2 Výběr dat podle výsledků klasifikací

Další možností, jak zjistit parametry jenž jsou vhodné pro klasifikaci je ta možnost, že necháme klasifikaci proběhnout pro všechny možné kombinace parametrů na vstupu a podle výsledků bude zřejmé, které kombinace parametrů jsou pro klasifikace vhodné a které méně. Je zde ale zapotřebí vzít v úvahu, zda vstupní datový soubor není takových rozměrů, že by potom doba klasifikace přesáhla meze naší trpělivosti.

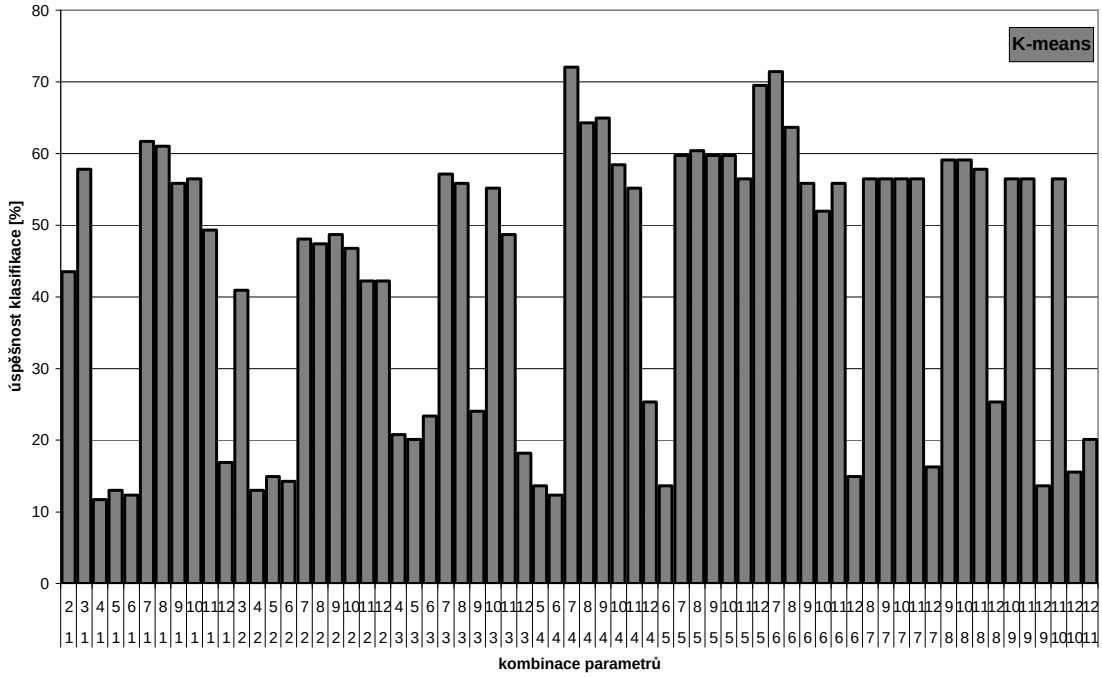
Pro demonstraci byl proveden test, kde se hledala kombinace dvou parametrů, které vedou ke zdárnému výsledku klasifikace pokud možno u většiny metod. Zvoleny byli dva parametry, protože budou moci být vykresleny grafy v pro nás přijatelném 2D prostoru.



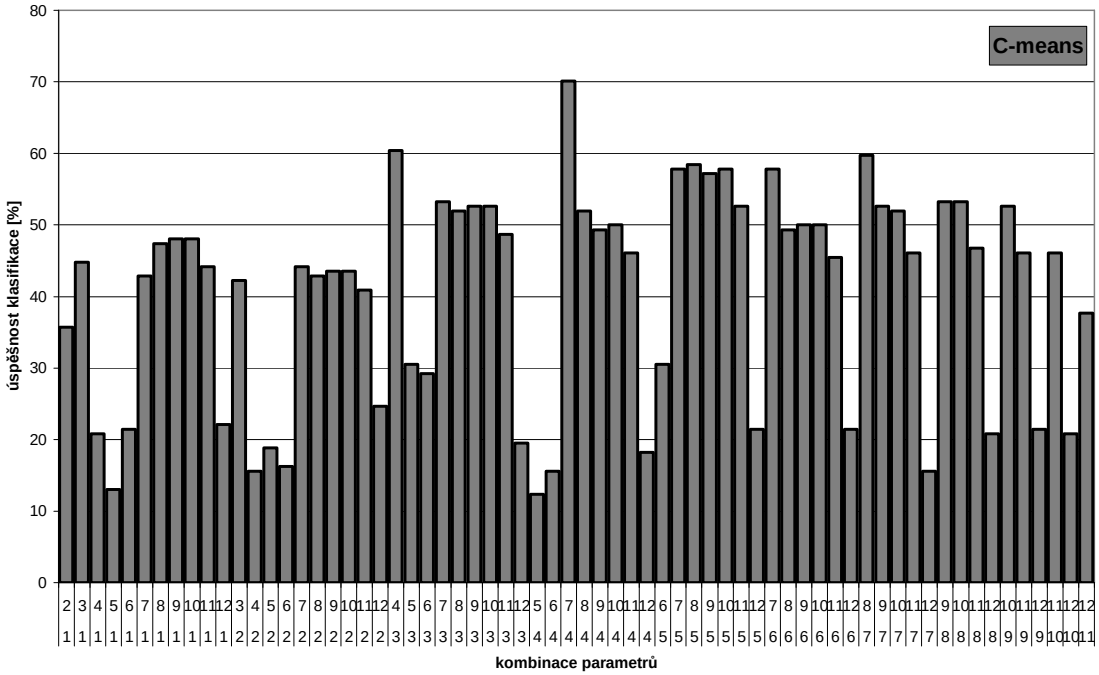
Obrázek 25 – úspěšnost metody pro kombinace parametrů



Obrázek 26 - úspěšnost metody pro kombinace parametrů



Obrázek 27 - úspěšnost metody pro kombinace parametrů



Obrázek 28 - úspěšnost metody pro kombinace parametrů

Z grafů na obrázcích 25 – 28 je patrné, jaké kombinace dvou parametrů jsou přínosem pro klasifikaci.

Na prvních dvou (hierarchické metody) je vidět, že metody jsou úspěšné nad 40% jen pro několik málo dvojic parametrů.

Pro metodu nejbližšího souseda, jde jen o 4 kombinace parametrů jež vedou k poměrně úspěšné klasifikaci. Průměrná úspěšnost metody přes všechny kombinace parametrů nečinila více nežli 10%.

Pro metodu nejvzdálenějšího soused je zde 11 dvojic parametrů a průměrná úspěšnost vzrostla na téměř 20%.

U metod nehierarchických (Obrázek 27a 28) byly klasifikace úspěšné pro více kombinací parametrů. Jejich průměrná úspěšnost byla kolem 40%.

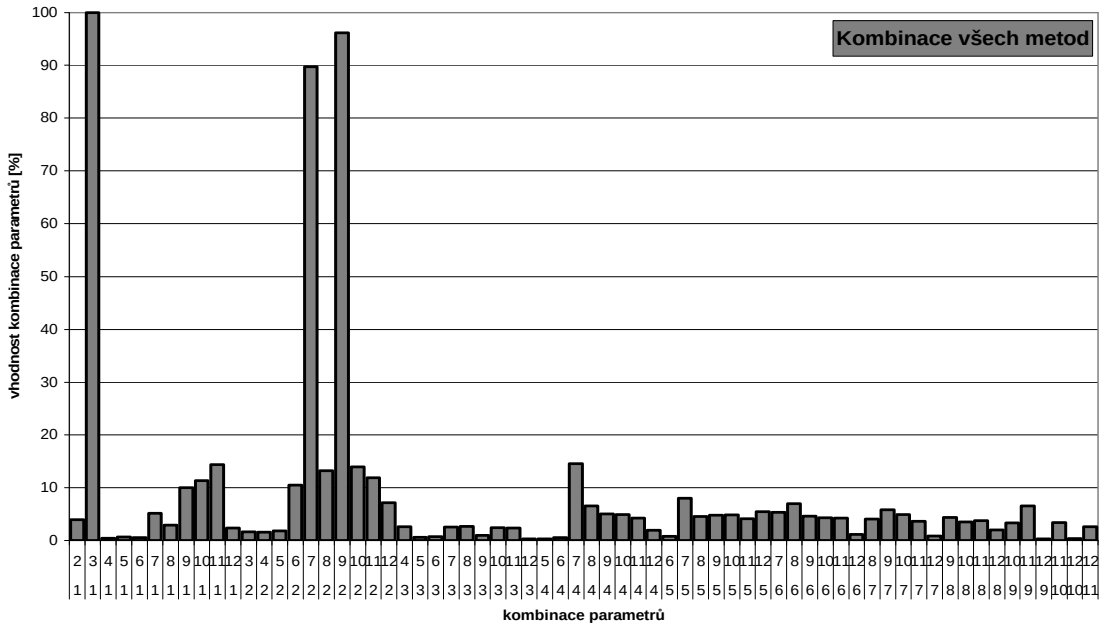
Pro výběr nejefektivnější dvojice parametrů jsem zvolil graf společný všem metodám, skóre F_{MAX} dvojice parametrů ***m*** a ***n*** je dáno:

$$F_X(mn) = F_{NN}(mn) \cdot F_{FN}(mn) \cdot F_{K-MEANS}(mn) \cdot F_{C-MEANS}(mn) \tag{19}$$

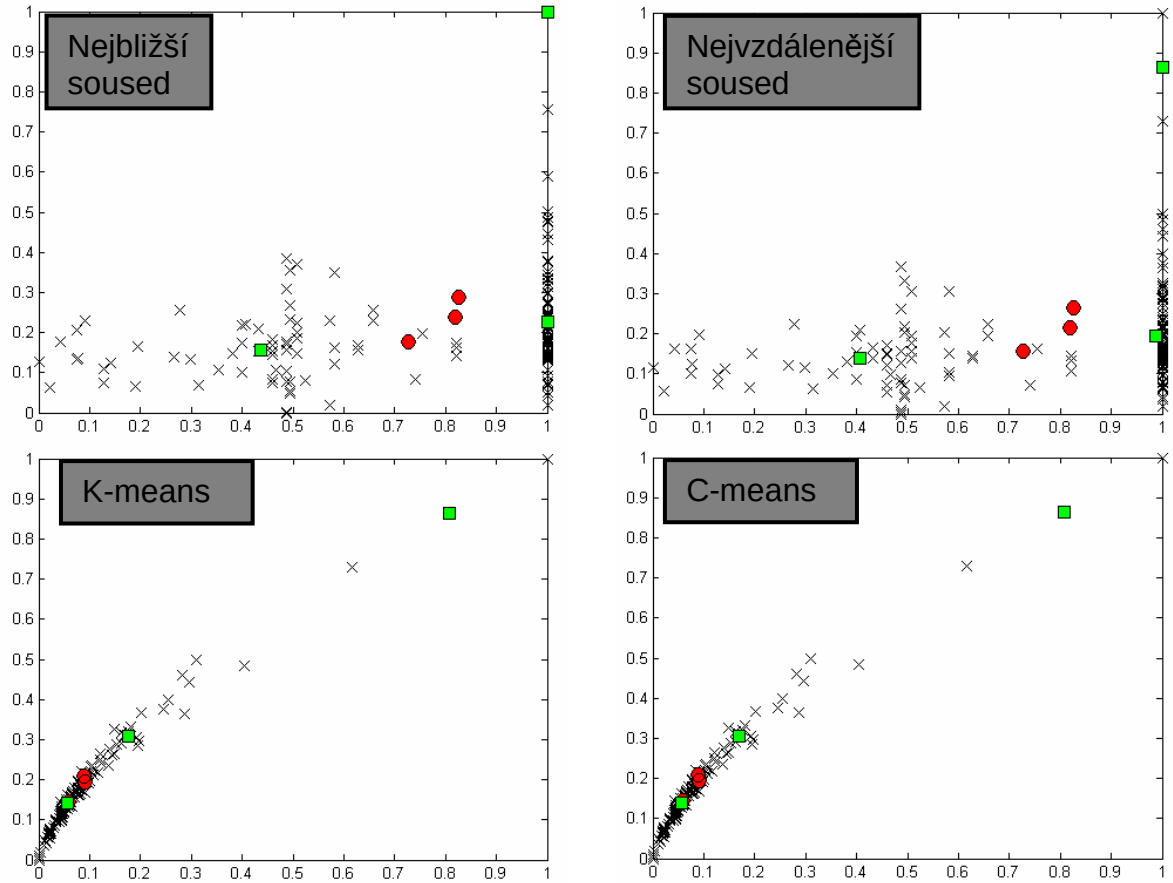
$$F_{MAX}(mn) = \frac{Fx(mn)}{\max(F_X(mn))} \cdot 100 \tag{20}$$

Násobením jednotlivých hodnot dosáhneme potlačení výsledných hodnot u nichž jedna z metod klasifikace má nízkou úspěšnost a naopak se vyzdvihnou parametry které jsou úspěšné u metod všech. Výsledná F_{MAX} je pak vynesena v grafu (Obrázek 29). Nejlepší kombinace parametrů má úspěšnost 100% a ostatní

jsou uvedeny v relativní formě vůči této maximální. Z grafu je zřejmé, že nejlépe hodnocené parametry jsou 1-3, 7-2 a 9-2.



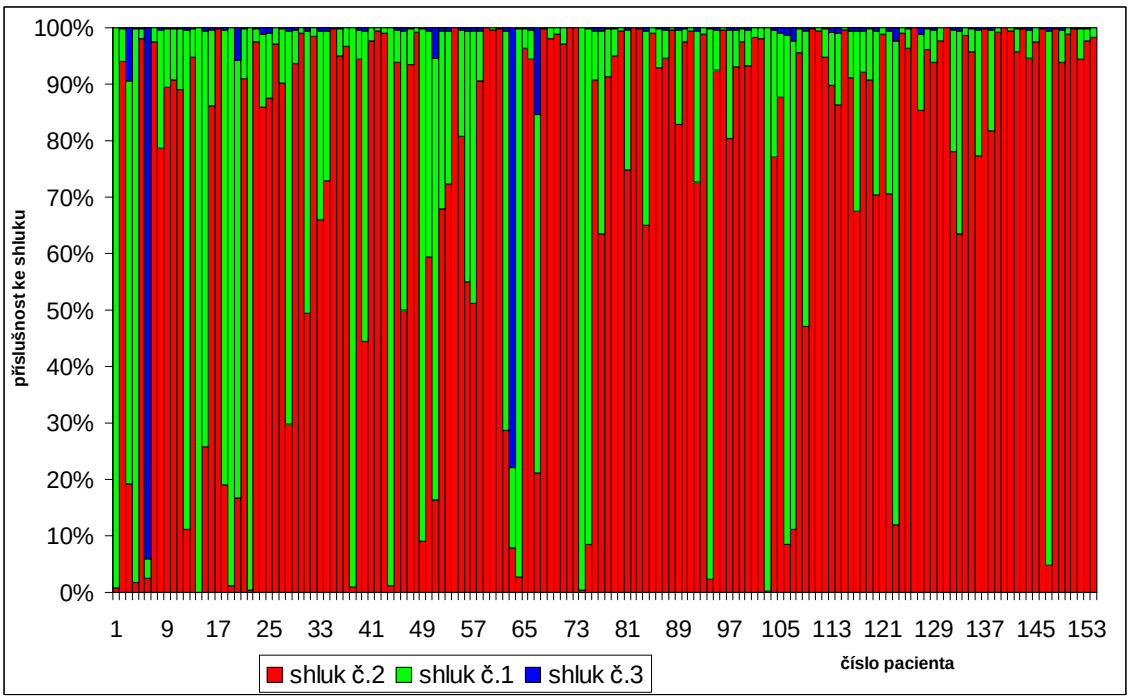
Obrázek 29 – výsledná kombinace



Obrázek 30 – metody při nejvhodnějších parametrech

V grafech (Obrázek 30) jsou vykresleny kombinace nejúspěšnějších parametrů pro danou metodu jako body v 2D prostoru. Tyto parametry vešly do klasifikace, ze které byli obdrženy středy tří klasifikovaných tříd (značeny zelenými čtverci). Pro porovnání byli vyznačeny i středy skutečných tříd (červenými kolečky). Je zde opravdová odlišnost v umístění center, ale klasifikace jsou přesto poměrně úspěšné, což je dané nejspíš podobným rozmístěním center.

Pro fuzzy techniku shlukování je názornost vhodnější využít plné její síly a zobrazit příslušnost k jednotlivým třídám pomocí stupně příslušnosti, jenž není tak triviální jako u ostatních metod. Předností této metody je to, že každý pacient patří ke každé diagnóze s určitou nenulovou pravděpodobností. Na bargrafu (Obrázek 31) je dobře čitelné, že největší podíl pacientů patří do shluku č.2 a naopak nejméně do shluku č.3.

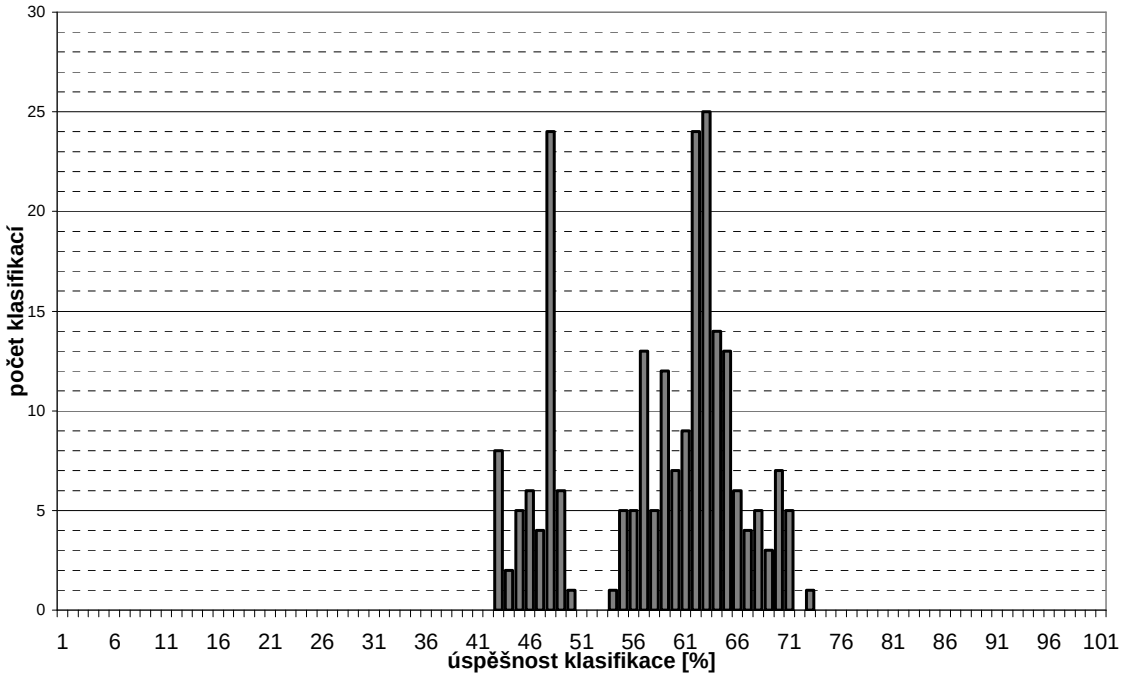


Obrázek 31 – bargraf pro C-means

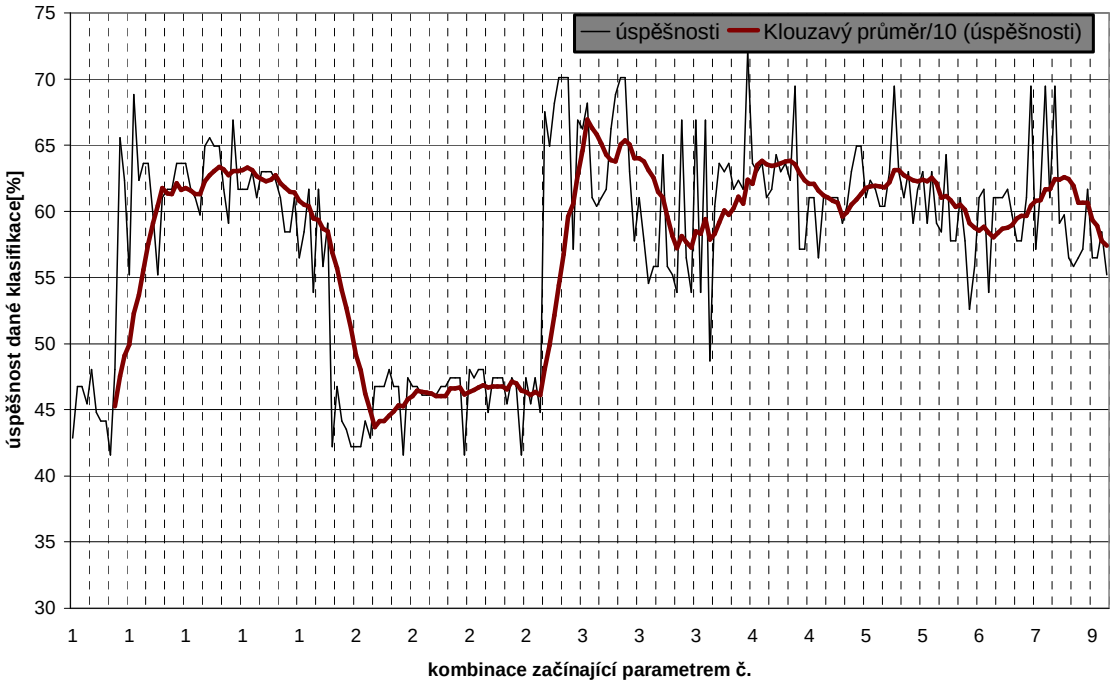
4.5.3 Prověření metody K-means pro 3 a 4 vstupní parametry

Aby byla klasifikace účinnější, ne možno zadat více nežli 2 vstupní parametry. Aby možné otestovat jak se chová výsledek klasifikace pro všechny kombinace parametrů, je potřeba nechat proběhnout klasifikační proces více nežli 200krát. Takový výpočet již značně záleží na efektivitě kódu. Pro metody nehierarchické kód není natolik efektivní a výpočet by pak trval téměř ½ hodiny pro 3 parametry a přes 4 hodiny pro parametry 4. Testována byla tedy metoda K-means, jakož to zástupce metod hierarchických.

Její výsledek zle popsat grafem četností (Obrázek 33) jednotlivých skóre metod, při různých kombinacích parametrů. Jako nejlepší kombinace parametrů zde vystoupila kombinace 4-6-7. Z grafu jednotlivých úspěchů klasifikace pro jednotlivé kombinace parametrů byl vyčten pokles úspěšnosti pro kombinace obsahující parametr č.2 .

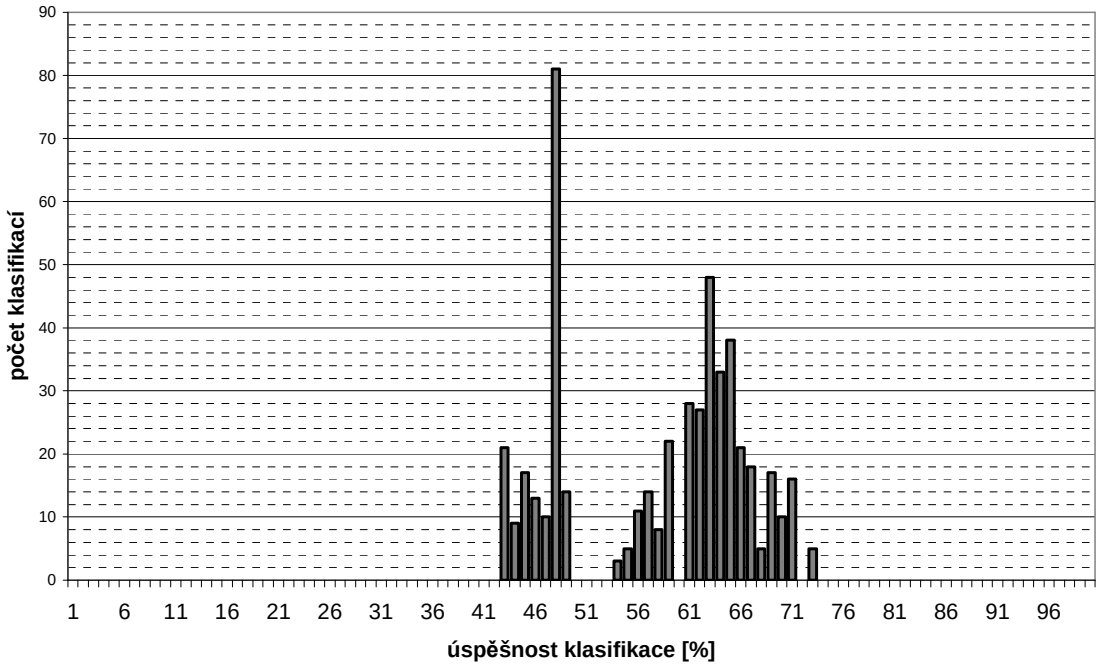


Obrázek 32 – úspěšnost klasifikace -3 parametry

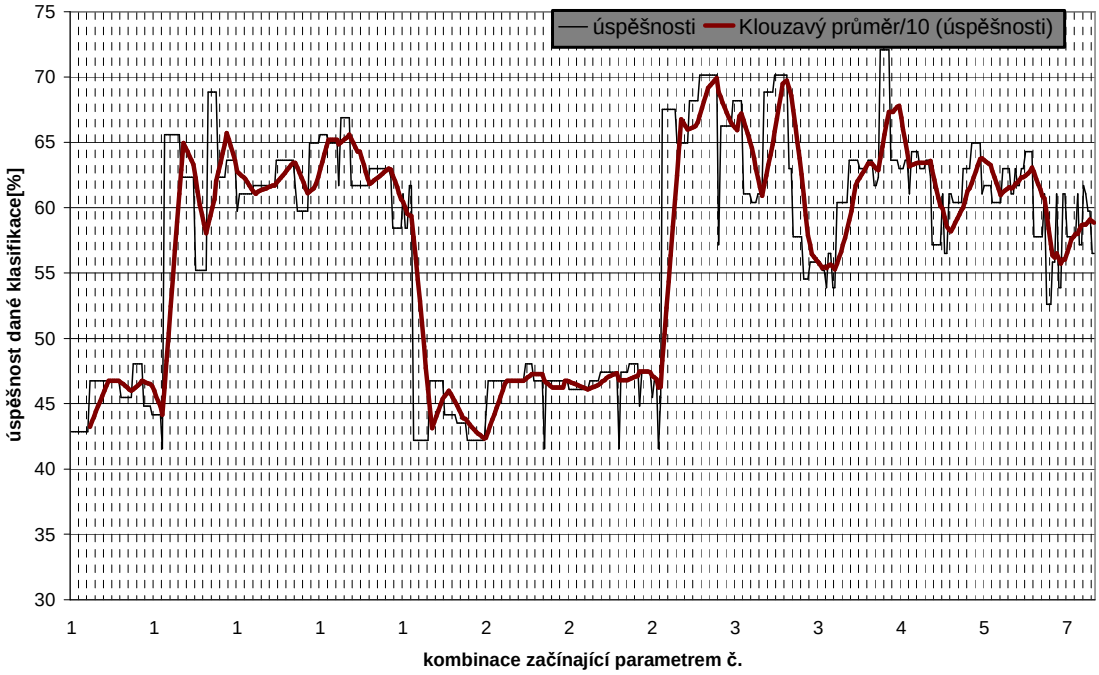


Obrázek 33 – úspěšnost klasifikace – 3 parametry

V další části je jen pro série stejných grafů pro K-means s kombinacemi 4 parametrů. Na obrazcích (Obrázek 34-35) je opět znát, že pacientský parametr č.2 snižuje kvalitu klasifikace.

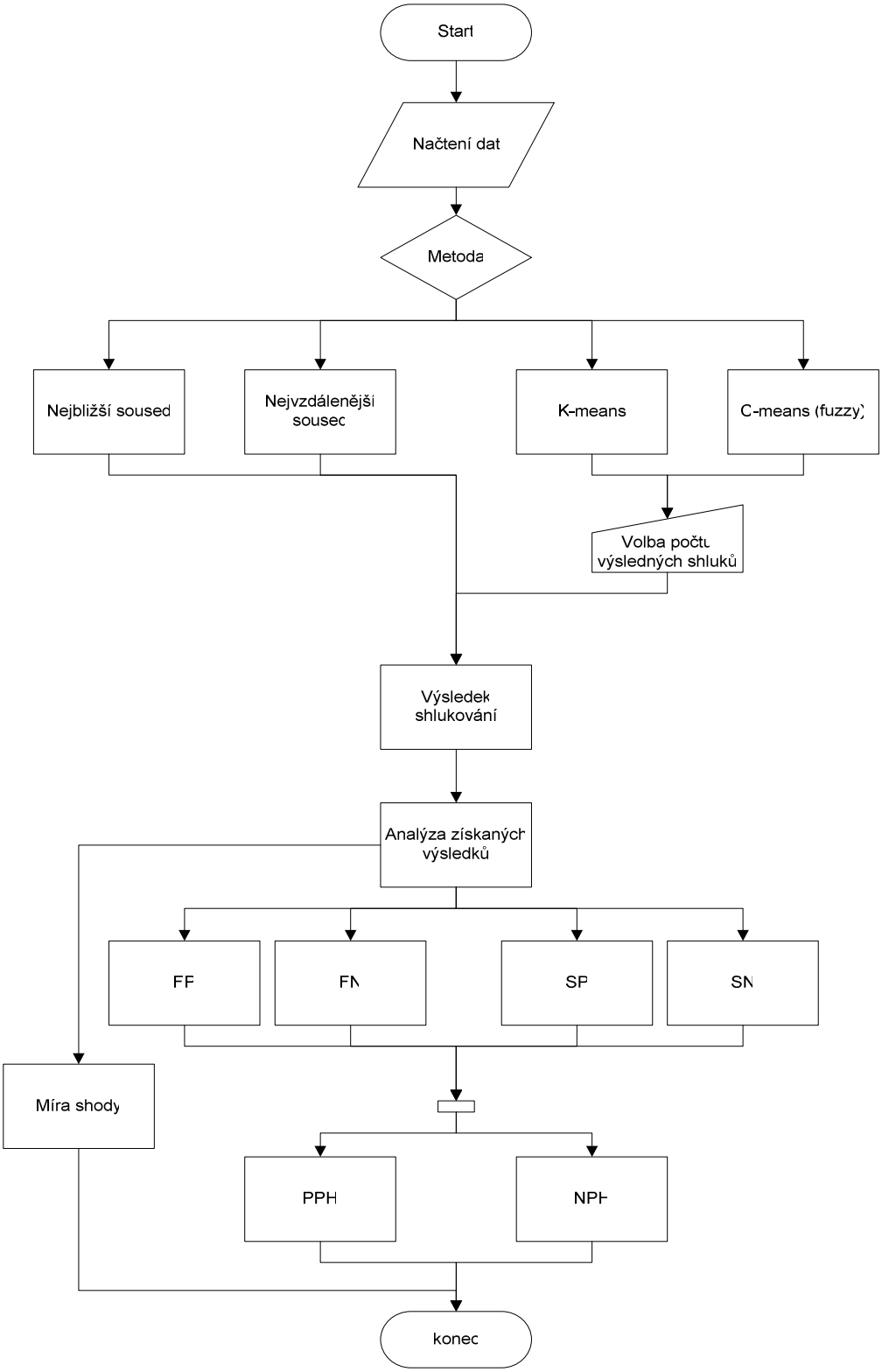


Obrázek 34 - úspěšnost klasifikace -4 parametry



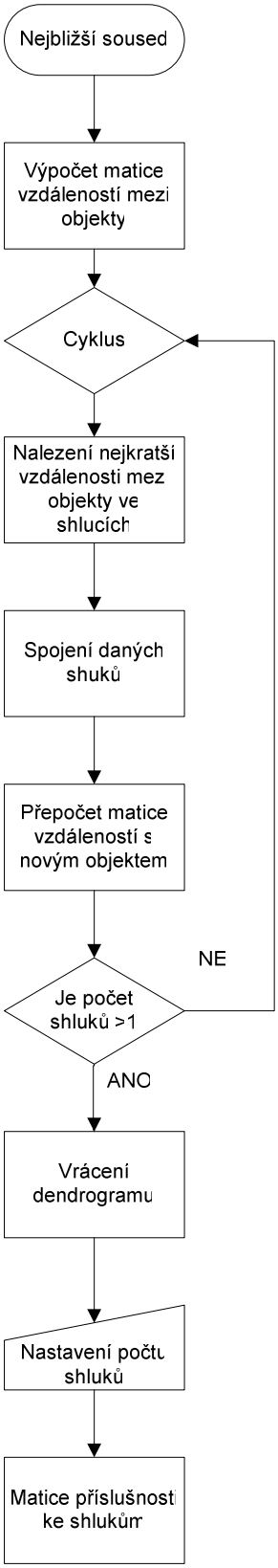
Obrázek 35 - úspěšnost klasifikace -4 parametry

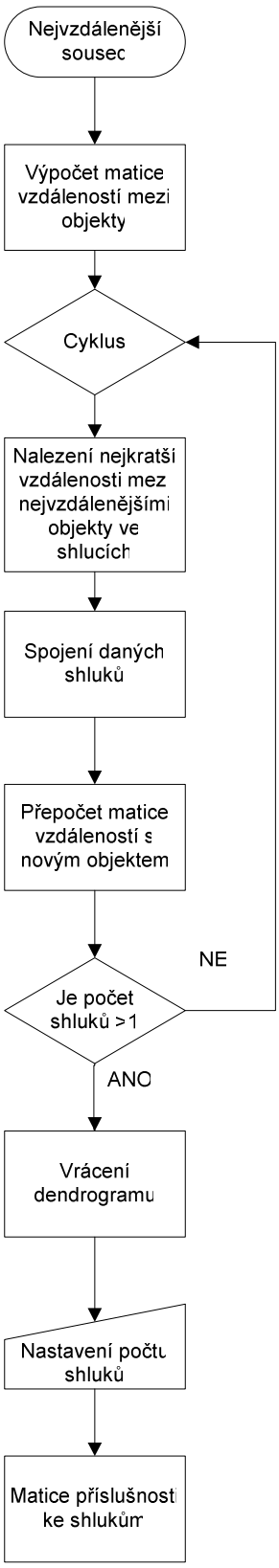
Vývojový diagram použitého algoritmu:



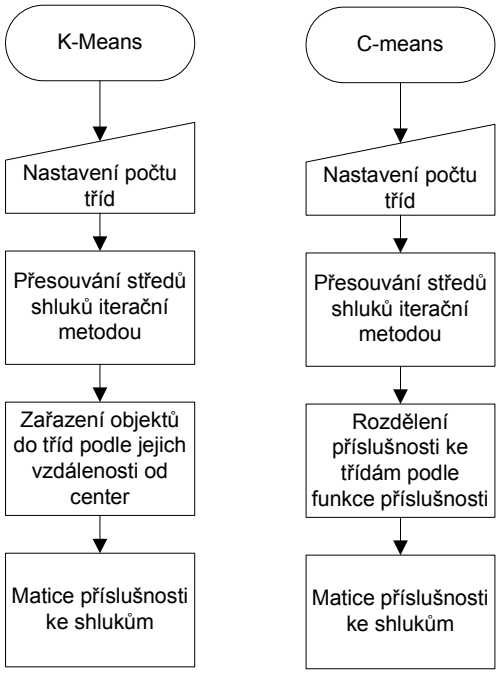
Celý algoritmus je možné uzavřít do cyklu a měnit v něm vstupní parametry.

Metoda nejbližšího a nevzdálenějšího souseda:





Metoda K-means a C-means



Algoritmus shlukovacích metod K-means a C-means se liší až v konečné fázi, kdy místo jednoznačného přiřazení do shluku, jak je tomu u K-means je u fuzzy alternativy vyjádřena pouze míra účasti objektu ve shluku (viz Tabulka 4)

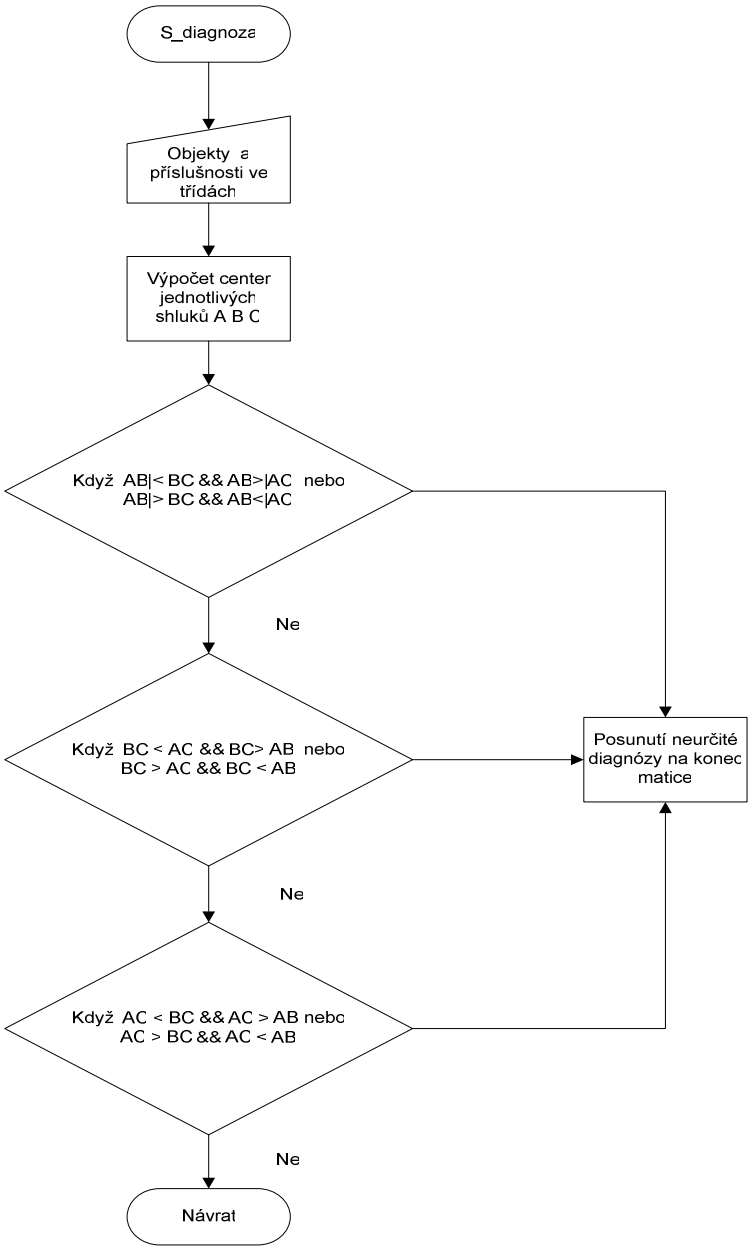
č.	pozitivní	negativní	neurčeno
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	1	0
6	0	0	1
7	0	1	0
8	1	0	0
9	1	0	0
10	1	0	0

č.	Pozitivní	negativní	neurčeno
1	0,270974	0,08783	0,641196
2	0,284032	0,669423	0,046546
3	0,172349	0,094186	0,733465
4	0,317713	0,086887	0,595399
5	0,020557	0,976203	0,00324
6	0,279149	0,205741	0,515111
7	0,445603	0,50464	0,049757
8	0,41231	0,205461	0,38223
9	0,923542	0,048169	0,028288
10	0,870884	0,075683	0,053432

Tabulka 4 – příslušnosti k třídám

(vlevo K-means a vpravo C-means)

Algoritmus, jenž určuje povahu shluku při třech diagnózách. Zjistí který ze shluků leží uprostřed mezi dvěma zbývajícími a prohlásí ho v případě diagnóz: pozitivní – negativní – neurčitá, za neurčitou. Základ tvoří porovnávání vzdáleností center jednotlivých shluků A B C. Čili vzdáleností |AB|, |BC| a |AC|. Použitelný je pro hodnocení kvality na základě senzitivity a specificity. U prostého porovnávání již nemá smysl.



5. ZÁVĚR

Byla rozebrána problematika týkající-se analýzy dat. Především bylo shrnutu, na jakých základních principech stojí diskriminační analýza (DA) a shluková analýza (SA), její modifikace, především metoda nejbližšího souseda, nejvzdálenějšího souseda a hierarchické modifikace K-means a C-means.

K jednotlivým metodám byl uveden i matematický základ v podobě výčtu metrik a jejich vzorců pro výpočet vzdálenosti. K DA a SA byl přiložen i optimální postup klasifikace.

Byl navržen algoritmus ve vývojovém prostředí MATLAB, který měl za úkol prověřit funkčnost jednotlivých metod a jejich chování při změně parametrů na dodaném datovém souboru.

Chování metod a jejich efektivita byli testovány dvěma způsoby. První test byl vhodný spíše pro medicínská data v jejichž výsledku jsou pouze dvě diagnózy. Tento tes ze zde neprojevil jako vhodný. Kdyby byla klasifikace upravena, abychom ve výsledku obdrželi pouze 2 třídy, metodu nejbližšího souseda by nebylo možné aplikovat, jelikož by vytvořila pouze jednu objemnou třídu a druhou s malým počtem jedinců. Je to způsobeno algoritmem shlukování této metody, čili to má v povaze.

Ve druhém testu, vhodnějším byla hodnocena procentuální shoda skutečné diagnózy a diagnózy vzešlé klasifikací. Některé metody dosahovali více nežli 70% shody. Toto číslo není nijak závratné, ale musíme vzít v potaz, že ne každá data jsou pro shlukování dobrá a lze na ně tyto algoritmy používat. Testovaná data měla jisou tendenci tvořit shluky, ale až po výběru správné kombinace parametrů.

Hledání parametrů bylo buď automatizováno na základě entropie datové řady nebo byl výběr uskutečněn po prozkoušení všech možných kombinací pro 2, 3 a 4 parametry na vstupu klasifikace. Tento druhý systém se zlál být sofistikovanějším, jelikož entropie volí i data která jsou očividně poruchovým signálem.

Výsledkem byla vždy úvaha, které parametry ve zkoušce jak obstály. Veškeré úvahy byly doplněny grafy.

Závěrem shrnuji, že klasifikace pomocí shluků u tohoto typu dat byla poměrně úspěšná, ale ne naprosto nepoužitelná v lékařském prostředí kde cifra 70% úspěšnosti by byla naprosto nepřijatelná. Pro některé diagnózy bude vždy kvalitnější jiná (nematematická) metoda, jako je chemický rozbor, některé zobrazovací techniky v kombinaci se zkušeností odborníka.

6. SEZNAM

Obrázky

OBRÁZEK 1 - KOS A LABUŤ (DVA PARAMETRY) 10
OBRÁZEK 2 – KOS A ŠPAČEK (DVA PARAMETRY) 10
OBRÁZEK 3 – EUKLEIDOVSKÁ VZDÁLENOST V 2D 11
OBRÁZEK 4 – HAMMINGOVA VZDÁLENOST V 2D[1] 11
OBRÁZEK 5 – TYPICKÝ DENDROGRAM PRŮMĚROVÉ METODY 19
OBRÁZEK 6 – TYPICKÝ DENDROGRAM CENTROIDNÍ METODY 19
OBRÁZEK 7 – NEJBLIŽŠÍ SOUSED 20
OBRÁZEK 8 – TYPICKÝ DENDROGRAM METODY NN 20
OBRÁZEK 9 – NEJVZDÁLENĚJŠÍ SOUSED 20
OBRÁZEK 10 – TYPICKÝ DENDROGRAM METODY FN 21
OBRÁZEK 11 – TYPICKÝ DENDROGRAM MEDIÁNOVÉ METODY 21
OBRÁZEK 12 – WARDOVA METODA 22
OBRÁZEK 13 – TYPICKÝ DENDROGRAM WARDOVY METODY 22
OBRÁZEK 14 – POČÁTEČNÍ ITERACE (NULTÁ) 28
OBRÁZEK 15 – 2. ITERACE 29
OBRÁZEK 16 – N-TÁ ITERACE 29
OBRÁZEK 17 – FUNKCE PŘÍSLUŠNOSTI 30
OBRÁZEK 18 – FUNKCE PŘÍSLUŠNOSTI 30
OBRÁZEK 19 – PRŮMĚT PARAMETRŮ(1-12) 36
OBRÁZEK 20 – HISTOGRAMY PRŮMĚTU PARAMETRŮ (1-12) 37
OBRÁZEK 21 – GRAFICKÉ VYJÁDŘENÍ (TABULKA 2) 38
OBRÁZEK 22 – NEJBLIŽŠÍ SOUSED – DENDROGRAM 39
OBRÁZEK 23 – NEJVZDÁLENĚJŠÍ SOUSED – DENDROGRAM 39
OBRÁZEK 24 – TEST VÝBĚRU DAT ENTROPIÍ 40
OBRÁZEK 25 – ÚSPĚŠNOST METODY PRO KOMBINACE PARAMETRŮ 41
OBRÁZEK 26 - ÚSPĚŠNOST METODY PRO KOMBINACE PARAMETRŮ 42
OBRÁZEK 27 - ÚSPĚŠNOST METODY PRO KOMBINACE PARAMETRŮ 42
OBRÁZEK 28 - ÚSPĚŠNOST METODY PRO KOMBINACE PARAMETRŮ 43
OBRÁZEK 29 – VÝSLEDNÁ KOMBINACE 44
OBRÁZEK 30 – METODY PŘI NEJVHODNĚJŠÍCH PARAMETRECH 44
OBRÁZEK 31 – BARGRAF PRO C-MEANS 45
OBRÁZEK 32 – ÚSPĚŠNOST KLASIFIKACE -3 PARAMETRY 46
OBRÁZEK 33 – ÚSPĚŠNOST KLASIFIKACE – 3 PARAMETRY 46
OBRÁZEK 34 - ÚSPĚŠNOST KLASIFIKACE -4 PARAMETRY 47
OBRÁZEK 35 - ÚSPĚŠNOST KLASIFIKACE -4 PARAMETRY 47

Rovnice

EUKLIDOVSKÁ VZDÁLENOST[1](1) 11
HAMMINGOVA VZDÁLENOST[1](2) 12
MINKOVSKÉHO VZDÁLENOST[1](3) 12
CHABYSHEVOVA VZDÁLENOST[1](4) 12
MAHALANOBISOVAVZDÁLENOST[1](5) 12
TĚTIVOVÁ VZDÁLENOST[1](6) 13
DISKRIMINAČNÍ PRAVIDLO[2](7) 14
ZOBECNĚNÍ DPFG[2](8) 14
LDA[2](9) 15
QDA[2](10) 15
SUMA ČTVERCŮ ODCHYLEK[1](11) 21
MEZISHLUKOVÁ SUMA ČTVERCŮ[1](12) 24
PROCENTO VARI[1](13) 24
SENZITIVITA(14) 33

SPECIFICITA(15) 33
PPH(16) 34
NPH(17) 34
ENTROPIE[5](18) 37
SKÓRE (19) 43
HODNOTA PRO GRAF (20) 43

7. POUŽITÁ LITERATURA

[1] MELOUN, M.; MILITKÝ, J. Statistická analýza experimentálních dat. Academia, Praha, 2004. 953 s. ISBN 80-200-1254-0.

[2] MELOUN, M.; MILITKÝ, J. Kompendium statistického zpracování dat. Academia, Praha, 2004. 766 s.

[3] STORK, D. G.; DUDA, R. O.; HART, P. E. Pattern Classification. 2nd edition. [s.l.] : Wiley-Interscience, 2000. 654 s. ISBN 0471056693.

[4] -red- ,Dendrogram. *Cojeco* [online]. Říjen 2000 [cit. 25. dubna 2008]. Dostupné na WWW: <http://www.cojeco.cz/index.php?detail=1&id_desc=18956&s_lang=2&title=dendrogram >.

[5] -,Entropy. *WikipediA* [online]. [cit. 15. května 2009]. Dostupné na WWW: <http://en.wikipedia.org/wiki/Entropy >.

[6] NOVÁK, V. Základy fuzzy modelování. BEN, Praha, 2000. 175s.

[7] ZAPLATÍLEK, K.;DOŇAR, B. MATLAB - Začínáme se signály. BEN, Praha, 2006. 271s.