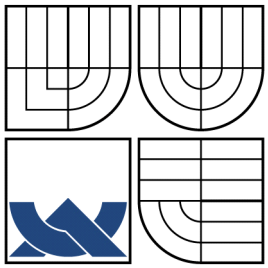
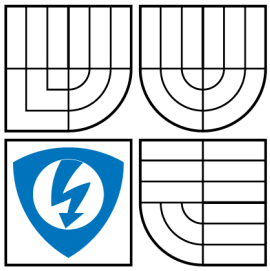


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A
KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV AUTOMATIZACE A MĚŘÍCÍ TECHNIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF CONTROL AND INSTRUMENTATION

# UČENÍ BEZ UČITELE
UNSUPERVISED LEARNING

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE — PETR KONČINSKÝ
AUTHOR
VEDOUCÍ PRÁCE — Ing. PETR HONZÍK, Ph.D.
SUPERVISOR

BRNO 2008

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ
Fakulta elektrotechniky
a komunikačních technologií
Ústav automatizace a měřicí techniky

# Bakalářská práce

bakalářský studijní obor
**Automatizační a měřicí technika**

***Student:*** Končinský Petr
***Ročník:*** 3

***ID:*** 89489
***Akademický rok:*** 2007/08

**NÁZEV TÉMATU:**

## Učení bez učitele

**POKYNY PRO VYPRACOVÁNÍ:**

Proveďte rešerši aplikací využívajících metody učení bez učitele. Dále vytvořte přehled používaných algoritmů a jejich principů. Popište základní metriky a demonstrujte jejich vliv na výslednou klasifikaci. Alespoň jeden z algoritmů naprogramujte.

**DOPORUČENÁ LITERATURA:**

Dle doporučení školitele.

***Termín zadání:*** 1.2.2008

***Termín odevzdání:*** 2.6.2008

***Vedoucí projektu:*** Ing. Petr Honzík, Ph.D.

**prof. Ing. Pavel Jura, CSc.**
*předseda oborové rady*


**UPOZORNĚNÍ:**

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být 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.

ANOTACE V ČESKÉM JAZYCE

Cílem zadání této práce bylo, seznámit se a prozkoumat problematiku učení bez učitele. Učení bez učitele se využívá pro sledování dat, kdy nejsou poskytována data od učitele, tj. chybí trénovací množina. Data, která se sledují, se vysvětlují pomocí matematických modelů-statistický přístup, deterministický přístup. Nejprve je nutné se zaměřit na nejdůležitější části tématu učení bez učitele. Jednou z nejpodstatnějších častí, jíž se celý projekt zabývá, je shluková analýza. Konečným výsledkem podrobného zkoumání a rozebrání shlukové analýzy je jednoduchý program realizovaný v jazyce C ++.

Podstatou analýzy je popis práce s jednotlivými proměnnými, s nimiž jsou prováděny různé matematické operace. Metriky jsou jednou z podstatných částí shlukové analýzy. Zde jsou řešeny vzájemné podobnosti mezi objekty a počítání jejich vzájemné vzdálenosti pro další shluky. Opomíjeny nemohou být ani další metody patřící do učení bez učitele. Další neméně významnou analýzou je analýza hlavní komponent nebo faktorová analýza. Snahou projektu bylo vytvořit průřez metodami, které se podílejí na strojovém učení, v našem případě u učení bez učitele.

KLÍČOVÁ SLOVA

Učení bez učitele

Shluková analýza

Faktorová analýza

Analýza hlavní komponent

ANOTACE V ANGLICKÉM JAZYCE

The aim of the (diploma) thesis was to identify problems of unsupervised learning. At first I focus on the most essential parts of unsupervised learning. One of the most fundamental parts, on which the project is focused on, is cluster analysis. I give a detailed description of cluster analysis and the final product is a single program which was programmed in C++ language. At analysis I describe tasks with particular variables which I use to run various mathematical operations. At metrics which belong to the most important part of cluster analysis I am concerned with mutual similarity relations between objects and I compute their inter-cluster distances for other clusters. Furthermore, I do not omit other methods belonging to unsupervised learning, principal components analysis and factor analysis are not less important. In my project I attempted to do a survey of methods which are involved in machine learning, in our case in unsupervised learning.

KEY WORDS
Unsupervised Learning
Cluster analysis
Factor analysis
Principal components analysis

# Bibliografická citace

KONČINSKÝ Petr. Učení bez učitele. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2008. s., příloh. Vedoucí práce.

# P r o h l á š e n í

„Prohlašuji, že svou diplomovou práci na téma "Učení bez učitele" 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: 02. 06. 2008 Podpis:

# P o d ě k o v á n í

Děkuji tímto Ing. Petru Honzíkovi Ph.D. za cenné připomínky a rady při vypracování diplomové práce.

V Brně dne: 02. 06. 2008 Podpis:

# SEZNAM OBRÁZKŮ



# SEZNAM TABULEK

# 1. ÚVOD

Práce se zabývá strojovým učením, které umožňuje různým systémům se učit. Především se zaměřuje na učení bez učitele, kde se podrobně zabývá popisem jednotlivých metod. Pod pojmem učení, je rozuměna speciální činnost přizpůsobení se ke specifickým podmínkám. Učení je možné definovat jak pro živé organizmy, tak i pro stroje. Podmínky mohou být různého charakteru a je na ně potřeba včas reagovat. Stanovení vhodných podmínek není snadné, ale značně usnadní analýzu. Pro dosažení co nejlepších výsledků jsou důležité metody použité při učení.

U metody učení bez učitele není předem vždy jisté, zda je objekt znám a zda patří do nějaké skupiny předem známých shluků. Reakce na nastalé podněty musí být co nejrychlejší a pokud je to možné tak, aby byly prováděny v reálném čase. Jednou z hlavních podmínek je co nejvíce omezit lidský faktor. Nemusel by vždy poskytnout adekvátní zpětnou vazbu nebo vyhovět fyzickým nárokům. Využitelnost strojového učení je např. v automatizovaných zařízeních, kde je požadavek na rychlou a nepředvídatelnou situaci. Vhodné použití metody „Učení bez učitele“ je také v souvislosti s prozkoumáváním známých procesů, kde z nějakých příčin nejsou dostačující jejich výsledky a je nutné nalézt jiný princip procesu. Cílem je klasifikovat všechny objekty zahrnuté do analýzy. Tento postup je pak obecně označován termínem shluková analýza.

Se strojovým učením nemusí přímo souviset zdroje k získávání dat. Data získáváme z různých databází pocházejících z odlišných oborů. Vzniklé výsledky pak mohou sloužit pro další vědecké zkoumání.

Jednotlivé postupy analýzy jsou zahrnovány do multivariačních technik, které jsou spojené se statistickou analýzou vícerozměrných dat. Účelem vícerozměrné statistické analýzy je porovnání vztahů mezi jednotlivými složkami náhodného vektoru. Analýza se snaží o vysvětlení vztahů mezi proměnnými. Multivariační metody nacházejí uplatnění v průzkumu dat, s cílem identifikovat malý počet zajímavých kombinací, které mohou být zkoumány z prostorového hlediska. Do multivariačních metod patří různé analýzy, které pracují na odlišných principech

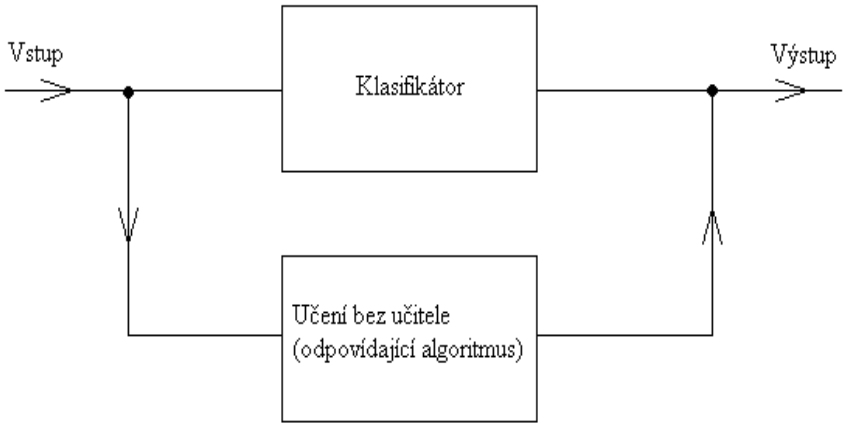
shlukování. Základní analýzy jsou: shluková analýza, analýza hlavní komponent, faktorová analýza, diskriminační analýza a kanonická korelační analýza.

Data mohou být pozorována a analyzována i v případě, že nejsou k dispozici potřebné informace. K získání potřebných informací jsou využitelné matematické modely.

Matematické modely se rozdělují na dvě skupiny:

a) Statistický přístup - určuje statistické modely dat

b) Deterministický přístup – určuje podle jiných směrů podobnosti, např. podle vzdálenosti, využití u shlukové analýzy




**Obr. 1.1.** Blokové schéma učení bez učitele

Pro samostatné učení se používá místo trénovací množiny výstup z klasifikátoru

# 2. SHLUKOVÁ ANALÝZA

Shluková analýza patří mezi metody učení bez učitele. Je to souhrn početních operací zaměřených na rozklad informací (dat). K rozkladu dochází pomocí matematických výpočtů na jednotlivé podmnožiny objektů (shluků).

Cílem metody je v dané množině objektů nalézt její podmnožiny, tzv. shluky objektů. Jednotlivé shluky musí být navzájem stejné, ale ne příliš podobné s objekty mimo tento shluk. Přitom shluky mohou být z různých oblastí, např. fauna, flora, nebo jako datové soubory určené k práci na PC. Pro zjednodušení se celá analýza roztřídí do několika podsouborů, tzv. klasifikaci, která má za následek vytvoření systému tříd. Tento způsob eliminuje jednoduchost úlohy tak, že jednotlivé proměnné se nahradí příslušností k nové třídě.

V jednotlivých třídách se nacházejí objekty, které jsou reprezentovány nějakou vlastností. V praxi se stanoví proměnné veličiny (znaky), které je potřeba sledovat. Proměnné mohou být různých typů:

a) číselné: délka, šířka, např. krabice
b) ordinální: hodnoty lze uspořádat, ale není podmínka, aby hodnoty nabývaly číselných hodnot. Intenzita silového pole může být slabá, střední, silná, případně žádná. Obdobně lze charakterizovat i velikost lidí - malý, střední, velký.
c) nominální: zde hodnoty uspořádat nejdou. Jako příklad lze použít zvířata, která budou charakterizována pomocí chování - hodná, zlá, smutná, veselá a podle potřeby je možné i více specifikovat – hodná, ale smutná.
d) dichotomická: má pouze dvě hodnoty. Ano, ne. „ +, - “, „0, 1“, popřípadě jiného slovního vyjádření – známá, neznámá

Lze také uvažovat popis objektů pomocí znaků, které se označují jako symbolické, jejichž hodnoty mohou být v intervalu nebo v pravděpodobnostním rozložení s určitými parametry. Nastane-li situace, kdy je známa jedna konkrétní hodnota, pak taková data označíme jako fuzzy [2].

# 3. MEZI SHLUKOVÉ VZDÁLENOSTI

## 3.1 SHLUKY

Pojem shlukování je spojován se shlukováním objektů a sleduje se zde podobnost vektorů, které tvoří řádky matice. Jedná se o více možností. Při analýze dat se mohou hledat také shluky proměnných. Se zmíněnými variantami je možné se setkat obvykle v oblasti informatiky. Aby celý proces mohl fungovat, je nutné si stanovit kritéria, která budou určovat základní pravidla při shlukování. Jedním z pravidel je stanovit si cíl shlukování. Vhodné rovněž je určit, jakých typů shluků je třeba dosáhnout. Požadavek je zařazen jako objekt do určitého počtu shluků, nebo vytvoří hierarchii shluků.

Podle požadavků se rozlišuje shlukování na:

- nehierarchické
- hierarchické

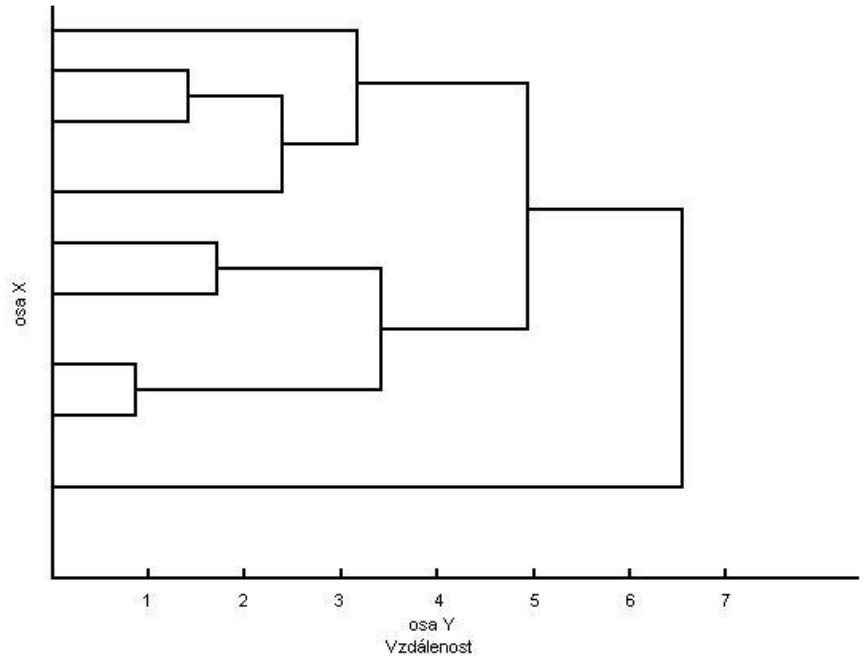
Kritérium pro stanovení počtu shluků vychází z předpokladu nehierarchického shlukování, u kterého je obvykle zadán počet shluků. Při zkoumání struktury v datech nemá většinou uživatel o počtu shluků žádnou základní informaci. Je tedy možné využít shlukování pro různé počty shluků. Ze získaných výsledků je určen optimální počet. Dosažený výsledek lze graficky zobrazit za pomocí dendrogramu. Jednou z možností je i určení počtu shluků podle grafu. Nastávají i situace, kdy není možné určit z grafu počet shluků, proto v případě velkého množství objektů přicházejí v úvahu metody určování optimálního počtu.

V některých případech bývá užitečné rozdělit objekty do různých počtů shluků. To má souvislost i s interpelací shluků, kdy každý z uvažovaného počtu shluků může mít reálnou interpelaci. Interpelace námi získaných výsledků je spjata se stanovením počtu shluků. Zmíněný princip se využívá při velkém počtu objektů. Nejprve je nahrazen velký počet objektů malým počtem shluků a následně je provedeno shlukování v každé skupině zvlášť.

Výsledek shlukování objektů je vždy ovlivněn výběrem proměnných, případně jejich vahami. Váhy proměnných jsou při shlukové analýze přiděleny.

Dále je výsledek závislý na tom, zda datový soubor obsahuje objekty odlehlé, tj. velmi odlišné od ostatních. Při použití základních shlukovacích algoritmů tvoří také objekty samostatné shluky.
Pokud jsou hojně zastoupeny, pak zvyšováním počtu shluků získáváme další a další jednoobjektové shluky. Neodlehlé objekty naopak tvoří stále jeden shluk.

Shluky se postupně spojují a znázorňují ve speciálním grafu, který se jmenuje dendrogram. Graf lze popsat jako stromový diagram znázorňující shlukování. Zobrazují se jak jednotlivé objekty, tak i shluky vytvořené předem. Dendrogram se zobrazuje v horizontální nebo vertikální poloze. V horizontální poloze jsou objekty uvedeny na ose Y a ve vertikální poloze na ose X. Bereme-li v úvahu, že dendrogram je popsán pomocí stromu, ale z opačné strany, tedy od kořene k větvím, pak jejich další rozdělování je možné popsat tak, že na ose Y je zakresleno *n* listů a z těchto listů vycházejí větve. Větvě, mezi kterými je nejmenší vzdálenost, se spojí do jedné a jejich spojením vznikne hodnota na ose X, což je vzdálenost [2].




**Obr. 3.1.** Dendrogram

## 3.2 OBJEKTY A ZNAKY

Pomocí klasifikace jsou určeny předměty nebo jevy a každý objekt pak je možné popsat *n*-ticí stavů předem stanovených *n*-stavů. Stavům se přiřazují číselné kódy a ty pak představují hodnoty znaků. Výsledkem shlukové analýzy je *n* rozměrný vektor čísel. Objekty mohou být například z oblasti fauny či flory nebo z běžného života [7].

Znaky lze zařadit do množin objektů, které budou následně klasifikovány do množin stavů. Znaky objektů a množiny jejich stavů mohou být:

- kvalitativní znaky: konečná množina bodů, kterým mohou být přiřazeny číselné kódy: nominální, např. barva, a ordinální (dají se uspořádat). Binární (dichotomické): pravda – nepravda, pohlaví: žena – muž
- kvantitativní znaky: interval nacházející se v oblasti reálných čísel, např. teplota

## 3.3 KLASIFIKACE OBJEKTŮ

Princip klasifikace objektů je založen na vlastnostech klasifikovaných objektů. Je-li použit nějaký náhodný vektor, který bude normálně rozdělen podle náhodného vektoru, v němž se nacházejí charakteristické objekty, pak se jedná o parametrické klasifikační metody. V opačném případě nastává nepravidelné rozdělení náhodného vektoru, a pak se jedná o neparametrické klasifikační metody.

Mezi objekty se tedy hledají vzájemné vazby, podle kterých se následně zařazují do již vzniklých tříd. Této problematice se věnuje diskriminační analýza. Nebo objekty mohou být neuspořádané a pomocí klasifikačních metod je uspořádáme do sourodých tříd-shluků. Klasifikace objektů patří do vícerozměrné statistické analýzy [1].

### 3.4 STANDARDIZACE DAT

Jednotlivé hodnoty znaků objektů mohou nabývat různých hodnot nebo se objeví v různých jednotkách. To se projeví, když se jednotlivé znaky (objekty), budou jevit jako dominantní nebo méně dominantní. Z toho důvodu se provádí standardizace, aby lidský faktor co nejméně ovlivnil celý proces shlukování. Data jsou upravena tak, aby všechny znaky byly co nejvíce stejné.

Pro příklad lze použít libovolnou matici $Z=(z_{ij})$ typu $n$ x $p$, jejíž řádky jsou $p$-rozměrné vektory čísel charakterizující $n$ objektů. Standardizace je pak provedena ve dvou krocích [7].

a) Nejprve se vypočte střední hodnota $\bar{z}_j$ $j$-tého znaku $z_j$ a směrodatnou odchylkou $s_j$ a pro $j$=1,2,… p podle vzorců:

$$\bar{z}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij} \tag{2.1}$$

$$s_j = \left[\frac{1}{n}\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2\right]^{\frac{1}{2}} \tag{2.2}$$

b) Původní hodnoty $z_{ij}$ j-tého znaku $i$-tého objektu se přepočtou na tzv. standardizované hodnoty:

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j} \tag{2.3}$$

pak znaky mají střední hodnotu rovnou 0 a rozptylu 1

### 3.5 NORMALIZACE OBJEKTŮ

Všechny objekty pro shlukovou analýzu jsou dány vektory o *p* složkách, představující hodnoty vybraných *p* znaků. Tyto vektory mohou mít nepříznivý vliv na výsledky kvantitativního hodnocení podobností objektů. Proto se v takovýchto případech zavádí normalizace, aby měly vektory stejnou normu, nejlépe jednotkovou [2].

### 3.6 PODOBNOST OBJEKTŮ

U shlukové analýzy nastávají nejčastěji problémy se vzájemným pojetím jednotlivých podobností objektů a následně jejich kvantitativním vyjádřením podobnosti. K určování podobnosti objektů se využívají míry podobnosti, ale častěji míry nepodobnosti. Míru podobnosti pro objekty $x_i$ a $x_j$ se zapíší jako $S(x_i, x_j)$ nebo ve zkráceném tvaru $S_{ij}$. Potom platí, že $S_{ij} = S_{ij,}$ a jedná se o vzájemný vztah. Může nastat varianta, že hodnoty budou nabývat z intervalu < 0; 1>, poté platí, že $S_{ij}=1$. Jestliže je vytvořena matice se vzájemným ohodnocením podobností pro všechny dvojice objektů, vznikne matice s jedničkami na diagonále [2].

Nepodobnost $x_i$ a $x_j$ se zapíše jako $D(x_i, x_j)$ nebo pro zjednodušení lze použít zápis ve zkráceném tvaru $D_{ij}$ a platí, že:

1. $D_{ij} \geq 0$
2. $D_{ij} = 0$ zde jsou v matici všechny prvky na diagonále nulové
3. $D_{ij} = D_{ij}$ matice je symetrická

Pro naše účely se vychází z míry nepodobnosti objektů, což vede k metrice představující míry nepodobnosti objektů.

### 3.7 METRIKY

Patří k nejběžnějším způsobům vyjádření podobností vztahů mezi objekty. Vychází z geometrické podobnosti dat. Je-li dána množina bodů a k nim přiřazen charakteristický znak *n* pro model bodů *n*-rozměrného Euklidovského prostoru $E_n$, je pro libovolné dva body *(a, b)* definována euklidovská vzdálenost [2]:

$$\delta(a,b) = \left[ \sum_{i=1}^{n} (x_{ai} - x_{bi})^2 \right]^{\frac{1}{2}} \tag{2.4}$$

Je-li metrika definována obecně na $E_n$ x $E_n$ a ke každé dvojici bodů *(a, b)* je přiřazeno číslo $\delta(a, b)$, pak splní tyto čtyři podmínky:

$\delta(a, b) = 0 \Leftrightarrow a = b$, jsou identické
$\delta(a, b) \geq 0$
$\delta(a, b) = \delta(b, a)$, jsou symetrické
$\delta(a, c) \leq \delta(a, b) + \delta(a, c)$, mají trojúhelníkovou nerovnost

Jedním z typů Euklidovské vzdálenosti je čtvercová euklidovská vzdálenost:

$$E(a_i, b_j) = \sum_{l=1}^{m} (x_{il} + x_{jl})^2 \tag{2.5}$$

Vzdálenosti lze také počítat i podle jiných kritérií, např. pomocí Manhattanské vzdálenosti:

$$M(a_i, b_j) = \sum_{l=1}^{m} \left| x_{il} - x_{jl} \right| = \left| x_i - x_j \right| \tag{2.6}$$

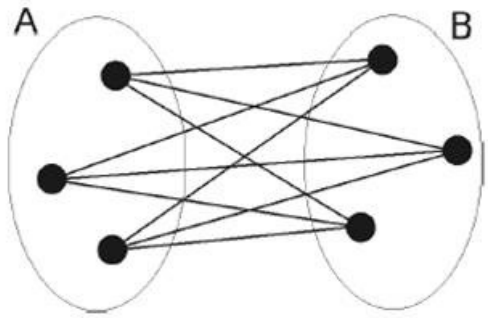
Nebo se použije Minkowského počítání vzdáleností. Tento algoritmus v sobě kombinuje dvě metody počítání vzdáleností. Kombinuje Euklidovskou vzdálenost a Manhattanskou vzdálenost:

$$M(a_i, b_j) = \sqrt[q]{\sum_{l=1}^{m} \left| x_{il} - x_{jl} \right|^q} \tag{2.7}$$

Koeficient *q* určuje, která z metod bude použita. Jestliže je za *q* dosazena 1, bude se počítat Manhattanská vzdálenost. V případě, že je za *q* dosazena 2, počítá se Euklidovská vzdálenost.

Pro vyjádření nových podobností shluků se používají nejčastěji tyto metody:

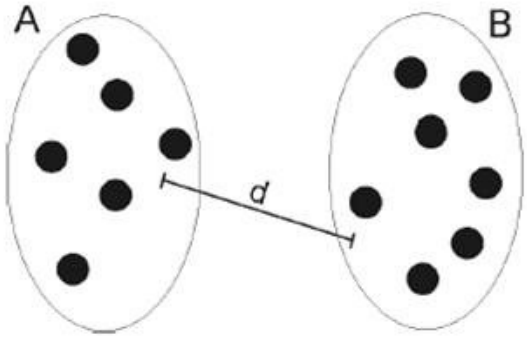
- Metoda průměrné vazby pro mezishlukové vzdálenosti – k výpočtu vzdálenosti mezi dvěma shluky se využívá aritmetický průměr. Aritmetický průměr je ze všech možných vzdáleností objektů, a to jeden patří do prvního shluku a druhý do druhého shluku.




**Obr. 3.2.** Grafické zobrazení průměrové metody

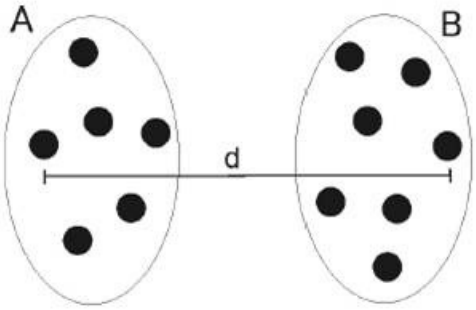
- Metoda průměrné vazby pro vnitroshlukové vzdálenosti – nejprve se určí oblast dvou shluků a spojí se do jednoho shluku a poté se počítá vzdálenost pomocí aritmetického průměru.

- Metoda nejbližšího souseda – pro počítání vzdálenosti shluků musí být dána minimální vzdálenost objektů.




**Obr. 3.3.** Grafické zobrazení metody nejbližšího souseda

- Metoda nejvzdálenějšího souseda – pro počítání vzdálenosti se využívá maximální vzdálenosti objektů




**Obr. 3.4.** Grafické zobrazení metody nejvzdálenějšího souseda

- Centroidní metoda – k výpočtu vzdálenosti se využívá euklidovská vzdálenost mezi centroidy, což jsou vektory aritmetických průměrů počítané u objektů uvnitř shluku.

Algoritmy se popisují pomocí vzdálenosti mezi *g*-tým shlukem a sjednocením shluků *h, h'*

Metoda průměrné vazby:

$$D_{g\langle hh'\rangle} = \frac{n_h}{n_h + n_{h'}} D_{gh} + \frac{n_{h'}}{n_h + n_{h'}} D_{gh'} \tag{2.8}$$

Metoda nejbližšího souseda:

$$D_{g\langle hh'\rangle} = \frac{1}{2}(D_{gh} + D_{gh'} - |D_{gh} - D_{gh'}|) \tag{2.9}$$

Metoda nejvzdálenějšího souseda:

$$D_{g\langle hh'\rangle} = \frac{1}{2}(D_{gh} + D_{gh'} - |D_{gh} - D_{gh'}|) \tag{2.10}$$

Centroidní metoda:

$$D_{g\langle hh'\rangle} = \frac{n_h}{n_h + n_{h'}} D_{gh} + \frac{n_{h'}}{n_h + n_{h'}} D_{gh'} - \frac{n_h n_{h'}}{(n_h + n_{h'})^2} D_{hh'} \tag{2.11}$$

# 4. METODY SHLUKOVÉ ANALÝZY

## 4.1 HIERARCHICKÉ SHLUKOVÁNÍ

Pracuje na principu podmnožin, kde průnikem dvou podmnožin-shluků je buď prázdná množina, nebo jeden z nich - systém je hierarchický. Získané sekvence vnořených rozkladů začínají jednoduchým rozkladem, kde každý prvek dané množiny objektů tvoří jednoprvkový shluk a končí jednoduchým rozkladem s jedním shlukem obsahujícím všechny objekty. Tedy jedná se o jakési zjemňování, větvení klasifikace [7].

U hierarchického shlukování se přístupy rozdělují:

- Monotetický - využívá se principu srovnání shluků do určité jednotné hladiny a z té se shluky vytvoří podle jedné z proměnných.

- Polytetický - současně se využívají u shluků všechny proměnné najednou.

Dalším kritériem pro členění je analýza podobnosti, která se dělí podle přístupu na:

- Divizní algoritmus – lze rozdělit do dvou kroků. V prvním kroku všechny objekty tvoří jeden shluk. Ve druhé části dochází k postupnému rozdělování shluků až do stavu, kde každý objekt je samostatným shlukem.

- Aglomerativní algoritmus - pracuje částečně na opačném principu než divizní algoritmy. Počáteční stav aglomerativního algoritmu je, když objekt je samostatným shlukem. S tímto shlukem se spojují další shluky a vytvářejí dvojice od nejvíce až k nejméně podobným. Konečným výsledkem je pak samostatný jeden shluk.

### 4.1.1 Divizní hierarchické shlukování

Divizní metoda rozkládá množiny objektů postupným rozdělováním existujících shluků. Při řešení divizního algoritmu se postupuje tak, že za počáteční shluk se považuje celá množina objektů. Dále se rozdělují existující shluky tak, až jsou všechny shluky jednoprvkové. Princip spočívá v postupném rozdělování každého existujícího shluku na dva nové. Rozdělují se tak, aby výsledný rozklad tohoto shluku byl optimální vzhledem k danému kritériu. Divizní metoda je optimální pro malý počet objektů. Najít optimální rozklad nějaké množiny o *n* prvků na dvě podmnožiny vyžaduje pravděpodobnost průzkumu na $2^{n-1}$-1 možností [2].

### 4.1.2 Aglomerativní hierarchické shlukování

Vychází se z možnosti, že na počátku je každý objekt samostatným shlukem. Při spojování shluků se vychází z předpokladu, že se spojí dva nejpodobnější shluky. V první časti je shluk vytvořen ze dvou objektů. Objekty se tvoří z matice podobnosti nebo nepodobnosti. Pro naše účely je výhodnější matice nepodobnosti. Potom počítáme nepodobnost shluků, např. aglomerativní algoritmy. Přiklad aglomerativních algoritmů: Metoda průměrné vazby, Metoda nejbližšího souseda, Metoda nejvzdálenějšího souseda, Centroidní metoda [2].

## 4.2 NEHIERARCHICKÉ SHLUKOVÁNÍ

Základem je nalézt optimální počet shluků. Pro hledání rozkladu se použije vhodný index. Hodnota indexu se nastaví tak, že je nalezena závislost na proměnném parametru *K*. K hledání závislostí se použije, např [4].:

a) Calinski-Harabascův index - pro optimální K se hledá maximum CH (K)

$$CH(K) = \frac{n-K}{K-1}\left[\frac{E_1^2}{E_K^2} - 1\right] \tag{3.1}$$

*n*-počet obrazů
*K*-počet shluků
$E_K^2$ - kvadrát přříslušné o funkcionálu

b) index je normalizovaný tvar Γ statistiky

$$\Gamma = \sum_{q=1}^{n-1}\sum_{r=q+1}^{n} d(q,r)c(q,r) \tag{3.2}$$

Kde *c(q, r)* = 1 platí pouze jestliže $x_{q\_}$ a $x_{r\_}$ jsou ve stejném shluku, v opačném případě *c(q, r)*=0.

*d(q, r)* je nepodobnost nebo Euklidovská vzdálenost mezi *q* a *r*.

$$C(K) = \frac{\Gamma - \min(\Gamma)}{\max(\Gamma) - \min(\Gamma)} \tag{3.3}$$

kde: min(Γ) je suma $a_k$ nejmenších nepodobností
max(Γ) je suma $a_k$ největších nepodobností

### 4.3 POČÁTEČNÍ ROZKLAD

Při provádění shlukové analýzy se vyskytuje problém při počátečním rozkladu. Bude-li určen počáteční rozklad *k*-shluků, lze zachovat stejný počet shluků nebo se provede změna počtu k v průběhu výpočtu v závislosti na řídících proměnných algoritmech. Důraz je kladen u těch algoritmů, kdy už nedochází ke změnám v počtu shluků k. Pomocí klasifikace jsou stanoveny vzorkovací objekty, např. v Euklidovském prostoru, kde je předpoklad vytvoření shluků. Kvalita rozkladu může být ovlivněna hned v počátečním stádiu tak, aby shlukování bylo co nejkvalitnější. Pro výběr bodů zavedeme co nejjednodušší kritéria [5]:

- První *k*-body z uspořádané množiny bodů jsou libovolně vybrány
- vytváří se *k* umělých bodů tak, aby souřadnice byla náhodné číslo z určitého intervalu.
- U prvních bodů se vygeneruje $(2^p + 1)$, kde p je počet rozměrů prostorů. První je těžiště celé množiny bodů a ostatní jsou vrcholy *p*-rozměrného kvádru se středem v těžišti a hranami délky $2s_i$, kde $s_i$ je směrodatná odchylka hodnot *i*-té souřadnice. Podmínkou je $2^p > n$, kde *n* je počet bodů.

Iterace se provádí do doby, dokud nedojde ke stabilní konfiguraci, rozkladu na výsledné shluky.

### 4.3.1 Dělící metody

Dělící metody, též někdy nazývané separační metody, mají za cíl rozčlenit objekty do shluků. Vždy se dělení provádí podle nějakého kritéria, nejčastěji se využívají podobnostní funkce, které pracují na principu vyhledávání globálního minima nebo maxima. Vyhledávání dat ve velkém prostoru je velmi náročné, a proto se optimalizuje velikost prostoru vhodnými algoritmy, aby bylo nalezeno globální optimum.

Dělící metoda pracuje tak, že se vyberou objekty, které budou představovat středy shluků. Středy shluků jsou těžiště objektů a počítají se jejich vzdálenosti od těžiště. Metoda je vhodná pro malý počet shluků z velkého množství objektů.
K nevýhodám patří volení počátečního počtu shluků. Při velkém počtu může vzniknout nežádoucí šum. Nejpoužívanější algoritmy jsou K-Menas, K-Medoids [4].

## 4.4 MŘÍŽKOVÁ METODA

Využívá dělení objektů do vícerozměrné mřížkové datové struktury. Dělení se provádí až na konečný počet buněk. Celý proces se provádí nad mřížkovou strukturou. Rychlost zpracování je velmi rychlá a operace probíhá nezávisle na počtu buněk v mřížkové struktuře [4].

## 4.5 METODA ZALOŽENÁ NA HUSTOTĚ

Metoda je založena na principu formování shluků podle hustoty objektů v daném prostoru. Hustotu ovlivňují dva parametry. První parametr ε určuje maximální počet sousedů. Druhý parametr MinPts udává minimální počet sousedů. Důležité u této metody je definovat si shluk. V tomto případě je shluk maximální množina propojených objektů.

Metoda založená na hustotě pracuje tak, že nejprve provede kontrolu okolí každého bodu, jestliže okolí splňuje podmínku prvního parametru MinPts, nebo druhého parametru ε se vytvoří nový shluk. Celý proces se ukončí až tehdy, když už se nedá přidat žádný bod do shluku. Výhodou je odolnost proti šumu a nalezení různých tvarů shluků. Shluk nemusí mít pravidelný tvar [5].

# 5. DALŠÍ METODY UČENÍ BEZ UČITELE

## 5.1 ANALÝZA HLAVNÍ KOMPONENT

Cílem této metody je určit základní proměnné tak, aby nedocházelo ke ztrátě informací. Analýza hlavní komponent pracuje na principu přeměny proměnných na nové proměnné, které se pak označují jako komponenty. Tyto komponenty jsou zařazeny podle informací nezávisle tak, aby srozumitelně vyjadřovaly celkový rozptyl proměnných. Nevýhodou metody je dodržování daných měřítek, protože je velmi citlivá na každou změnu. Je proto nutné vždy provádět normalizaci původních dat.

Analýzu charakterizují dva hlavní komponenty, které udávají její vlastnosti. První komponenta udává u proměnných největší část přeměny. Druhá komponenta udává u proměnných druhou největší část přeměny. Proces probíhá až do té doby, dokud nejsou dokončeny všechny přeměny proměnných.

Důraz se klade na hlavní komponent a snahou je mu pro lepší pochopení přiřadit nějaký skutečný význam. Komponenty, které stojí v pozadí, reprezentují zobecněné vlivy a vyvolávají variabilitu a ovlivňují strukturu závislosti proměnných. Při vysvětlování se využívá především korelace s původními proměnnými.

Celá analýza hlavní komponent je transformace z původního stavu do nového stavu, kde základem přeměny jsou komponenty [6].

## 5.2 FAKTOROVÁ ANALÝZA

Faktorová analýza vyhledává proměnné i v oblastí, kde je větší pravděpodobnost nějakého rušení. Nachází i nepozorovatelné proměnné a následně z nich vytváří faktory. Ve faktorové analýze předpokládáme, že každou vstupující proměnnou můžeme vyjádřit jako lineární kombinaci malého počtu společných skrytých faktorů a jediného chybového faktoru. U analýzy je každá hodnota porovnávána s proměnou, která má alespoň nějaký společný faktor. Jednotlivé faktory určují změnu dvou proměnných [6].

## 5.3 POROVNÁNÍ ANALÝZY HLAVNÍCH KOMPONENT A FAKTOROVÉ ANALÝZY

Faktorová analýza a analýza hlavní komponent se snaží zredukovat rozměrnost různých skupin dat. Hlavní rozdíl mezi oběma analýzami je, že analýza hlavní komponent vysvětluje veškerou změnu mezi proměnnými a faktorová analýza se zaměřuje pouze na společné proměnné. Je tedy možné říci, že cílem analýzy hlavní komponent je odvození malého množství kombinací (hlavních komponentů) z množiny proměnných při zachování co nejvíce informací obsažených v původních proměnných. Cílem faktorové analýzy je vysvětlit rozdíly mezi proměnnými pomocí malého množství proměnných.

## 5.4 DISKRIMINAČNÍ ANALÝZA

Diskriminační analýza se zabývá závislostí mezi množinami nezávisle proměnných a kvalitativně závislými proměnnými. Nezávislé proměnné se nazývají diskriminátory. Výhodné je, že objekty můžeme zařadit rovnou do již existujících tříd. Jednotlivé objekty jsou přiřazovány do tříd pomocí míry podobnosti. Nejčastěji se využívá metoda nejmenší Mahalanobisovy vzdálenosti. Diskriminátory, které vstupují do analýzy, mají již dány hodnoty zařazené do objektů v primárních třídách. Následně vznikají nezařazené objekty a pro ně se bude hledat nové zařazení do tříd.

Zařazování objektů do tříd se řídí, tzv. diskriminačními pravidly. Pravidla vyčíslují hodnoty diskriminačních funkcí. Funkce pak slouží ke zjednodušení zařazování objektů do primárních tříd. Výsledky lze také použít k roztřídění nezařazených objektů do předem známých tříd. Všechny třídy jsou popsány nějakou funkcí, např. primární třídy - bývají často popsány funkcí hustoty pravděpodobnosti [1].

## 5.5 KANONICKÁ KORELAČNÍ ANALÝZA

Analýzu zařazujeme do vícerozměrných metod. Používá se především ke zkoumání dvou skupin proměnných. První skupina proměnných bude pro názornost označena *y,* a ta tvoří závislé proměnné, druhá skupina bude označena *x,* a ta tvoří nezávislé proměnné. Principem analýzy je nalézt v každé skupině proměnných koeficienty *a* a *b*, aby platilo pro všech *n* objektů vyčíslení kanonických proměnných. Ty pak musí určovat maximální prahový koeficient. Jestliže jsou nalezeny tyto proměnné, hledají se další lineární kombinace kanonických proměnných a ty mají druhý největší korelační koeficient. Cílem koeficientů *a* a *b* je, aby maximálně hledaly korelace mezi proměnnými. Koeficienty *a* a *b* jsou námi libovolně zvolené pro lepší vysvětlení analýzy.

První kanonická korelace je maximální korelace mezi kombinacemi nezávisle proměnných *x* a závisle proměnných *y*. Je možné si představit, že se jedná o analogii vícenásobného korelačního koeficientu, kde probíhá korelace mezi jedinou závislou proměnnou *y* a souborem nezávislých proměnných *x* [1].

# 6. SHRNUTÍ A VYUŽITÍ METOD U UČENÍ BEZ UČITELE

U učení bez učitele se využívá mnoho různých metod, jak rozpoznávat data. Jednou ze základních analýz patřících do učení bez učitele je shluková analýza. Tato analýza pracuje na principu hledání dat, která jsou si nějak podobná, ale zároveň se odlišují jinak od ostatních skupin dat. Slučuje objekty do shluků se stejnými vlastnostmi a vytváří tak skupiny dat, která mají stejné vlastnosti než jiné skupiny. Jednotlivé vlastnosti jsou stanoveny podle určitých pravidel – metrik. Nejjednodušší metriku u shlukové analýzy tvoří míry vzdálenosti, nebo - li podobnosti, definované na *n* rozměrném Euklidovském prostoru. Zde se vychází ze vzájemných geometrických podobností. Celá shluková analýza slouží k získávání klasifikace, což lze využít k rozdělování prostředků podle cílů, k nimž směřují, a není nutné využívat matematických prostředků. Proto se rozlišují dvě skupiny metod shlukové analýzy: hierarchické a nehierarchické metody. Například, je-li použita hierarchická shluková analýza, počítají se vzdálenosti - podobnosti v matici, kde se porovnávají dvojice proměnných. Dále se porovnávají v matici vzdálenosti mezi jednotlivými proměnnými a už vzniklými shluky proměnných. Tento postup je velmi výhodný k použití a ke zjišťování, jak podobnosti proměnných, tak i podobností objektů. Shluková analýza je prvním stupněm použitým k roztřídění dat a nemá za úkol stanovit konečné závěry. Spektrum použití shlukové analýzy je velmi rozsáhlé a zasahuje do různých vědních oborů. Hlavním pilířem je ekonomické odvětví, které je využíváno ke statistickému rozdělování dat. Vhodné je i použití ve zdravotnictví, kde se uplatňuje k vyhledávání vzájemně podobných šroubovic DNA, ale uplatnění nachází i v lesnictví.

Shluková analýza je jen jedna z řady metod, které patří do učení bez učitele. Další metodou je faktorová analýza. Pomocí jiných principů zjišťuje podobnosti kvantitativních proměnných. Ke zjišťování využívá další metodu, a to metodu hlavní komponent, která snižuje rozměry proměnných. Celé to pracuje tak, že z velkého

množství proměnných se vybere menší počet hlavních komponent a ty se použijí pro další výpočty.

Další metody, které můžeme zařadit do učení bez učitele jsou diskriminační analýza a kanonická korelační analýza. Diskriminační analýza se zabývá metodami zkoumání mezi závisle proměnnými a nezávisle proměnnými. Její výhoda spočívá v tom, že dokáže zařadit objekty - shluky do předem existujících tříd. Kanonická korelační analýza pracuje částečně na stejném principu jako diskriminační analýza. Rozdíl je, že kanonická analýza porovnává závislosti dvou skupin proměnných. První skupina je, např. závisle proměnná a druhá nezávisle proměnná.
Metody zařazené v učení bez učitele se využívají i v běžném životě a při tom si to ani neuvědomujeme. V každodenních rutinách jsou využívány různé přístroje, které zpracovávají i velké množství informací, např. záznam obrazu pomocí videokamery nebo hlasový záznamník.

# 7. PŘÍKLAD NA SHLUKOVOU ANALÝZU

Postup práce při shlukování je popsán na velmi jednoduchém příkladu, který má pouze za úkol objasnit základní principy při shlukování. Celý proces shlukování je realizován pomocí Euklidovské vzdálenosti, která je ještě pro příklad nahrazena Manhattanskou vzdáleností. Výpočet nových shluků je počítán metodou aritmetického průměru a metodou nejbližšího souseda. Algoritmy a metody příkladu jsou naprogramovány v jazyce C++.

Při vytváření příkladu bylo postupováno tak, že bylo náhodně vytvořeno a zakresleno pět shluků do grafu, na které byly aplikovány algoritmy a metody. Atributy mohou být získány i z jiných zdrojů, není podmínkou je odečítat z grafu. Shluky byly pro jednoduchost rozděleny na dvě části. První část shluků označena jako L, je tvořena třemi shluky. Druhou část tvoří pouze dva shluky s označením S. Poté se pro každý samostatný shluk odečtou z grafu příslušné atributy, se kterými se dále počítá.

Postup:

Odečtené atributy pro L a S z grafu – v tabulce:

**Tab. 7.1.** Tabulka se vstupními atributy

| | X | Y |
|---|---|---|
| $L_1$ | 2 | 9 |
| $L_2$ | 3 | 7 |
| $L_3$ | 4 | 8 |
| $S_1$ | 6 | 5 |
| $S_2$ | 5 | 4 |

Z tabulky je vytvořena vstupní matice, ze které bude následně počítána Euklidovská vzdálenost nebo Manhattanská vzdálenost.

$$\begin{bmatrix} 2 & 9 \\ 3 & 7 \\ 4 & 8 \\ 6 & 5 \\ 5 & 4 \end{bmatrix}$$

Nejprve je proveden výpočet pomocí Euklidovské vzdálenosti podle vztahu:

$$E = \sqrt{\sum_{l=1}^{m} (x_i - x_j)^2} \tag{6.1}$$

Poté jednotlivé spočítané vzdálenosti jsou:

$$L_1, L_2 = \sqrt{(2-3)^2 + (9-7)^2} = 2{,}23$$
$$L_1, L_3 = \sqrt{(2-4)^2 + (9-8)^2} = 2{,}23$$
$$L_1, S_1 = \sqrt{(2-6)^2 + (9-5)^2} = 5{,}65$$
$$L_1, S_2 = \sqrt{(2-5)^2 + (9-4)^2} = 5{,}83$$
$$L_2, L_3 = \sqrt{(3-4)^2 + (7-8)^2} = 1{,}41$$
$$L_2, S_1 = \sqrt{(3-6)^2 + (7-5)^2} = 3{,}6$$
$$L_2, S_2 = \sqrt{(3-5)^2 + (7-4)^2} = 3{,}6$$
$$L_3, S_1 = \sqrt{(4-6)^2 + (8-5)^2} = 3{,}6$$
$$L_3, S_2 = \sqrt{(4-5)^2 + (8-4)^2} = 4{,}12$$
$$S_1, S_2 = \sqrt{(6-5)^2 + (5-4)^2} = 1{,}41$$

Z jednotlivých vzdáleností je sestavena matice, tzv. matice vzdáleností, která má na diagonále 0.

$$\begin{bmatrix} 0 & 2,23 & 2,23 & 5,65 & 5,83 \\ 2,23 & 0 & 1,41 & 3,6 & 3,6 \\ 2,23 & 1,41 & 0 & 3,6 & 4,12 \\ 5,65 & 3,6 & 3,6 & 0 & 1,41 \\ 5,83 & 3,6 & 4,12 & 1,41 & 0 \end{bmatrix}$$

Z matice vzdáleností se vybere nejmenší vzdálenost mezi shluky. V našem případě vznikla situace, kde existují dvě stejné nejmenší vzdálenosti a to mezi $L_2$,$L_3$ s $S_1$,$S_2$ jejichž hodnota je 1,41. Nezáleží na tom, která vzdálenost je zvolena. Například lze vybrat $L_2$,$L_3$.

Ze shluku $L_2$,$L_3$ se spočítá nový shluk, se kterým se budou počítat další vzdálenosti. Výpočet nového shluku se provede pomocí aritmetického průměru nebo metodou nejbližšího souseda. Hodnoty $L_2$,$L_3$ se získají ze vstupní matice.

Hodnoty $L_2$,$L_3$ : $L_2$ [3 7]
$L_3$ [4 8]

Nový shluk se počítá pomocí aritmetického průměru:

výpočet atributu X:

$$\frac{3+4}{2} = 3,5$$

výpočet atributu Y:

$$\frac{7+8}{2} = 7,5$$

Ze získaných nových atributů se vytvoří nový shluk a označí se, např. jako A[3,5 7,5]

Tento postup vyeliminoval původní vstupní matici na novou matici, která bude o jeden shluk kratší. S novou vstupní maticí se počítají opět vzdálenosti stejným způsobem jako doposud.

Nová vstupní matice:

$$\begin{bmatrix} 2 & 9 \\ 3{,}5 & 7{,}5 \\ 6 & 5 \\ 5 & 4 \end{bmatrix}$$

Výpočet vzdáleností opět pomocí Euklida:

$$L_1, A = \sqrt{(2-3{,}5)^2 + (9-7{,}5)^2} = 2{,}12$$
$$L_1, S_1 = \sqrt{(2-6)^2 + (9-5)^2} = 5{,}65$$
$$L_1, S_2 = \sqrt{(2-5)^2 + (9-4)^2} = 5{,}83$$
$$A, S_1 = \sqrt{(3{,}5-6)^2 + (7{,}5-5)^2} = 3{,}53$$
$$A, S_2 = \sqrt{(3{,}5-5)^2 + (7{,}5-4)^2} = 3{,}8$$
$$S_1, S_2 = \sqrt{(6-5)^2 + (5-4)^2} = 1{,}41$$

Nyní je opět vytvořena matice vzdálenosti s nulami na diagonále.

$$\begin{bmatrix} 0 & 2{,}12 & 5{,}65 & 5{,}83 \\ 2{,}12 & 0 & 3{,}53 & 3{,}8 \\ 5{,}65 & 3{,}53 & 0 & 1{,}41 \\ 5{,}83 & 3{,}8 & 1{,}41 & 0 \end{bmatrix}$$

Z této matice je proveden výběr nejmenší vzdálenosti, která je $S_1$,$S_2$ s hodnotou 1,41.

Hodnoty $S_1$,$S_2$: $S_1$ [6 5]
$S_2$ [5 4]

Nový shluk:

výpočet atributu X:

$$\frac{6+5}{2} = 5{,}5$$

výpočet atributu Y:

$$\frac{5+4}{2} = 4{,}2$$

Vytvořený nový shluk je označen, např. B [5,5 4,2]

Vstupní matice je opět o jeden shluk kratší:

$$\begin{bmatrix} 2 & 9 \\ 3{,}5 & 7{,}5 \\ 5{,}5 & 4{,}2 \end{bmatrix}$$

Počítání vzdáleností:

$$L_1, A = \sqrt{(2-3{,}5)^2 + (9-7{,}5)^2} = 2{,}12$$
$$L_1, B = \sqrt{(2-5{,}5)^2 + (9-4{,}2)^2} = 5{,}94$$
$$A, B = \sqrt{(3{,}5-5{,}5)^2 + (7{,}5-4{,}2)^2} = 3{,}8$$

Matice vzdáleností:

$$\begin{bmatrix} 0 & 2,12 & 5,94 \\ 2,12 & 0 & 3,8 \\ 5,94 & 3,8 & 0 \end{bmatrix}$$

Je vybrána nejmenší vzdálenost $L_1$,A s hodnotou 2,12.

Hodnoty $L_1$,A: $L_1$ [2 9]
A [3,5 7,5]

Nový shluk:

výpočet atributu X:

$$\frac{2+3,5}{2} = 2,75$$

výpočet atributu Y:

$$\frac{9+7,5}{2} = 8,25$$

Vytvořený nový shluk se označí, např. C [2,75 8,25]

Vstupní matice:

$$\begin{bmatrix} 5,5 & 4,2 \\ 2,75 & 8,25 \end{bmatrix}$$

Vzdálenost:

$$L_1, C = \sqrt{(5,5-2,75)^2 + (4,2-8,25)^2} = 4,8$$

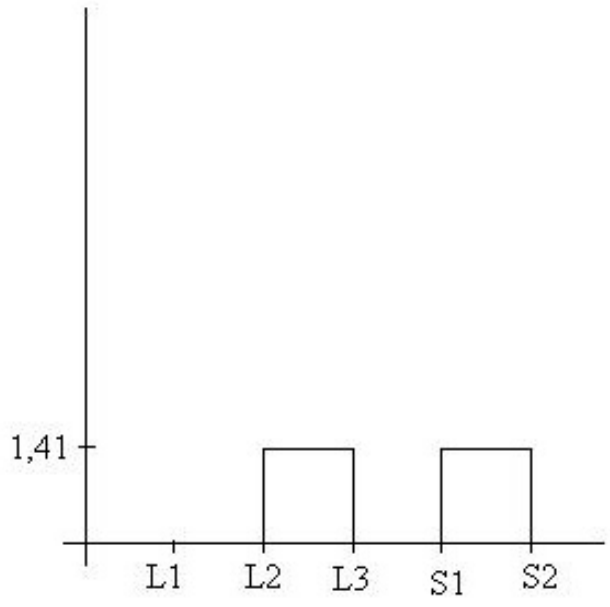
Poslední matice vzdálenosti:

$$\begin{bmatrix} 0 & 4{,}8 \\ 4{,}8 & 0 \end{bmatrix}$$

Konečným výsledkem je graf, tzv. dendrogram. Dendrogram znázorňuje graficky konečný stav shlukování a tím je celý proces srozumitelnější a názornější. Konstrukce dendrogramu je úkol poměrně náročný, a proto se princip dendrogramu uvede na předchozím přikladu.

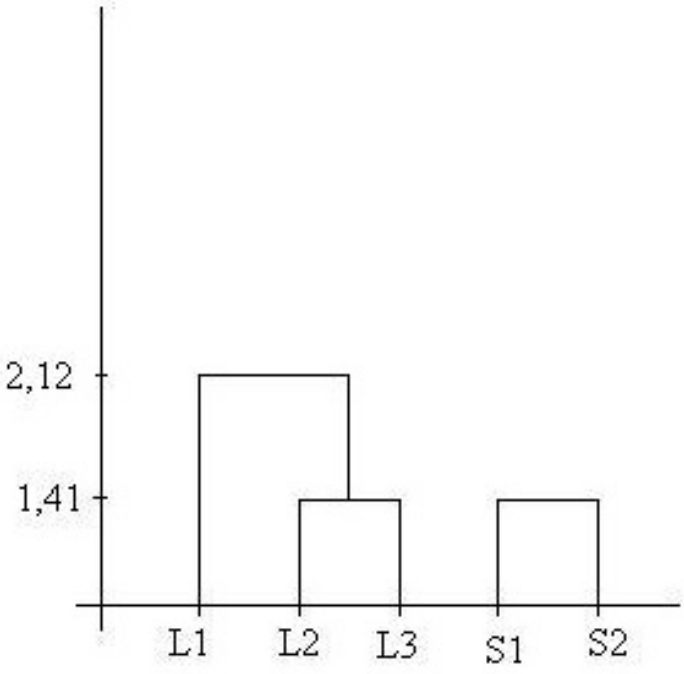
Princip dendrogramu:

Shluky $L_1$,$L_2$,$L_3$,$S_1$,$S_2$ jsou výchozí shluky. Dendrogram se zakreslí do kartézské soustavy**,** kde na osu X jsou vyneseny shluky $L_1$,$L_2$,$L_3$,$S_1$,$S_2$ a na osu Y budou + vyneseny s počítané vzdálenosti. V prvním kroku počítání vzdáleností vyšla jako nejmenší vzdálenost mezi shluky $L_2$,$L_3$ a $S_1$,$S_2$ s hodnotou 1,41 a ty se spojí dohromady.



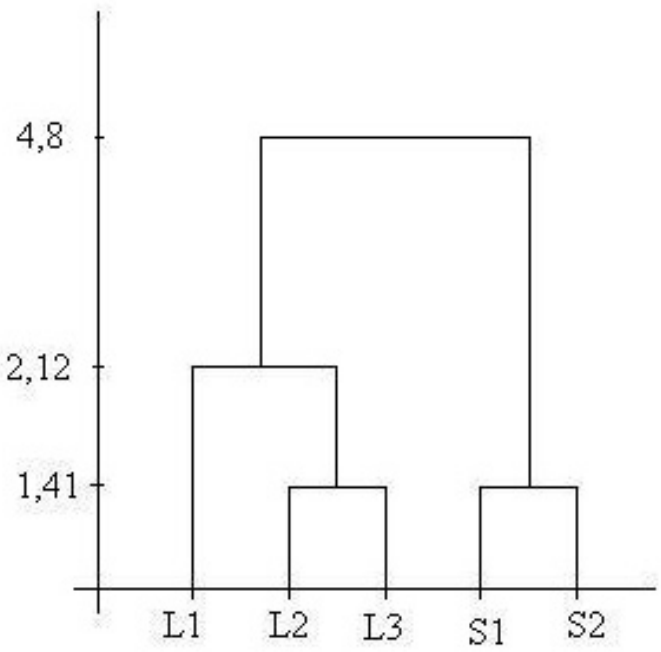

**Obr. 7.1.** Počáteční stupeň vykreslení dendrogramu

V dalším kroku počítání vzdáleností vyšla nejmenší vzdálenost mezi shluky A a $L_1$ s hodnotou 2,12. Shluk A je nový vypočítaný shluk, který se počítal se shluků $L_2$,$L_3$. Po zakreslení:




**Obr. 7.2.** Střední stupeň vykreslení dendrogramu

Poslední počítaná nejmenší vzdálenost je mezi shluky B a C s hodnotou 4,8 (je jediná a tedy nejmenší). Lze předpokládat, že shluk B je složen ze shluků $S_1$,$S_2$ a C je ze shluků A, $L_1$.




**Obr. 7.3.** Koncový stupeň vykreslení dendrogramu

Uvedený příklad, zpracovává dvě metody počítání shlukové analýzy. Z použitých metod je to již zmiňovaná metoda počítání vzdálenosti mezi shluky pomocí euklidovské metody a počítání nových shluků je realizováno aritmetickým průměrem. Jako jednu z dalších metod pro počítání shlukové analýzy je možné zvolit metodu Manhattanskou.

Pomocí Manhattanské metody se počítají vzdálenosti mezi shluky a pro počítání nových shluků je možné zvolit metodu nejbližšího souseda. Pro lepší orientaci a vysvětlení lze tyto nové dvě metody aplikovat na stejném příkladě jako doposud.

Zjednodušený postup realizace shlukové analýzy pomocí metody Manhattanské a nejbližšího souseda:

Vstupní matice je použita jako v předchozím příkladě:

$$\begin{bmatrix} 2 & 9 \\ 3 & 7 \\ 4 & 8 \\ 6 & 5 \\ 5 & 4 \end{bmatrix}$$

Vzdálenosti se počítají pomocí Manhattanské metody:

$$M(a_i, b_j) = \sum_{l=1}^{m} \left| x_{il} - x_{jl} \right| = \left| x_i - x_j \right| \tag{6.2}$$

Jednotlivě spočítané vzdálenosti:

$$\begin{aligned}
L_1, L_2 &= |(2-3)| + |(9-7)| = 3 \\
L_1, L_3 &= |(2-4)| + |(9-8)| = 3 \\
L_1, S_1 &= |(2-6)| + |(9-5)| = 8 \\
L_1, S_2 &= |(2-5)| + |(9-4)| = 8 \\
L_2, L_3 &= |(3-4)| + |(7-8)| = 2 \\
L_2, S_1 &= |(3-6)| + |(7-5)| = 5 \\
L_2, S_2 &= |(3-5)| + |(7-4)| = 5 \\
L_3, S_1 &= |(4-6)| + |(8-5)| = 5 \\
L_3, S_2 &= |(4-5)| + |(8-4)| = 5 \\
S_1, S_2 &= |(6-5)| + |(5-4)| = 2
\end{aligned}$$

Matice vzdáleností, kde na diagonále jsou 0:

$$\begin{bmatrix} 0 & 3 & 3 & 8 & 8 \\ 3 & 0 & 2 & 5 & 5 \\ 3 & 2 & 0 & 5 & 5 \\ 8 & 5 & 5 & 0 & 2 \\ 8 & 5 & 5 & 2 & 0 \end{bmatrix}$$

Poté jsou opět vybrány dva shluky nejmenší vzdálenosti, např. $L_2$,$L_3$ a počítá se nový shluk pomocí nejbližšího souseda.

Nové shluky je možné realizovat metodou nejbližšího souseda:

$$D_{g\langle hh'\rangle} = \frac{1}{2}(D_{gh} + D_{gh'} - |D_{gh} - D_{gh'}|) \tag{6.3}$$

Například je možné provést změnu v prvním kroku, kde se vybere nejmenší vzdálenost, tj. $L_2$,$L_3$

Hodnoty $L_2$,$L_3$: $L_2$ [3 7]
$L_3$ [4 8]

Nový shluk:

výpočet atributu X:

$$D = \frac{1}{2}(3 + 4 - |3 - 4|) = \frac{1}{2}(7 - 1) = 3$$

výpočet atributu Y:

$$D = \frac{1}{2}(7 + 8 - |7 - 8|) = \frac{1}{2}(15 - 1) = 7$$

Z těchto výpočtů se vytvoří nový shluk, např. A [3  7] a poté se vše opakuje, dokud není znám konečný výsledek. Výsledkem je opět dendrogram a princip jeho stavby je stejný.

Při hledání nových shluků je možné postupovat tak, že se např. vyberou nejmenší vzdálenosti jako ve výše zmíněném případě, ale lze také naopak hledat největší vzdálenosti. Rozdíl mezi jednotlivými metodami je jen velmi malý a principiálně jsou téměř stejné. Oba zmíněné postupy počítaní shlukové analýzy patří do aglomerativních algoritmů, kde každý samostatný objekt je shlukem a celé se řadí do hierarchického shlukování.

## 7.1 UKÁZKA FUNKCE PROGRAMU

Nejprve se zadá počet prvků v matici a poté se napíší jednotlivé atributy x, y.

Naše vstupní matice

```
Zadej pocet prvku matice:5

Zadej radek: 1 ve formatu x,y
2,9
Zadej radek: 2 ve formatu x,y
3,7
Zadej radek: 3 ve formatu x,y
4,8
Zadej radek: 4 ve formatu x,y
6,5
Zadej radek: 5 ve formatu x,y
5,4
----------------------------
Zadana matice 1 [ 2 , 9]
Zadana matice 2 [ 3 , 7]
Zadana matice 3 [ 4 , 8]
Zadana matice 4 [ 6 , 5]
Zadana matice 5 [ 5 , 4]

----------------------------
```

**Obr. 7.4.** Ukázka programu – zadaná vstupní matice.

Realizace výpočtu vzdáleností.

```
Vypocet Euklidovske vzdalenosti
[1,1]=0
[1,2]=2,24
[1,3]=2,24
[1,4]=5,66
[1,5]=5,83
[2,1]=2,24
[2,2]=0
[2,3]=1,41
[2,4]=3,61
[2,5]=3,61
[3,1]=2,24
[3,2]=1,41
[3,3]=0
[3,4]=3,61
[3,5]=4,12
[4,1]=5,66
[4,2]=3,61
[4,3]=3,61
[4,4]=0
[4,5]=1,41
[5,1]=5,83
[5,2]=3,61
[5,3]=4,12
[5,4]=1,41
[5,5]=0
```

**Obr. 7.5.** Ukázka programu – výpočet euklidovské vzdálenosti

Výpis matice vzdáleností.

```
Matice Euklidovske vzdalenosti
¦ 0 ; 2,24 ; 2,24 ; 5,66 ; 5,83¦
¦ 2,24 ; 0 ; 1,41 ; 3,61 ; 3,61¦
¦ 2,24 ; 1,41 ; 0 ; 3,61 ; 4,12¦
¦ 5,66 ; 3,61 ; 3,61 ; 0 ; 1,41¦
¦ 5,83 ; 3,61 ; 4,12 ; 1,41 ; 0¦
```

**Obr. 7.6.** Ukázka programu – matice vzdálenosti

Zde je vybrán nejmenší shluk a je vypsána nová vstupní matice.

```
Nejmensi shluk
¦ 6 ; 5 ¦
¦ 5 ; 4 ¦

----------------------------
Nova matice 1 [ 2 , 9]
Nova matice 2 [ 3,5 , 7,5]
Nova matice 3 [ 5,5 , 4,5]
```

**Obr. 7.7.** Ukázka programu – nejmenší shluk, nová vstupní matice

Nový výpočet vzdáleností s vypsanou maticí vzdáleností.

```
Vypocet Euklidovske vzdalenosti
[1,1]=0
[1,2]=2,12
[1,3]=5,7
[2,1]=2,12
[2,2]=0
[2,3]=3,61
[3,1]=5,7
[3,2]=3,61
[3,3]=0

------------------------------
Matice Euklidovske vzdalenosti
¦ 0 ; 2,12 ; 5,7¦
¦ 2,12 ; 0 ; 3,61¦
¦ 5,7 ; 3,61 ; 0¦
```

**Obr. 7.8.** Ukázka programu – výpočet vzdálenosti, matice vzdálenosti

Opět vypsaná matice vzdáleností s dalším výpočtem Euklidovské vzdálenosti s konečnou maticí vzdáleností.

```
Nova matice 1 [ 2,75 , 8,25]
Nova matice 2 [ 5,5 , 4,5]

------------------------------
Vypocet Euklidovske vzdalenosti
[1,1]=0
[1,2]=4,65
[2,1]=4,65
[2,2]=0

------------------------------
Matice Euklidovske vzdalenosti
¦ 0 ; 4,65¦
¦ 4,65 ; 0¦
```

**Obr. 7.9.** Ukázka programu – matice vzdáleností, výpočet vzdáleností, matice vzdáleností.

Dále ještě program určí nejmenší shluk (tedy jediný) a vytvoří konečnou vstupní matici, se kterou už není možné už dále pracovat.

```
Nejmensi shluk
¦ 5,5 ; 4,5 ¦
¦ 2,75 ; 8,25 ¦

------------------------------
Nova matice 1 [ 4,125 , 6,375]
------------------------------
```

**Obr. 7.10.** Ukázka programu – nejmenší shluk, nová konečná vstupní matice

Realizace Manhatanské metody a Metody nejbližšího souseda. Ukázka spočítané vzdálenosti.

```
Vypocet Manhattanská vzdálenost
[1,1]=0
[1,2]=3
[1,3]=3
[1,4]=8
[1,5]=8
[2,1]=3
[2,2]=0
[2,3]=2
[2,4]=5
[2,5]=5
[3,1]=3
[3,2]=2
[3,3]=0
[3,4]=5
[3,5]=5
[4,1]=8
[4,2]=5
[4,3]=5
[4,4]=0
[4,5]=2
[5,1]=8
[5,2]=5
[5,3]=5
[5,4]=2
[5,5]=0
```

**Obr. 7.11.** Ukázka programu – výpočet manhatenské vzdálenosti

Výpis matice vzdáleností s nejmenším shlukem a určením nejbližšího souseda a vypsaná vstupní matice.

```
Matice Manhattan vzdalenosti
¦ 0 ; 3 ; 3 ; 8 ; 8¦
¦ 3 ; 0 ; 2 ; 5 ; 5¦
¦ 3 ; 2 ; 0 ; 5 ; 5¦
¦ 8 ; 5 ; 5 ; 0 ; 2¦
¦ 8 ; 5 ; 5 ; 2 ; 0¦

------------------------------
Nejmensi shluk
¦ 3 ; 7 ¦
¦ 4 ; 8 ¦

------------------------------
Nejblizsi soused: A[3,7]
------------------------------
Nova matice 1 [ 2 , 9]
Nova matice 2 [ 3 , 7]
Nova matice 3 [ 6 , 5]
Nova matice 4 [ 5 , 4]
```

**Obr. 7.12.** Ukázka programu – manhatenské matice vzdáleností, nejmenší shluk s nejbližším sousedem, nová matice vzdáleností

Proces dále pokračuje jako v předchozím případě, tedy: výpočet vzdáleností, matice vzdáleností, výběr nejmenšího shluku, nová vstupní matice.

```
Vypocet Manhattanská vzdálenost
[1,1]=0
[1,2]=3
[1,3]=8
[1,4]=8
[2,1]=3
[2,2]=0
[2,3]=5
[2,4]=5
[3,1]=8
[3,2]=5
[3,3]=0
[3,4]=2
[4,1]=8
[4,2]=5
[4,3]=2
[4,4]=0

------------------------------
Matice Manhattan vzdalenosti
¦ 0 ; 3 ; 8 ; 8¦
¦ 3 ; 0 ; 5 ; 5¦
¦ 8 ; 5 ; 0 ; 2¦
¦ 8 ; 5 ; 2 ; 0¦

------------------------------
Nejmensi shluk
¦ 6 ; 5 ¦
¦ 5 ; 4 ¦

------------------------------
Nejblizsi soused: A[5,4]
------------------------------
Nova matice 1 [ 2 , 9]
Nova matice 2 [ 3 , 7]
Nova matice 3 [ 5 , 4]
```

**Obr. 7.13.** Ukázka programu – realizace vzdálenosti, nová vstupní matice

Program dále počítá dle zavedeného principu a na konec vypíše nejmenší shluk a vytvoří konečnou vstupní matici, se kterou už není možné už dále pracovat.

```
Vypocet Manhattanská vzdálenost
[1,1]=0
[1,2]=3
[1,3]=8
[2,1]=3
[2,2]=0
[2,3]=5
[3,1]=8
[3,2]=5
[3,3]=0

------------------------------
Matice Manhattan vzdalenosti
¦ 0 ; 3 ; 8¦
¦ 3 ; 0 ; 5¦
¦ 8 ; 5 ; 0¦

------------------------------
Nejmensi shluk
¦ 3 ; 7 ¦
¦ 2 ; 9 ¦

------------------------------
Nejblizsi soused: A[2,7]
------------------------------
Nova matice 1 [ 2 , 7]
Nova matice 2 [ 5 , 4]

------------------------------
Vypocet Manhattanská vzdálenost
[1,1]=0
[1,2]=6
[2,1]=6
[2,2]=0

------------------------------
Matice Manhattan vzdalenosti
¦ 0 ; 6¦
¦ 6 ; 0¦

------------------------------
Nejmensi shluk
¦ 5 ; 4 ¦
¦ 2 ; 7 ¦

------------------------------
Nejblizsi soused: A[2,4]
------------------------------
Nova matice 1 [ 2 , 4]
```

**Obr. 7.14.** Ukázka programu – dokončení všech výpočtu týkající se programu

# 8. ZÁVĚR

Práci můžeme rozdělit na dvě podstatné části. V první části se zaměřuje na kapitoly zabývající se objasněním jednotlivých výrazů. Na základě dostupné literatury je provedena rešerše pojmů. Nejpodstatnější část práce je zaměřena na shlukovou analýzu, která je základním pilířem metody učení bez učitele, a proto se zabývá jednotlivými částmi podrobněji.

Druhá část práce je směřována na praktičtější využití shlukové analýzy. Byl vytvořen program, který uživateli umožňuje si vyzkoušet shlukovou analýzu v praxi. Program porovnává samostatné shluky a pomocí vzdáleností určuje vzájemnou podobnost jednotlivých shluků. Daný program lze použít pro demonstraci porovnávání více subjektů a získané informace už nám pomohou posloužit k dalšímu rozhodování. Za určitých podmínek by bylo možné ho uvést do reálné praxe. Dále jsou hodnoceny a porovnávány analýzy mezi sebou. Cílem práce bylo vytvořit celkový pohled na problematiku spojenou s učením bez učitele.

# 9. LITERATURA

[1] Lukasová, A., Šarmanová, J.: *Metody shlukové analýzy*. Praha 1985.

[2] Řezáková, H., Húsek, H., Snášel, V.: *Shluková analýza dat*. Praha 1990

[3] Honzík, P.: *Poznámky z přednášek*. Brno 2007.

[4] Hlaváč, V.: *Učení bez učitele*. Praha 2001.

[5] Horák, J.: *Shluková analýza – zdroj internet*. Praha 2002.

[6] Rimarčík, M.: *Faktorová analýza – zdroj internet*. Košice 2008.

[7] Kelbel, J., Šilhán, D.: *Shluková analýza.* Praha 2007