1. DOLOVÁNÍ ZNALOSTÍ Z DAT

Čas ke studiu: 2 hodiny

Cíl Po prostudování této kapitoly budete umět

- definovat a vysvětlit pojem dolování znalostí z dat
- popsat a zdůvodnit etapy procesu získávání znalostí a etapy dolování
- popsat využití výsledků dolování dat z různých oblastí reality

Výklad

1.1. Data, informace, znalost

Realita jako prostor znaků

Svět, realita je pojem příliš obecný a široký. Reálnou skutečnost poznáváme po malých částech, často se navzájem překrývajících podle toho, proč nás právě tato část světa zajímá. Předměty našeho dílčího zájmu nazýváme **objekty**. Mohou to být lidé, zvířata, věci, jevy, procesy, vztahy či závislosti mezi nimi. Objekty nejčastěji poznáváme a popisujeme pomocí jejich vlastností - **atributů**, znaků, příznaků, údajů, v matematických disciplínách je pak nazýváme proměnnými. *(Chceme-li popsat člověka, o kterém budeme mluvit, řekneme například: ten malý blonďák v červené bundě ze 2. patra).* Výběr atributů pro dobrý popis objektů je obvykle problémem, který budeme diskutovat níže. Ideální by bylo, kdybychom v případě potřeby dalších údajů o pozorovaných objektech je mohli jednoduše získat. To však není obvykle reálné.

Ve smyslu výše uvedených úvah formulujme následující předpoklad:

> **Každá množina reálných objektů a jevů má své zákonitosti, své zařazení do hierarchie světa, svou klasifikaci na podtypy, své vztahy k okolí. Také podmnožiny atributů mohou mít mezi sebou důležité vztahy – asociace: korelace, příčiny a následky, skryté faktory apod**.

Poznávání všech těchto skutečností je naším cílem.

Poznávání reality

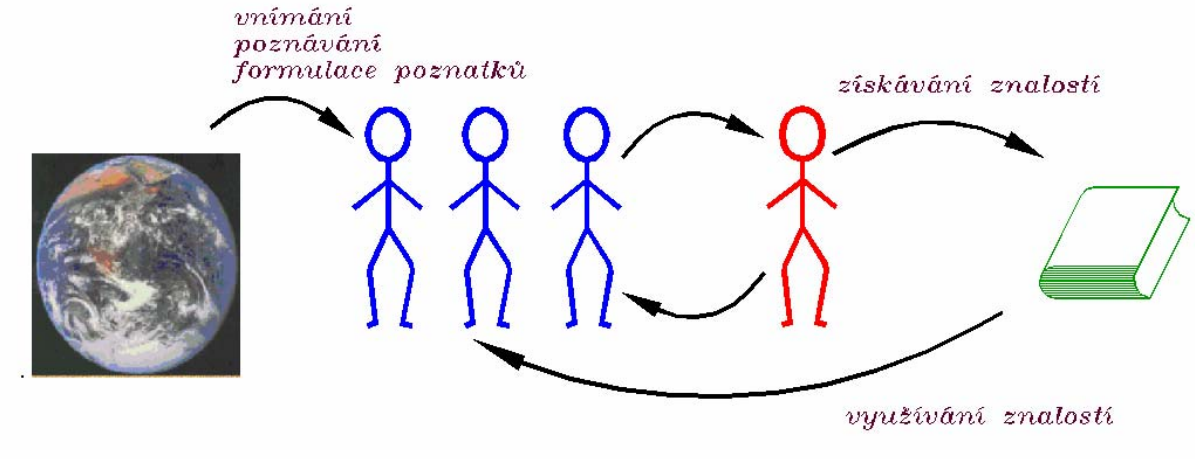
Dávno předtím, než dolování v datech dostalo své jméno, lidé získávali své znalosti pozorováním okolního světa a zobecňováním jednotlivých poznatků: rozdělováním pozorovaných objektů do podobných skupin, formulováním pravidel o tom, že má-li objekt vlastnosti A, má většinou i vlastnosti X; objeví-li se nový objekt s vlastnostmi B, patří nejspíš ke skupině C atd.

Pomineme dobu, kdy hlavním paměťovým médiem byla lidská paměť a získané znalosti byly předávány ústně (některé jsou dodnes, od přísloví a pranostik až po "nesahej na žehličku, pálí"). Jakmile začnou hloubavější z nás zkoumat některé jevy systematicky, často začínají shromažďováním údajů a zkoumáním toho, jaká fakta o údajích platí. Ověřují, jestli se z faktů dají formulovat obecně platná pravidla, nebo dokonce dokázat některé (přírodní, společenské, …) zákonitosti.

Ve výzkumné praxi při zkoumání objektů zatím příliš málo známých či příliš složitých, u nichž dosud neumíme popsat jejich vlastnosti a chování, často také vycházíme z jejich pozorování. Sbíráme nebo

měříme o nich údaje, které pozorovat můžeme, provádíme experimenty s různě nastavenými podmínkami apod. Tak o nich nashromáždíme množství údajů a z nich se pokoušíme vydedukovat jejich další, obecnější nebo skrytější vlastnosti.

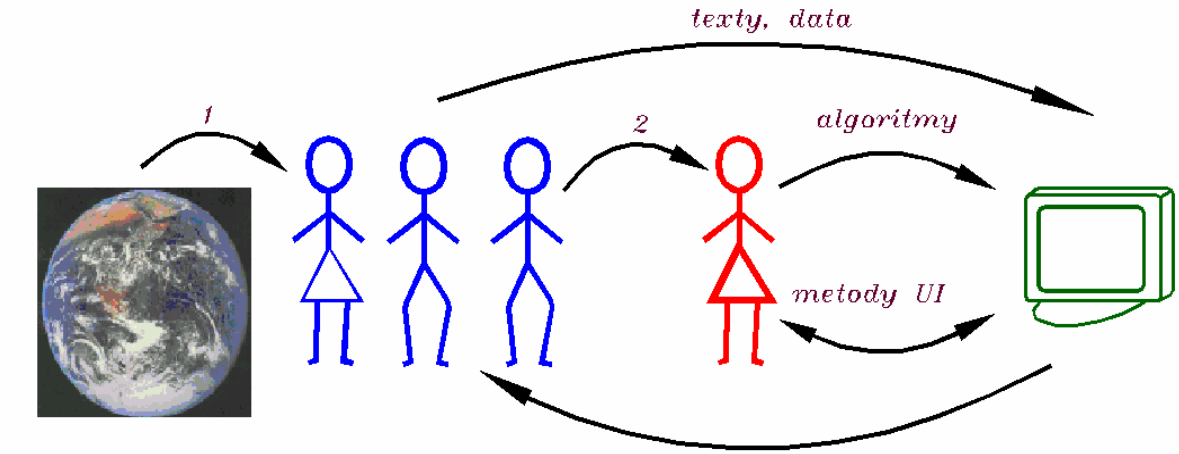
Tento postup prováděli lidé odpradávna. Zprvu jim za paměťové médium sloužila jen vlastní hlava, později papír, ještě později paměť počítačů. Pro vyhodnocování takových údajů a odvozování nových sloužil zprvu jen selský rozum, později hlavně matematické disciplíny - především matematická logika a statistika a konečně mnohé metody explorační analýzy dat, umělé inteligence, případně neuronových sítí. Metody analýzy dat vznikaly již dávno v době předpočítačové pro výzkum v nejrůznějších oborech (psychologie, biologie, technika atd.), doba počítačů jim však teprve dala velké možnosti využití.



Obrázek 1.1. Cyklus klasického poznávání světa

Sbírání údajů a jejich vyhodnocení používá nejen vědecký výzkum, ale i formálně jednodušší úlohy, například různé sociologické průzkumy pro účely politické, reklamní, marketingové, dopravní, psychologické atd.

Víme, že k ověřování formulovaných hypotetických pravidel již dlouho slouží matematická statistika. Není ale obecně známo, že i pro etapu zobecňování jednotlivých faktů – pro etapu formulování hypotéz nad nimi, existuje již mnoho desetiletí řada metod.



Obrázek 1.2. Automatizované poznávání reality

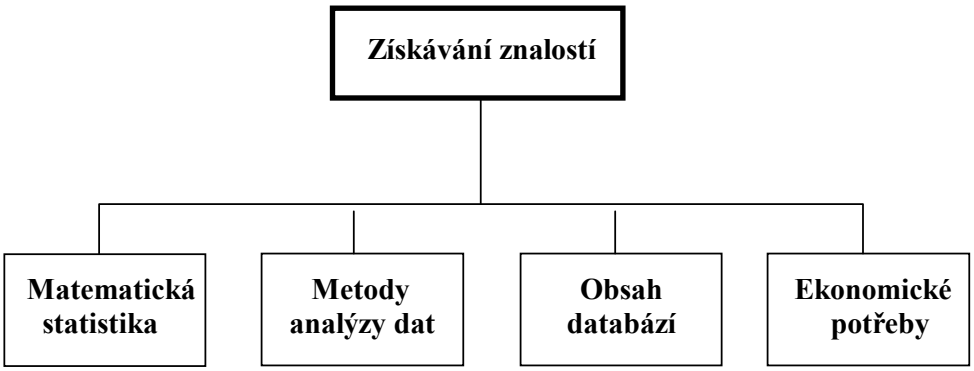
Nezávisle na tomto zkoumání člověk ve shodě s potřebou udržovat pořádek v evidencích o kdečem – o skupinách lidí, věcí, jevů apod., zakládal databáze údajů. Když údaje zestárly, zrušil je nebo archivoval a ukládal do evidence údaje nové. Největší rozsah údajů pravděpodobně postupně nashromáždily firmy o svých výrobách, obchodech, … Dobrému fungování firem pomáhají znalosti z ekonomie. Ovšem nejen firmy mají v databázích mnoho evidovaných dat, také nemocnice, školy, zemědělci, technici, sociologové, výzkumníci nejrůznějších oborů.

❑ Získávání znalostí jako multioborová disciplina

Teprve počátkem 90. let 20. století přišel nápad využít především údajů z počítačových databází, byť určených původně jen k evidenčním účelům, také jako zdroj automatizovaného získávání znalostí. Jsou k tomu využívány staré známé metody i metody nově vznikající.

Hlavním impulsem pro rozvoj nového oboru byl zájem firem zpracovávat svá data za účelem získání lepších informací o fungování firmy a umět tak lépe a rychleji reagovat na potřeby trhu, být konkurenceschopnější. (*Zvažme příklad obchodní firmy, evidující některé osobní údaje o svých zákaznících a jejich nákupech – době nákupu, seznamu zboží, způsobu platby. Analýzou takových údajů získají například skupiny podobných „typických" zákazníků a mohou jim efektivněji na míru nabízet své reklamní či věrnostní akce, čímž ušetří náklady na masovou reklamu a v zákaznících posílí dojem, že jsou tou správnou firmou pro jejich potřeby).*

Na rozdíl od ostatních majitelů dat velké firmy mohly rozvoj metod a jimi získané informace zaplatit. Tím výrazně přispěly ke vzniku nového integrovaného oboru, jeho pojmenováním byl proces vzniku završen.



Obrázek 1.3. Zdrojové obory pro disciplínu získávání znalostí z dat

Již jsme si uvedli, že mnohé úlohy a metody zařazované k dolování dat vznikly před řadou desetiletí. Byly nazývány **metodami analýzy mnohorozměrných dat**. Konkrétně se o historii každé metody zmíníme u jednotlivých úloh. V současnosti se pro celou disciplínu, patřící do počítačových věd, používá několik názvů. Z angličtiny převzatý **Knowledge Discovery in Databases** (KDD) nebo u nás i v originále používaný **Data Mining** (DM), v překladu pak v několika provedeních jako **získávání znalostí z dat** (jakýchkoliv, nejen databázových) či **získávání znalostí z databází,** také **dolování dat** nebo některé podobné názvy. Pokud se metody DM používají k získání užitečných znalostí z firemních dat, schopných podpořit rozvoj firmy, používá se často názvu **Business Intelligence** (BI).

Níže si uvedeme, že vlastní metody získávání znalostí z dat jsou součástí většího procesu, zahrnujícího výběr a přípravu dat, vlastní dolování, vizualizaci a interpretaci výsledků.

Statistika a metody získávání znalostí z dat

Mějme tedy následující obecnou úlohu: předmětem našeho zájmu jsou reálné objekty či jevy, k jejich zkoumání máme k dispozici jejich množinu (zřídka úplnou, obvykle více nebo méně rozsáhlý vzorek objektů, často jen jejich náhodný výběr). Každý objekt je popsán vektorem svých atributů, které získáme zaznamenáním, pozorováním, měřením, experimentováním apod. Úkolem je zjistit o objektech, jejich množině a jejich atributech vše, co se na základě dat zjistit dá.

Pro metody analyzující data budeme dále používat zkratku DM (Data-miningové metody). Protože je možno zkoumat data z různých hledisek, existuje i celá řada typů metod analýzy dat. Dávají výsledky různých typů, společný mají zdroj informací. DM jako disciplína je někdy zařazována do mnohorozměrné statistiky, jindy mezi metody umělé inteligence. Obě zařazení mají svá opodstatnění, protože některé konkrétní metody patří více tam, jiné onam.

Někdy se v této souvislosti mluví o rozdělení metod patřících k matematické statistice (též někdy ke konfirmační analýze dat) nebo k explorační analýze dat (DM).

Konfirmační analýza

- zahrnuje metody klasické matematické statistiky,
- používá metody, odpovídající na analytikovy otázky typu "je pravda, že ... ", tj. testují zadané hypotézy nad daty,
- nahrazují neznámé matematické vztahy mezi veličinami zadaným typem funkce, testují míru takové závislosti apod.,
- analytik formuluje požadavek, statistika mu dá na tento požadavek odpověď; na nevyslovené otázky neodpoví,
- v procesu poznávání světa patří na začátek etapy formulování nových znalostí.

Explorační analýza (explore = prozkoumat, probádat)

- provádějí tzv. orientační studie nad daty,
- samy si automaticky generují otázky zadaného typu a ihned testují výsledky,
- dávají odpovědi s menší vahou, ale odpovídají i na otázky nevyslovené typu "co je v datech zajímavé, neprůměrné, ... ",
- objevují i nové poznatky,
- jako výsledek formulují hypotézy podporované zkoumanými daty a předkládají je k dalšímu bádání, ověřování, testování,
- patří spíš k metodám umělé inteligence, v procesu poznávání světa patří do etapy poznávání, zobecňování.

Informace získané z databází

Databází tedy budeme chápat nejen firemní databáze, případně nad nimi vytvořené datové sklady, shromažďující a integrující i data archivní a další, ale jakákoliv data uspořádaná do (relačních) tabulek, kde v řádcích tabulky jsou údaje o jednom z množiny sledovaných objektů, ve sloupcích jsou atributy těchto objektů. Pro jednoduchost budeme předpokládat jedinou, někdy dokonce nenormalizovanou tabulku.

> Je dána tabulka **X** popisující množinu objektů $\mathbf{O} = \{O_1, O_2, \ldots, O_n\}$ zadaných pomocí atributů $\mathbf{A} = \{A_1, A_2, \ldots, A_m\}$, každý s doménou z $\mathbf{D} = \{D_1, \ldots D_m\}$.

Z těchto dat je možno získat užitečné informace prostředky klasických informačních systémů. Jsou to různé komentované výběry dat, agregované hodnoty – sumy, průměry atd., jejich řady – časové, prostorové atd., jejich hierarchie. Všechny vznikají na základě požadavků uživatele, jejich nalezení je součástí IS nebo existuje jiná možnost získat, bývají doplněny vhodným formátováním, grafickým zobrazením. I vyšší úroveň databází, datové sklady, nabízejí uživatelům předzpracovaná data.

Mimo to však mohou být v datech rozptýleny další, dosud nepoznané skutečnosti, z nichž některé mohou být užitečné člověku a doplnit jeho dosavadní znalost světa.

Příklad 1.1. *Je dána tabulka dat za posledních 10 let o studentech gymnázia se strukturou:*

Student (skol_rok, trida, rod_cis, jmeno, vek, vyska, vaha, znam_CJ, …, skok_vys, …)

Tabulka obsahuje redundance, například opakující se (rod_cis, jmeno).

Klasickými SQL dotazy získáme informace o jednotlivých studentech a jejich atributech, jako například jakou známku má Jiří Novák z matematiky, jaký je průměr známek z matematiky v jednotlivých třídách apod.

Metodami matematické statistiky dostaneme odpověď i na další otázky, jako například „existuje korelace mezi známkami z matematiky a hudební nauky?“ a mnohé další.

Z datového skladu nad touto tabulkou můžeme zobrazit například časový průběh průměrných známek z matematiky v jednotlivých ročnících nebo podle jednotlivých učitelů atd.

Metodami získávání znalostí z dat pak můžeme objevit i další, v datech rozptýlené a na první pohled neviditelné zákonitosti, jako například „má-li student známku = 1 z občanské nauky, pak skok_vys > 110 cm se spolehlivostí 70%”. Na takovou souvislost se pravděpodobně nikdo nebude ptát a přesto se v datech může objevit,

♦

❑ Typy úloh dolování dat

Zopakujme si, že realitu poznáváme prostřednictvím reálných objektů a jejich atributů. Odhalování a poznávání vztahů mezi nimi je naším cílem.

Úlohy DM můžeme rozdělit z několika hledisek:

- ❑ Podle období platnosti výsledků na úlohy
  - **deskriptivní,** popisující obecné vlastnosti zadaných dat (*například jejich rozdělení do skupin podobných objektů, platné vztahy mezi jejich atributy apod.*)
  - **prediktivní**, umožňující na základě analýzy zadaných dat předpovídat chování – vlastnosti dat budoucích (*například nalezením pravidel, kterými se řídila skupina zákazníků dosud předpovědět jejich chování v následujícím období a přizpůsobit tak nabídku).*
- ❑ Jinou klasifikací můžeme rozdělit metody DM na
  - **charakterizace dat**, popisující obecné vlastnosti analyzovaných množin objektů nebo atributů; *(například najdeme silný vztah příčina - následek mezi atributy dlouholetý kuřák a onemocnění jednou z nemocí rakovina plic nebo infarkt),*
  - **diskriminace dat**, vyhledávající pro některé podmnožiny objektů nebo atributů, v čem se liší od ostatních dat, v čem jsou výjimečné; budeme je též nazývat analýzou výjimek *(například hledáme, co způsobuje sice velmi řídký, ale nepříjemný výskyt kazů materiálu při výrobě).*
- ❑ Podle toho, který typ vztahů zkoumají, rozdělujeme úlohy DM na 3 nejčastěji používané typy úloh:
  - **asociační**, hledající vztahy mezi podmnožinami atributů,
  - **shlukovací**, hledající vztahy mezi objekty, rozdělení objektů do podobných podmnožin a případně hierarchii těchto podmnožin,
  - **klasifikační**, hledající pravidla, podle nichž se data rozdělují do definovaných tříd.

Na základě všech výše uvedených informací budeme rozumět následující, nejčastěji uváděné definici.

Definice 1.

Dolováním znalostí z dat nazýváme proces netriviálního získávání implicitní, dříve neznámé a potencionálně užitečné informace z dat.

Netriviálním procesem rozumíme to, že informaci nelze získat jednoduše, například vhodným SQL dotazem, ale je nutné použít vhodnou speciální metodu. *Dříve neznámá* informace je informace v datech skrytá, rozptýlená, neviditelná na první pohled. *Potenciálně užitečná* informace je ta, která má význam jako nová vědecká hypotéza, jako podklad pro manažerské rozhodnutí, jako upozornění na neobvyklou skutečnost apod.

❑ Rozdíl mezi OLAP a dolováním znalostí

Někdy se zaměňuje nebo nedostatečně rozlišuje OLAP a metody dolování znalostí. Obě skupiny metod provádějící výpočty nad daty mají za úkol vyhledávat pro uživatele nové užitečné informace.

OLAP je soubor metod, které provádějí výpočty uživatelem požadované a prezentují výsledky ve srozumitelné a publikovatelné podobě (tabulky, grafy, slovní formulace apod.). Pracují téměř výhradně nad datovým skladem. Jejich podstatou jsou převážně agregované hodnoty různých hierarchických stupňů. Z nich se vytváří časové řady, řady podél jiných dimenzí nebo podél více dimenzí současně. Tytéž výsledky je možno prezentovat na různých hierarchických úrovních a různými metodami. Úkolem je maximálně zpřístupnit pochopení výsledků uživateli a maximálně zjednodušit ovládání programu.

K dolování znalostí patří mnoho metod, které vyhledávají v datech nové, dosud neznámé a také nedotazované znalosti různých typů. Pracují (podle implementace) nad různými formáty dat, do DS jsou zařazovány spíše jako nadstavba, produkující něco navíc proti metodám OLAP.

Příklad 1.2.

Mějme část datového skladu banky s daty za 10 let evidence o poskytování a splácení úvěrů. Nástroji OLAP se může uživatel – marketingový ředitel – dotazovat na

- *dlouhodobý časový vývoj poskytovaných úvěrů a splátek celkem, podle zákazníků apod.,*
- *rozložení zákazníků a jejich splátek podle regionů, podle velikosti firem apod.*

Pokud má k dispozici i příslušné metody dolování, může například pomocí nich zjistit, že

- *zákazníci s úvěrem nad 50 mil. a s ručením některé státní organizace nesplácejí,*
- *že nesplácejí ještě některé další specifické podmnožiny zákazníků*

♦

1.2. Využití dolování znalostí

❑ Potřeby uživatelů metod dolování

Nejčastěji uváděná definice dolování znalostí z dat jako “proces netriviálního získávání implicitní, dříve neznámé a potencionálně užitečné informace z dat” je často doplňována sloganem “za účelem získání obchodní výhody”.

V našem širším pohledu na tento proces rozdělíme potřeby dolování z hlediska uživatelů na

- průzkum - marketing, bankovnictví, výroba, pojišťovnictví, … (získání obchodních výhod)
- výzkum – medicína, biologie, hutnictví, … (získání nových odborných znalostí, hypotéz)
- sociologický průzkum – veřejné mínění, sčítání lidu, lokální věcné problémy, … (získání politických výhod)

Každá z těchto skupin má jiné nároky a potřeby na data, proces jejich dolování, míru spolehlivosti výsledků, prezentaci apod. Můžeme si je představit následovně:

	Marketing	**Výzkum**	**Sociologie**
DATA			
Zdroj	Databáze	Sběr + databáze	Sběr
Rozsah	Velký	Menší	Malý
Přírůstky	Časté	Řídké – žádné	Žádné
Struktura	Stálá	Různá	Různá
ZPRACOVÁNÍ			
Předzpracování	Automatické	Na míru	Žádné
Analýza potřeb	Jednorázová	Na míru	Standardní
Rychlost	Vysoká-on line	Menší	Menší
Úplnost, kvalita	Informativní	Vysoká	Informativní
Výstupy	Graf, tabulka, text	Pracovní	Graf, tabulka, text
Výsledky	Aktuální	Dlouhodobé	Aktuální
UŽIVATEL			
Primární	Manažer	Výzkumník	Sociolog
Sekundární		Odborná veřejnost	Veřejnost

To vše znamená, že datové sklady firemní a pro byznys – inteligenci na jedné straně a datové sklady pro využití v ostatních oblastech se budují rozdílně. Přitom současná literatura se zabývá téměř výhradně první kategorií, i když příklady užití (obvykle pro lepší pochopení čtenářů) čerpá odjinud.

1.3. Životní cyklus procesu získávání znalostí z dat

❑ Etapy procesu získávání znalostí z dat

Dolováním znalostí nazýváme vlastní proces hledání konkrétních typů znalostí pomocí konkrétních metod. Ovšem to je jen částí rozsáhlejšího procesu, do něhož se spojují v integrovaný celek i další části.

Celý proces je nazýván procesem vyhledávání znalostí v databázích (KDD). Tvoří jakýsi životní cyklus komplexní analýzy dat, obsahující postupně etapy formulace řešeného problému a problémové analýzy, datové analýzy - výběru (nebo sběru) relevantních dat, jejich předzpracování - integrace do jednotného formátu, transformace a odvozování dat, dále vlastní dolování nových hypotéz, interpretace výsledků a konečně využití a zhodnocení celého procesu.

Obdobně, jako u každé složité činnosti, jsou pro projekty dolování znalostí definována mnohá teoretická pravidla: etapy projektu, řešitelské týmy, metody, algoritmy, SW nástroje, způsoby prezentace a interpretace výsledků, doporučení dalších postupů. Prakticky všechny se shodují v následujících etapách a následujících typech spoluřešitelů:

1. **Formulace problému** znamená věcné zadání výzkumné úlohy. Specifikuje podstatu úseku reality, která bude zkoumána a účel, pro který bude zkoumána. Provádí ji odborník na problémovou oblast (dále jen **expert**, řešitel). Vychází z potřeby přeměnit data na užitečnou informaci a znalost.

2. **Věcná analýza** úlohy se dělí na **datovou** analýzu (výběr objektů a jejich atributů relevantních vzhledem k zadanému účelu zkoumání, způsob kódování atributů nečíselných ap.) a analýzu **problémovou** (formulace problémů, formalizovaných otázek nebo jejich typů). Provádí ji odborník znalý dat (dále jen expert nebo též **majitel dat**), vhodná je konzultace s analytikem - odborníkem na metody analýzy dat (dále jen **analytik**). Výsledkem je seznam atributů každého sledovaného objektu i s jejich obory hodnot, často ve formě formuláře, dotazníku, tabulky ap., do nichž se údaje budou zaznamenávat. Dalším výsledkem je návrh metod, které budou aplikovány.

3. **Sběr údajů** je etapa z formálního hlediska zřejmá. Prakticky může znamenat třeba jen jednoduché sesbírání údajů (např. sociologický výzkum na základě vyplněných dotazníků), nebo provádění náročných experimentů (např. technických, technologických, biologických), nebo dlouhodobé sbírání údajů z praxe (např. v medicíně, meteorologii), výběr relevantních údajů z databáze (nejčastěji dat komerčních, ale i všech jiných typů) atd. Existují techniky a kritéria pro plánování experimentu, metody pro sběr dat, výběr reprezentativních vzorků ap., které tuto etapu podporují. Nasbírané údaje se zaznamenají pro další automatizované zpracování do počítače.

4. **Hrubá filtrace** dat je etapa předcházející vlastním výpočtům nad daty a nutná pro správnost výsledků. Jde o vyhledání dat chybných, chybějících, irelevantních, redundandních apod. Chyby mohou vzniknout na několika místech sběru: přímo při experimentu, při záznamu o jeho výsledku, při přepisu do počítače. Chybějící nenaměřené údaje některé metody následujícího zpracování nepřipouštějí, proto je třeba řešit situaci např. jejich vypuštěním, doplněním apod. Konečně pokud data nejsou navržena či kódována na míru problému, např. pochází z jiného zdroje, nebo věcná analýza nebyla provedena důsledně, mohou obsahovat data irelevantní, s konstantními hodnotami, s nadbytečnou informací, nevhodně zakódovaná apod. Výsledkem etapy filtrování dat mají být data bezchybná, relevantní, zakódovaná v souladu s následným zpracováním a s jednotným zakódováním údajů chybějících, prostě data formálně i věcně správná, konzistentní.

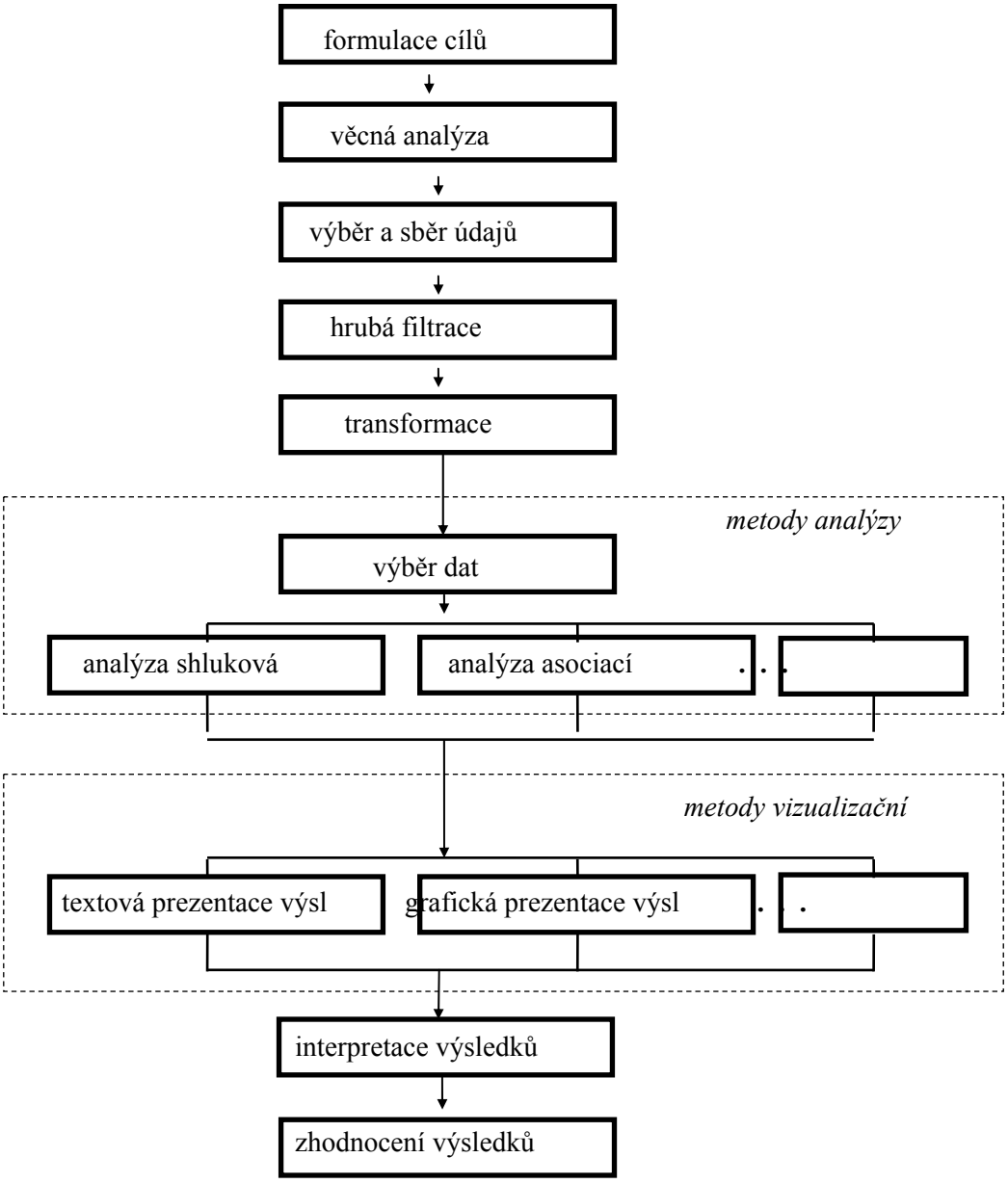
5. **Předzpracováním dat** rozumíme některé další úpravy dat před jejich analyzováním, které si vyžadují následně použité metody z hlediska věcné správnosti. Patří sem různé transformace údajů jako standardizace atributů pro odstranění závislosti na jednotkách měření nebo normalizace pro odstranění závislosti dat na velikosti objektů, někdy i dichotomizace či kategorizace údajů, transponování matice dat (i když tyto úpravy mohou být řazeny k předcházející etapě), transformace souboru dat do nových souřadnic odpovídajících hlavním komponent apod. Tyto transformace obvykle přímo souvisí s následným použitím některé metody.

6. **Vlastní analýzy** dat obsahují řadu metod, jejichž výsledky jsou cílem celého procesu. Patří sem metody matematické statistiky a příbuzných disciplín (gnostika, fuzzy) a patří sem metody explorační analýzy. Pro konkrétní data se obvykle provádí řada výpočtů realizujících jednotlivé metody. Jejich výběr souvisí se zadáním z etapy problémové analýzy.

7. Prezentace výsledků sice nepřináší nové výsledky, ale jejich nové zobrazení, **vizualizace** dat a výsledku analýz může výrazně ulehčit jejich pochopení a následnou interpretaci. Výsledky výpočtů nad daty mohou mít různou formu. Nejjednodušší forma numerická, byť uspořádaná do sestav, tabulek apod. obvykle znamená pro odborníka ještě mnoho práce při „překladu“ do vlastní odborné terminologie. Mnohem názornější jsou doplňující výstupy grafické nebo textové. Forma prezentace výsledků, jejich nové zobrazení může výrazně ulehčit jejich pochopení a následnou interpretaci. V systémech pro podporu rozhodování je na prezentaci výsledků kladen obzvláště velký důraz a všem uživatelům samozřejmě ušetří mnoho manuální práce.

8. **Interpretace výsledků.** Celá analýza nespočívá jen ve vlastních výpočtech nad daty. Aby byly výpočty užitečné, musí být provedena interpretace výsledků, jejich věcná slovní formulace. To je etapa, kterou obvykle provádí analytik společně s odborníkem, protože jde opět o rozhraní mezi jazykem matematiky a věcnou problematikou.

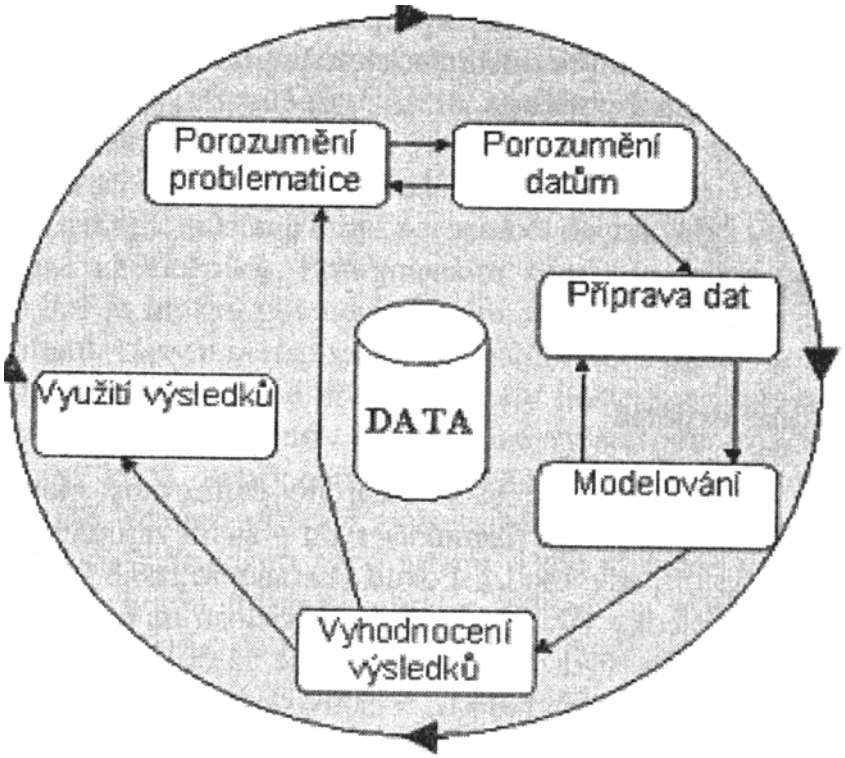
9. **Vyhodnocení průzkumu** je poslední fází zpracování, které provádí odborník, případně opět ve spolupráci s analytikem. Podle typu výzkumu vyhodnotí například, zda přinesl očekávanou kvalitu výsledků, zda se bude pokračovat v dalším sběru dat a dalších analýzách nebo jsme u konce výzkumu, zda průzkum neodhalil nic nového, zda jsou výsledky uspokojivé a zůstaneme u tohoto zjištění, zda jsou výsledky takové, že budou dále testovány metodami matematické statistiky, zda jsou výsledky natolik průkazné, že je možno formulovat zákonitosti (matematické, přírodní ap.) a přejít tak k vyšší fázi zkoumání a poznávání reality.



Obrázek 1.4. Životní cyklus procesu získávání znalostí

❑ Metodologie CRISP – DM

Často je používána jako standardní metodologie CRISP – DM (**CR**oss **I**ndustry **S**tandard **P**rocess model for **D**ata **M**ining). Jde opět o typ životního cyklu, neliší se v podstatě ničím od výše uvedeného postupu a je znázorňována takto:



Obrázek 1.5. Metodologie CRISP – DM

Metodologie SEMMA

Jednou z dalších metodologií často citovaných je SEMMA. Pomáhá snížit závislost na analytických expertech a zároveň slouží jako průvodce při implementaci data mining pro určitý problém.

Zkratka SEMMA znamená pět fází v data mining procesu:

Sampling – vzorkování, identifikování množiny vstupních dat

Exploring - prozkoumávání množiny dat statisticky a graficky

Modofocation – úprava dat, příprava na analýzu

Modeling – modelování, výběr vhodného prediktivního modelu

Assestment – zhodnocení, srovnání konkurenčních prediktivních modelů

Role analytika při dolování znalostí

Při cíleném výzkumu nebo při dolování z databází je nutno provést analýzu řešeného problému:

- **analýzu datovou** (jakou část reality popisují data, které údaje jsou či budou k dispozici pro analýzu, jakých syntaktických i sémantických typů jsou jednotlivé údaje, jejich rozdělení na dimenze pro agregování a fakty atd.),
- **analýzu problémovou** (jak je nutno data připravit – zakódovat, filtrovat, transformovat, než je možno použít konkrétních metod analýzy, jaké jsou požadovány typy výsledků, jaké typy metod jsou pro tato data použitelné)
- **analýzu forem prezentace výsledků**

❑ SW pro podporu dolování

V současném softwarovém světě je nabízena řada produktů, jejichž slogany zní velmi slibně jak z hlediska automatických postupů, tak uživatelského prostředí. Žádný z nich neslibuje dřinu, nutnost mnoho studovat, dlouho analyzovat, znovu a znovu analyzovat, probírat se množstvím banalit často krásně barevně provedených, 95% svých pracně získaných výsledků zahodit a jen při nemalém štěstí se dobrat malého množství skutečně nových znalostí. A to vše za ceny produktů desetitisícové, statisícové i milionové.

Podrobněji se o některých SW produktech zmíníme později.

Shrnutí pojmů 1.

Realita a její poznávání.

Statistika, analýzy mnohorozměrných dat, konfirmační a explorační analýzy dat.

Úlohy dolování znalostí z dat.

OLAP a dolování znalostí.

Využití dolování znalostí v praxi.

Typy uživatelů metod dolování a jejich potřeby.

Zdrojová data, použité metody pro různé oblasti praxe.

Využití dolování znalostí pro získání obchodních výhod, pro získání nových odborných znalostí, pro politické a sociální průzkumy.

Životní cyklus procesu získávání znalostí z dat. Etapy procesu získávání znalostí z dat

Metodologie CRISP – DM, metodologie SEMMA.

Role analytika při dolování znalostí.

Otázky 1.

1. Co je dolování znalostí z dat?
2. Jak souvisí metody dolování znalostí s metodami matematické statistiky?
3. Které metodologie pro dolování znalostí znáte a jak se liší?
4. Jaký je rozdíl mezi OLAP a dolováním znalostí?
5. Jaká je role analytika při dolování znalostí?

Semestrální projekt Analýza 1

Najděte si vlastní data pro analýzy.

Abychom měli na co aplikovat metody, se kterými se v MAD seznámíme, musíme si nějaká data opatřit. Možností, jak data získat, je několik:

1) Nejjednodušší možnost je najít a stáhnout nějaká data z internetu. Na internetu existují archívy dat, které se požívají pro testování dataminingových metod. Například poměrně známý archív UC Irvine Machine Learning Repository si lze prohlédnout na adrese http://archive.ics.uci.edu/ml/. Nevýhodou tohoto přístupu je, že data jsou již mnohokrát analyzována, a tudíž se toho o nich již mnoho nového prostřednictvím našich analýz nedozvíme. Další nevýhoda může nastat v případě, že se nedostaneme do kontaktu s autorem dat – pokud nám pak tedy v popisu dat o nich něco není jasné, jen obtížně se k vysvětlení dostáváme. *Pochopení dat a významů atributů v nich totiž hraje v celém analytickém procesu zásadní roli.* Pokud nechápeme, co vlastně analyzujeme (o čem přesně data jsou), tak **zaprvé** zřejmě povedeme celý analytický proces špatným směrem a **zadruhé**, což je ještě horší, buďto nebudeme vůbec schopni výsledky analýz interpretovat anebo je budeme interpretovat špatně, což je ta nejhorší možná varianta. Je proto dobré stáhnout si taková data, která buď budou velmi dobře a pochopitelně popsána v nějakém přiloženém souboru, nebo data, u kterých máme kontakt na jejich autora, který je ochoten k datům poskytnout nějaké dodatečné informace.
2) Druhou možností, jak získat data, je oslovit známé, kolegy či kamarády, kteří s nějakými daty pracují, a požádat je, aby nám dali data k dispozici. Tato varianta je poněkud náročnější (určitě nemáme takový výběr z mnoha souborů dat, jako na internetu, má však obrovskou výhodu v tom, že jsme v kontaktu s autorem dat, tudíž nejasnosti ohledně dat je možno s ním prokonzultovat. Další výhoda je ta, že výsledky analýz mohou být pro tuto osobu užitečné, což může být další motivace pro práci.
3) Třetí možností je udělat vlastní sběr dat, např. formou dotazníků (papírových či elektronických), v oblasti, která nás zajímá. Tato forma získání dat je nejpracnější, má ale samozřejmě své výhody – autorem dat jsme my, tudíž datům rozumíme a zřejmě nás zajímají i výsledky analýz, což nás motivuje k tomu, abychom analýzy provedli pořádně a získali nějaké seriózní výsledky.

Pokud jsme si tedy opatřili data kteroukoli formou, je potřeba se s nimi velmi dobře seznámit, zjistit případně již známé vztahy mezi atributy, popřípadě pokud existují, seznámit se s výsledky předchozích analýz nad těmito daty.

2. METODY PŘEDZPRACOVÁNÍ DAT

Čas ke studiu: 3 hodiny

Cíl Po prostudování této kapitoly budete umět

- rozdělovat data z hlediska možností budoucího použití dolovacích algoritmů
- filtrovat, transformovat a odvozovat data pro následné dolování
- pomocí výpočtu hlavních komponent data transformovat, nebo určit jejich charakteristické atributy

Výklad

2.1. Data pro analýzy

Data a jejich typy

Jak jsme uvedli, zkoumané objekty a jevy (z hlediska zákonitostí jejich chování a vztahů v realitě) popisujeme souhrnem jejich vlastností, atributů. Ze všech atributů objektu se pro zamýšlený výzkum vybírají ty, které s problémem souvisejí. Výběr relevantních atributů se provádí v rámci datové analýzy problému. Jde o úkol často velmi náročný a rozhodující pro kvalitu výsledku, ale pravděpodobně nealgoritmizovatelný. Expert na věcnou problematiku provede výběr samozřejmě efektivněji než laik. Při výzkumu dosud neznámé oblasti je vhodné volit v případě nejistoty o tom, zda atributy se zkoumanou skutečností souvisejí, raději atributů více.

Data pro analýzy můžeme dělit podle několika hledisek.

- **Z hlediska syntaktického** se dělí data podle standardních datových typů na numerická, textová, datumová a časová, logická, ostatní nestrukturované typy obvykle hromadně označované OLE (Object Linking and Embedding).

  V posledních desetiletích vznikly databáze nejen s klasickou strukturovanou informací (relační databáze), ale s uloženou informací mnoha dalších datových typů. Patří mezi ně například

  - transakční databáze obsahující data o obchodních transakcích, zahrnující mimo několik strukturovaných údajů o transakci dále seznam položek, které transakci tvoří; například seznam nakoupeného zboží při jednom nákupu (někdy jsou tato data nazývána nákupním košíkem),
  - objektově-relační databáze, které pracují s objekty s neatomickými atributy, jako seznamy, tabulkami atd.
  - textové databáze obsahující mimo strukturovanou informaci také nestrukturované rozsáhlé dokumenty,
  - prostorové databáze, například databáze geografických informačních systémů, obsahujících také data prostorová, jako zeměpisné souřadnice objektů,
  - temporální databáze obsahující také časové údaje a funkce pro práci s nimi,

- multimediální databáze obsahující data obrazová, audiosekvence, videosekvence, časové řady a další neatomické položky,
- heterogenní databáze obsahující všechny možné typy dat, například Web,
- a další.

Poznámka:

V tomto předmětu se budeme dále zabývat jen daty strukturovanými a vhodnou konverzí převeditelnými na data numerická (konverze viz níže). Rozvíjející se metody dolování z dat textových, obrazových, multimediálních a dalších zde probírat nebudeme. Často však tyto pokročilé metody dolování v první fázi převedou zkoumané údaje na data numerická a pak používají klasické metody dolování.

Pro dolování znalostí rozlišujeme údaje jemněji podle významu.

❑ **Z hlediska sémantického** dělíme dále data na

Numerické údaje

- **binární** (též dvouhodnotové, dichotomické, alternativní), nabývají pouze dvou hodnot

  Příklad 2.1. *{0,1}, {ano, ne}, {true, false}, {kuřák, nekuřák}, {muž, žena}, ...* ♦

- **kategoriální** (též kvalitativní, klasifikační, nominální), nabývají hodnot malého konečného počtu hodnot a znamenají příslušnost k jisté kategorii, udané svým očíslováním {0,1,...,k} bez významu kvantitativního, tedy bez uspořádání podle velikosti;

  Příklad 2.2. národnost $\in$*{česká,slovenská,polská,...} očíslována 1 = česká, 2 = slovenská, ...* ♦

- **ordinální** (též pořadové), nabývají také hodnot {0,1,...,k}, je u nich dáno přirozené uspořádání, případně bez významu vzdálenosti mezi hodnotami; obvykle se s nimi pracuje jako s kategoriálními, někdy jako s celočíselnými reálnými

  Příklad 2.3. známky *{1 až 5} ve škole jsou uspořádány, ale vzdálenosti mezi nimi obecně nejsou stejné, ...* ♦

- **reálné** (též reálněhodnotové, intervalové, kvantitativní), nabývají reálných hodnot z intervalu <a,b>, jsou zaznamenány s danou přesností; hodnota má absolutní význam v daných jednotkách; z hlediska významového se nerozlišuje, zda je přesnost na 0 desetinných míst a jsou tudíž celočíselná, nebo mají desetinnou část;

  někdy se uvádí jako samostatný typ údaje **poměrové**, obvykle udávající podíl dvou absolutních údajů; nabývají opět hodnot z <a,b> zaznamenané s danou přesností; hodnota má význam relativní, jejich stupnice není lineární; rozlišení s reálnými daty se projeví pouze při interpretaci výsledku.

  Příklad 2.4. *reálné* věk, výška, váha, cena ..., *poměrové* .... procento daně ♦

Nenumerické údaje

- **časové**, vyjadřující absolutní datum nebo čas události, případně časový úsek; používají se obvykle k přepočtu na časový úsek (údaj reálný) nebo pro kategorizaci.

  Příklad 2.5. *časové údaje* datum narození, termín splatnosti, začátek výukové hodiny, okamžik startu, *nebo doba události* délka letu, doba běhu na 100m, počet dnů školení *apod.* ♦

- **textové**, vyjadřující slovně nebo znakovým kódem popisovanou vlastnost; podle okolností se používají k výběru podmnožin objektů, je-li to možné, tak se číselně zakódují a zpracovávají jako data kategoriální či ordinální, případně se zpracovávají dalšími metodami.

Příklad 2.6. *typické texty bez možnosti jednoduchého zakódování jako* anotace článku, text článku, poznámka o čemkoliv, *..., ale případně také snadno zakódovatelné texty jako* rodinný stav *{svobodný, ženatý, ...}, nebo jinak upravitelné údaje jako* rodné číslo *(lomítko se zruší nebo zapíše jako desetinná tečka) apod.* ♦

- **grafické, zvukové, ostatní neatomické údaje,** obecně označované **OLE**; pokud se účastní zpracování, předzpracují se obvykle náročnějšími metodami, které jsou obsahem samostatných disciplín a kódují se na některý z předchozích typů.

  Příklad 2.7. *záznam EKG, záznam hudby, foto, videozáznam, ...* ♦

Většina metod analýzy mnohorozměrných dat pracuje pouze s numerickými údaji a před zpracováním se ostatní datové typy vhodně kódují. Způsob kódování je součástí etapy analýzy a souvisí s potřebami úlohy a budeme je probírat v rámci transformací údajů.

2.2. Metody filtrace a integrace dat

❑ Zdroje dat

Data vstupující do zpracování mohou pocházet z různých typů zdrojů. To může mít vliv na způsob jejich dalšího zpracování.

Buď jsou dobře navržena, zakódována a sesbírána přímo pro potřeby tohoto průzkumu, pak se v nich mohou objevit jen chybné údaje způsobené chybami při jejich měření, sběru, záznamu na médium počítače.

Druhou možností je, že byla data pořízena bez předchozí analýzy a obsahují údaje sice související s problémem, ale nejsou dosud ve formátu vhodném pro analýzy. Prakticky všechny dále uváděné metody vyžadují data numerická.

V praxi se vyskytující datové typy, časové, datumové, textové, grafické či zvukové, pokud se mají účastnit dalšího zpracování a nebýt jen informativním doplňkem pro zkoumané objekty, je nutno pro ně dodatečnou datovou analýzou navrhnout způsoby, jak z nich získat údaje numerické. Nebo jsou data vzniklá pro jiný účel a dodatečně zvolena pro analýzu. Pak je potřeba navíc vybrat údaje pro zamýšlené zpracování relevantní a z formálního hlediska vhodná.

Pokud data získává nějaký automat, například pomocí čidel, chápeme to jako speciální případ první možnosti a data pak nejsou zatížena lidskými chybami.

❑ Filtrace dat

Filtrací dat nazýváme řadu akcí, vedoucích k získání dat připravených ke spuštění dolovacích algoritmů. Je to činnost jen zdánlivě jednoduchá, skoro vždy však spotřebuje mnohem více času a práce, než vlastní dolovací výpočty.

Pro informovanější rozhodování je vhodné provést nejprve výpočty základních statistických charakteristik každého atributu.

Filtrace pak zahrnují

- výběr atributů vhodných k analýzám
- ošetření nebo vyloučení dat chybných, chybějících, redundandních, irelevantních, konstantních
- sjednocení formátů, měrných jednotek
- numerické zakódování některých dat, sjednocení kódování, kategorizace a dichotomizace dat; některé z těchto úprav mohou být řazeny již k transformacím dat.

Výsledkem filtrace mají být data numerická, formálně i věcně správná, konzistentní. Numerické atributy budeme rozlišovat na reálné, ordinální a kategoriální. Některé atributy se mohou opakovat v různých podobách (např. reálný věk a věk kategorizovaný do tříd, viz též odvozená data) pro různé typy metod dolování.

Jednotlivé metody filtrací:

- **Integrace dat, sjednocení formátů, měrných jednotek**

  je první operací při získávání dat z různých zdrojů. Musí se navrhnout současně s výběrem dat a na míru zdrojovým datům. Operace pro integrování dat je vhodné uložit do metadat pro pozdější opakující se načítání přírůstků dat nebo pro opakované zpracování na jiném vzorku dat.

- **Výběr relevantních dat pro analýzy – tzv. volba modelu**

  výběr dat, která se mohou používat pro různé typy dolovacích metod; výběr může být pro každou metodu poněkud jiný, výběr se provádí především pro atributy, někdy i pro objekty. Obvykle se tento proces nazývá tvorbou modelu pro dolování.

  Dle zvolené metody rozdělení vybraných atributů podle významu na

  dimenze, fakty (viz datové sklady)
  antecedenty, sukcedenty (předpoklady a následky, viz asociace, rozhodovací stromy)
  ovlivnitelné, neovlivnitelné (viz asociace i jiné metody)

- **Statistické charakteristiky atributů**

  Výpočet charakteristických hodnot atributů - průměr, minimum a maximum, standardní odchylka, medián, četnosti výskytů jednotlivých hodnot (frekvencí), počty chybějících údajů.

  Je užitečné tyto statistické charakteristiky zařadit k metadatům a uschovat. Jsou využívány u mnoha metod dolování. V etapě předzpracování jsou velmi užitečné k prvnímu „náhledu“ na data pomocí vizualizačních metod i k odhalení některých chyb v datech.

- **Statistické charakteristiky dvojic atributů**

  některé statistické charakteristiky se již používají pro analýzy, například

  korelační koeficienty
  regresní křivky
  frekvenční (kontigenční) tabulky

  Pro data **X** a jejich kategoriální veličiny A1 a A2 má frekvenční tabulka tvar

A1 \ A2	0	1	2	...	k2	
0	a_{00}	a_{0}			a_{0h2}	a_{0}
1	a_{10}	a_{11}			a_{1h2}	a_{1}
2						
...						
k1	a_{h10}	a_{h11}	a_{h12}		a_{h1h2}	$a_{h1.}$
	$a_{.0}$	$a_{.1}$	$a_{.2}$		$a_{.h2}$	m

- **Odhalení dat chybných**
  - pro záznam dat do databáze použít vstupní program s kontrolami syntaktickými i logickými, ne textový editor,
  - použít kontrolní funkce při prvotním načítání dat do DM systému,
  - opakovaná kontrola celého procesu přípravy dat,
  - využití statistických charakteristik – například minima a maxima jednotlivých atributů porovnat se zadanými doménami atributů.
- **Řešení dat chybějících**
  - statistiky odhalí chybějící údaje, důležité jednotné označení v celém souboru a ve shodě s potřebami analýz,
  - pokud metody neumí zpracovat soubor s chybějícími údaji a data neúplná jsou, je nutné situaci řešit; pohodlnou možností je vyloučit taková data ze zpracování - celý atribut, objekty; soubor se zmenší,
  - jinou možností je doplnit chybějící údaje „vymyšlenými“ hodnotami; strategie optimistická (maximálními hodnotami atributu), pesimistická (minimálními), neutrální (průměrnými), jiná; soubor dat zůstane celý, ovšem s jistou nepřesností,
  - některé metod umí zpracovat nebo eliminovat chybějící data a stačí, když jsou správně označena.

Příklad 2.8.

Jsou dána data o pacientech interního nemocničního oddělení, obsahující údaje:

Pacient (pohlavi {muž, žena},
vek [rok],
stav {svobodny, zenaty, rozvedeny, vdovec},
prijem [datum příjmu],
puls [počet],
tlak [horni, dolni],
zlozvyky {nic, kouří, pije, kouří i pije},
diagnoza [nemoc],
vysledek {propuštěn domů, přemístěn na jiné odděl, zemřel})

Navrhněte metody předzpracování těchto dat.

Řešení: atributy **pohlavi, stav, zlozvyky** *a* **vysledek** *se budou jednoduše kategorizovat = kódovat na 0 / 1 a 0 / 1 / 2 / 3 / 4,*

vek *bude vhodné kategorizovat po konzultaci s lékařem do nového atributu* **vek_k**, *například do intervalů 1-3,4-6,7-10,11-14,15-17,18-25,26-35,36-50,51-60, ...*

z atributu **prijem** *se odvodí několik atributů* den, mesic, rok, …

diagnóze *se po konzultaci s lékařem přidělí numerický kód, odpovídající nemoci a skupinám nemocí, případně se použije celostátní číselník diagnóz.*

♦

2.3. Metody transformací dat

Data filtrovaná nemusí být ještě vhodně připravena pro všechny metody dolování. Například při shlukování se často měří podobnost objektů Eukleidovskou vzdáleností a v případě atributů různých měrných jednotek a jejich řádově rozdílných hodnot by vliv některých atributů silně negativně ovlivnil výsledek. Proto je někdy nutné odstranit vlivy měrných jednotek, jindy odstranit vliv velikosti objektů.

K transformacím dat řadíme jednak převádění dat mezi datovými typy – pokud to použité metody vyžadují, jednak složitější transformace, řešící například problémy se vzájemnou porovnatelností informací.

❑ Převody datových typů

Různé metody dolování

- **Kategorizace (diskretizace) reálných dat**
  - kategorizace reálných údajů – ekvidistantní intervaly, dle rozložení četností v intervalech; kategorie se očíslují a data se transformují; ruční možnost nepravidelného rozložení kategorií dle konkrétní situace.
  - automatická kategorizace, je-li to vhodné vzhledem k daným hodnotám atributu, volí se počet kategorií
- **Kategorizace nenumerických dat**
  - textových, datumových, časových, dalších,

    pokud atribut nabývá jen malého počtu hodnot, je možno okódovat hodnoty, uložit je do číselníku (metadat) a dále pracovat s numerickými kategoriálními nebo ordinálními údaji.
- **Dichotomizace dat**
  - atribut kategoriální o **k** kategoriích převést na **k** atributů s hodnotami z {0,1} - pro každou kategorii znamená jedna nová hodnota údaj {nabývá, nenabývá}.
  - pomocí výrazu [h1, . . ., hc], kde hi $\in$ {0,1,2, . . ., k} , hi $\neq$ hj pro i $\neq$ j. Výraz

    Oi (Aj) [h1, . . ., hc]

    znamená, že Oi (Aj) nabývá nebo nenabývá jedné z hodnot h1, ..., hc. Pak můžeme z kategoriální Aj vytvořit binární Ak podle pravidla Oi (Aj) [h1, . . ., hc] = 1 právě, když Oi (Aj) nabývá jedné z hodnot h1, . . ., hc, jinak Oi (Aj) [h1, . . ., hc] = 0

Příklad 2.9.

V datech o pacientech jsou mj. atributy věk, národnost*, ... Pokud budeme data zpracovávat metodou pro reálné atributy, můžeme věk použít beze změn,* národnost *(kategoriální) použít nemůžeme. Pokud použijeme metodu pro kategoriální data, můžeme atribut* věk *kategorizovat. Při datové analýze navrhneme způsob kategorizace, zde pro pacienty například do zadaných intervalů, které nemusí být pravidelné ani stejně četné, ale mají smysl ze zdravotního hlediska:*

věk < 10	věk_k = 1
věk $\in$ <11,15>	2
věk $\in$ <15,18>	3
věk $\in$ <18,26>	4
věk $\in$ <27,35>	5
atd.	

Pokud metoda vyžaduje data binární a chceme pomocí ní zpracovávat i věk, *můžeme kategorizovanou hodnotu* věk_k *převést na několik binárních atributů* věk_b_1, věk_b_2, ..., *kde každá hodnota nabývá hodnoty 0 nebo 1 podle toho, zda věk patří nebo ne do příslušné kategorie (intervalu).*
Jiný způsob dichotomizace můžeme podle potřeby použít například seskupováním kategorií:

věk_k ∈ <1, 2, 3>	věk_b_1 = 1, jinak věk_b_1 = 0
věk_k ∈ <4, 5,6>	věk_b_2 = 1, jinak věk_b_2 = 0
věk_k ∈ <7, 8, 9>	věk_b_3 = 1, jinak věk_b_3 = 0

♦

❑ Transformace dat

Data vhodných datových typů, bezchybná, formálně správná, nemusí být ještě vhodným vstupem pro všechny typy analytických metod.

- **Normalizace objektů**

Normalizací objektů s reálnými atributy rozumíme odstranění závislosti dat na velikosti objektů; objekty se normalizují podle transformačního vztahu (v matici dat **X** je původní hodnota j-tého atributu u i-tého objektu označena x_{ij}, přepočtená hodnota z_{ij}

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{n} x_{ij}^2}} = \frac{x_{ij}}{P}$$

kde P je norma objektu.

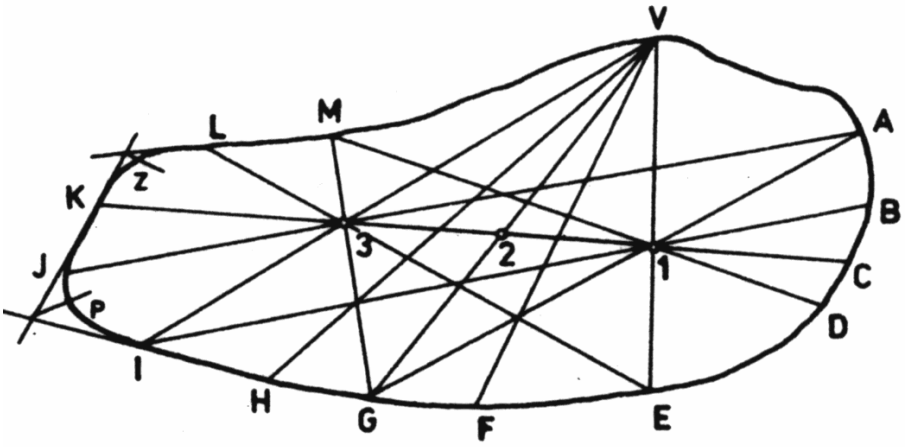
X

	A1	**A2**	**…**	**Aj**	**…**	**An**
O1	x11	x12		…	…	
	…	…				
Oi	xi1	xi2				
	…	…				
Om	xm1	xm2				

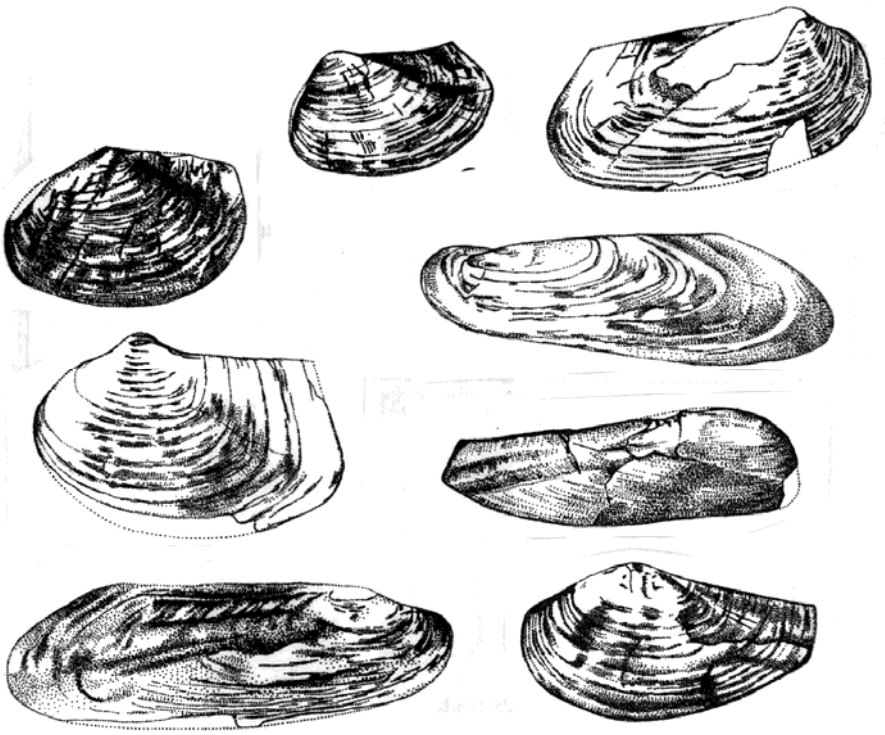
Příklad 2.10. [13]

Je dána matice dat, kde objekty jsou mlži několika rodů, popsané 25 rozměry (v mm) dle obrázku. Hledají se tvarově blízké množiny objektů.

Ke stejnému rodu zřejmě patří tvarově stejní, ale velikostí rozdílní mlži. Pokud bychom nechali absolutní rozměry v mm pro jejich popis a shlukování (hledání skupin podobných), zřejmě by velikost měla zkreslující vliv. Proto je nutné před shlukováním objekty – mlže normalizovat.



Obrázek 2.1. 25 rozměrů mlže popisujících tvar



Obrázek 2.2. Několik ukázek nalezených mlžů

♦

- **Standardizace atributů**

Standardizací reálného znaku rozumíme odstranění závislosti jednotlivých atributů na jednotkách měření. To je nutné prakticky u všech atributů s reálnými hodnotami. Metody založené na pojmu vzdálenosti, použité bez předchozí standardizace, by dávaly výsledky zkreslené vlivem rozdílných jednotek. Výsledná hodnota atributu po standardizaci je zhruba v intervalu <-1,1>.

X

	A1	A2	…	Aj	…	An	Aj_s
O1	x11	x12		x1j	…		z1j
	…	…		…			…
Oi	xi1	xi2		xij			zij
	…	…		…			…
Om	xm1	xm2		xmj			zmj

Atributy (sloupce) se normalizují podle transformačního vztahu

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

$$\bar{x}_j = \frac{\sum_{i=1}^{m} x_{ij}}{m} \qquad s_j = \sqrt{\frac{\sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2}{m}}$$

Příklad 2.11.

Jsou dány údaje různých druhů o sledovaných studentech:

Student (věk, výška, váha, známka_MAT, známka_DEJ, ..., skok_vys, skok_dal, beh_100m, ...)

Úkolem může být i zkoumání, jestli nějak spolu souvisí tělesné a duševní výkony. Pak intervaly známek <1,5> a skoků <0, 800> by byly nesouměřitelné, absolutní hodnoty skoků v cm by měly větší „váhu" než známky. Po standardizaci bude váha obou stejná.

♦

- **Transponování matice dat**

Při hledání optimální podmnožiny atributů, charakterizujících daný objekt je možno použít shlukování vlastností, tedy původních atributů. K tomu je třeba matici dat předem transponovat, zvolit za „objekty" původní aspekty. Výsledky metod analýzy dostanou nový význam.

Příklad 2.12.

U shlukovacích metod si uvedeme příklad hledání skupin podobných lidí na základě odpovědí na otázky v dotazníku. To provedeme klasickým použitím shlukovací metody.

Když se ale datová matice transponuje, shlukováním dostaneme skupiny podobných si vlastností. Je to jiný výsledek, dává jinou informaci.

♦

- **Odstranění lineárních závislostí mezi atributy**

Může se stát, že v datech některé atributy jsou vzájemně závislé, přičemž tato závislost není předem známa. Výsledky tak mohou zkreslené a je vhodné takové závislosti před dolováním odhalit a korigovat. Tímto problémem se zabývá **metoda hlavních komponent.**

Protože jde o metodu, která se řadí jak k transformačním metodám předzpracování dat, tak k metodám analýzy dat, věnujeme jí podrobně následující kapitolu.

2.4. Odvozování dat

❑ Generování odvozených atomických údajů

U úloh dolování je často užitečné z existujících atributů pro analýzy odvodit nové atributy, byť jsou jejich údaje redundandní. Je sice pravda, že v případě jejich potřeby je možné je vždy znovu vypočítat a nerozšiřovat tak velikost databáze, ale předpočítané údaje jsou pro využití pohodlnější a hlavně nezdržují časově náročné dolovací algoritmy.

Neexistuje zde jednotný návod k použití, záleží na potřebách analýz a nápaditosti analytika. Je velmi užitečné, když používaný SW má k dispozici nástroje pro zadání vzorců pro výpočet odvozených veličin, případně zabudovány funkce pro některé standardní často používaná odvození.

Příklad 2.13.

Datumové a časové položky (které v databázích mají pravidelně své datové typy) se samy o sobě metodami dolování zpracovávat nedají (nejsou to numerické údaje), ale jsou bezpochyby pro analýzy potřebné. Proto se z nich odvozují mnohé numerické údaje, jako den, měsíc, rok, den v týdnu, roční období, znamení zvěrokruhu, …, *nebo ze dvojice časových atributů délku období apod.*

Klasická odvození jsou věk, pohlaví, datum narození z rodného čísla, počet dnů, měsíců, roků *ze dvou datumových údajů*, *kategorizace* rodinného stavu slovně vyjádřeného apod.

♦

❑ Agregace údajů

Zvláštním případem odvozených dat jsou data agregovaná:

sumy, průměry, minima a maxima, počty případů, relativní četnosti apod.

Z nahromaděných dat za různá časová období, za různé geografické celky či jiné dimenze jsou tyto údaje bezpochyby užitečné samy o sobě a někdy jejich grafické či tabulkové prezentace bývají vydávány i za výsledky dolování.

V dalších kapitolách se seznámíme s tím, jak agregované údaje mohou využívat některé dolovací algoritmy.

Jinak (jak již víme) jsou agregované údaje, i hierarchicky uspořádané, hlavním zdrojem výsledků, které produkuje OLAP se svými prezentačními metodami.

Příklad 2.14.

Údaje o pacientech a jejich nemocech dlouhodobě sbírané mohou sloužit k vědeckým a statistickým účelům. Nebudou pak základním objektem zájmu jednotlivé případy nemocí, ale

- *počty případů nemocí po dnech (při epidemii), po týdnech a měsících či rocích, v jednotlivých okresech, krajích, zemích apod.,*
- *průměrné dávky léků podávané jednotlivým věkovým, profesním a jiným skupinám pacientů,*
- *součty množství spotřebovaných léků podle druhů, zemí,*
- *minimální a maximální dávky nových léků podávaných testované skupině pacientů*
- *atd.*

♦

Shrnutí pojmů 2.

Zdroje dat pro analýzy.
Datové typy syntaktické.
Datové typy sémantické.
Integrace dat, sjednocení významu atributů, formátů a měrných jednotek.
Metody filtrace dat.
Numerické zakódování údajů.
Metody transformací dat.
Převody datových typů. Kategorizace, dichotomizace.
Normalizace a standardizace. Transponování matice dat.
Odvozování dat.
Generování odvozených atomických údajů.
Agregace údajů a jejich využití.

Otázky 2.

1. Co všechno zahrnuje etapa předzpracování dat?
2. Jak rozlišujeme z hlediska sémantiky údaje pro zpracování a jak numerické údaje?
3. Co znamená integrace dat a proč se provádí?
4. Z jakých hledisek je potřeba zkontrolovat data a provést operace s daty při jejich filtraci?
5. Proč se převádí některé sémantické datové typy na jiné?
6. Jak se provádí převody datových typů?
7. Co znamenají transformace dat a které transformace znáte?
8. Co je standardizace, pro jaká data a proč se provádí?
9. Co je normalizace, pro jaká data a proč se provádí?
10. Co jsou hlavní komponenty pro datovou matici?
11. Kdy se používá výpočet hlavních komponent a k čemu?
12. Jaký význam pro dolování mají odvozené údaje?
13. Jaké typy odvozených údajů znáte?

Úlohy k řešení 2.

1. Jsou dána data z evidence sportovního gymnázia s atributy: školní rok, třída, učitel [jméno], jméno [studenta], pohlaví [chl / dív], věk, výška, váha, dále maximální výkony sportovní za tento rok ve skoku vysokém [cm], dalekém [cm], běhu -100m [12.3 sec], běhu - 400m a závěrečné známky z češtiny, cizího jazyka [ruština a později angličtina], matematiky, fyziky, dějepisu a zeměpisu. Data jsou pořizována za dobu 30 let.

   Navrhněte úplné předzpracování těchto dat.

2. Jsou dána data BANKA, obsahující 1027 záznamů za minulých 5 let. Jejich struktura je

Banka (pohlavi [m / z], vek [roků], svobodny [a / n], nezamestnany [a / n], cim_ruci [auto, dum, PC, kolo], problematicky_region [a / n], mes_prijem [Kc], hotovost_u_banky [Kc], pocet_mesicu_splatky, pocet_roku_u_souc_firmy, dostal_uver [a / n], splacel_bezproblemove [a/n])

Navrhněte úplné předzpracování těchto dat.

3. Jsou dána data Pacienti s atributy infarkt [ano / ne], ang_pectoris [ano / ne], berc_vred [ano / ne], vaha [kg], vyska [cm], kurak [ano / ne], pohlavi [muž / žena], vek [roků], město [ano / ne], duchodce [ano / ne], stres [ano / ne].

   Navrhněte úplné předzpracování těchto dat.

4. Data pocházejí ze sociologického výzkumu (ankety). Tato anketa se týká vědního oboru, který se nazývá antroponymie. Je to vlastně součást onomastiky (onomatologie), zkoumající vlastní jména živých bytostí. Pro informaci onomastika se dále dělí na toponomastiku (zkoumá vlastní jména neživých věcí) a chrématonomastiku, zkoumající vlastní jména lidských výtvorů a zařízení.
   Anketa byla prováděna v několika základních školách s polským jazykem vyučovacím v pohraničí (polsko - českém), a jejími účastníky byli nejen žáci těchto škol, ale i jejich rodiče. Cílem ankety bylo popsat antroponymické struktury ve školním i mimoškolním prostředí. Co všechno se pod tímto pojmem skrývá, nám nejlíp popíše výčet jednotlivých otázek pro žáky“

   1. Křestní jméno, příjmení
   2. Věk
   3. Základní škola v: Třída
   4. Bydliště: okres: Město/Vesnice
   5. Máš sourozence? Počet:
   6. Jakou formu tvého jména používal (používá)
      a) matka (v předškolním věku, současně)
      b) otec (v předškolním věku, současně)
      c) babička
      d) dědeček
      e) sourozenci
      f) kamarádky
      g) kamarádi
      h) učitelé
   7. Znáš přezdívku týkající se
      a) jména tvého, kamarádky, kamaráda, učitele
      b) příjmení tvého, kamarádky, kamaráda, učitele
   8. Znáš říkadla týkající se tvého jména, jména kamarádek, kamarádů?
   9. Jaké formy jmen sourozenců, kamarádů, kamarádek používáš?
   10. Jaké formy jmen používají rodiče při vzájemném oslovování?
   11. Jakým způsobem se nejčastěji oslovují kamarádi, kamarádky u vás ve třídě (jméno, příjmení, přezdívka, nevím)?
   12. Jakým způsobem učitelé nejčastěji oslovují žáky u vás ve třídě (jméno, příjmení, přezdívka, nevím)?
   13. Jsi spokojen se svým jménem?

Semestrální projekt Analýza 2

Pro svá data navrhněte úplné předzpracování.

3. HLAVNÍ KOMPONENTY

Cíl Po prostudování této kapitoly budete vědět

- co jsou latentní proměnné
- jak se dají vypočítat
- k čemu slouží výsledky výpočtu hlavních komponent
- kdo transformaci do hlavních komponent navrhuje a na základě čeho

Výklad

Skryté proměnné

Mezi transformační metody je možno zařadit i metodu hlavních komponent, i když její výsledky dávají i užitečné informace o datech. Jde o součást lineární faktorové analýzy, patřící mezi tzv. modely s latentními proměnnými.

Obsahuje-li datová matice závislé atributy, mohou být některé analýzy zkreslené. Metoda vychází z předpokladu, že dvě proměnné mohou být závislé proto, že obě měří tutéž skrytou společnou veličinu, nazývanou společný faktor nebo hlavní komponenta.

Příklad 3.1.

neměřitelná veličina = diagnóza,
měřitelné atributy = teplota, tlak, nález v krku, ...

nebo

neměřitelná veličina = nadání matematicko-logické
měřitelné atributy = známka z matematiky, známka z fyziky, ...

♦

Hlavních komponent může být v datech více a lze je matematicky zkonstruovat.

Vzájemná závislost a nezávislost znaků

Některé znaky (atributy) lze považovat za náhodné veličiny ve smyslu matematické statistiky. Pak můžeme při posuzování jejich vzájemné závislosti používat prostředků matematické statistiky.

Ideálně by měly být zkoumané objekty charakterizovány atributy vzájemně nezávislými. Protože vzájemná závislost atributů může ovlivnit výsledky analýzy, je vhodné takové závislosti předem vyloučit, nebo alespoň o nich vědět a kvantitativně je ohodnotit.

Závislost atributů může mít nejrůznější charakter a obecný postup pro její nalezení neexistuje. Pokud však předpokládáme závislost alespoň přibližně lineární, lze použít pro její vyjádření koeficient kovariance nebo koeficient korelace.

Kovariance je míra vzájemné závislosti mezi dvěma náhodnými veličinami. Protože je její hodnota závislá na hodnotách zkoumaných atributů, používá se častěji tato hodnota standardizovaná - **koeficient korelace**. Ten pro dva atributy Ai, Aj nabývá hodnot z intervalu < -1, 1>, přičemž pro hodnoty blízké nule považujeme atributy Ai, Aj za nezávislé, pro hodnoty blízké -1 nebo 1 považujeme oba atributy za lineárně závislé.

Pro matici dat X s reálnými atributy Ai, Aj

X

	A1	**A2**	**Ai**	**…**	**Aj**	**An**
O1	x11	x12	x1i	…	x1j	
	…	…			…	
Oi	xi1	xi2	xii		xij	
	…	…			…	
Om	xm1	xm2	xmi		xmj	

je definována kovariance $\mathbf{k_{ij}}$ a korelace $\mathbf{r_{ij}}$ vztahy

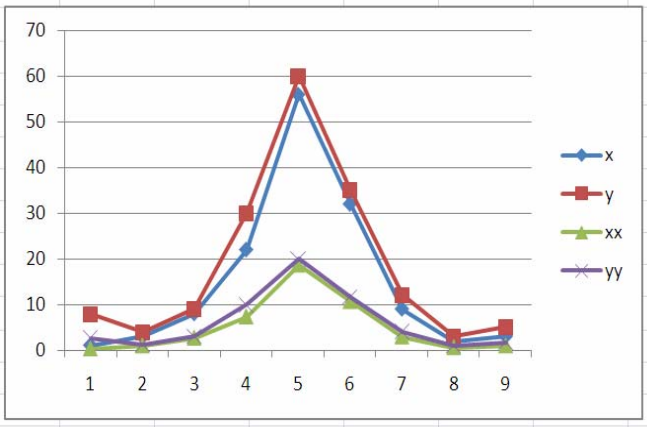
$$k_{ij} = \frac{1}{n}\sum_{i=1}^{m} x_{li} \,.\, x_{lj} - \bar{x}_i \,.\, \bar{x}_j \qquad\qquad r_{ij} = \frac{k_{ij}}{s_i \,.\, s_j}$$

kde x_{ij} je původní hodnota j-tého atributu u i-tého objektu v matici dat **X**,
$\underline{x}_i$ je střední hodnota i-tého atributu,
s_i je standardní odchylka i-tého atributu.
k_{ij} je hodnota kovariance i-tého a j-tého atributu
r_{ij} je hodnota korelace i-tého a j-tého atributu.

Příklad 3.2.

Jsou dány atributy x, y, platí xx = x/3, yy = y/3. Závislost obou dvojic atributů (x, y), (xx, yy) je tedy stejná.

x	y	xx	yy
1	8	0,333333333	2,666667
3	4	1	1,333333
8	9	2,666666667	3
22	30	7,333333333	10
56	60	18,66666667	20
32	35	10,66666667	11,66667
9	12	3	4
2	3	0,666666667	1
3	5	1	1,666667
covar(x,y)=	316,951	corel(x,y)=	0,992
covar(xx,yy)=	37,316	corel(xx,yy)=	0,992



Kovariance dvojic (x, y) a (xx, yy) jsou různé, závisejí na absolutních hodnotách atributů, standardizované korelace jsou shodné. Na grafu vidíme vysokou závislost obou dvojic.

♦

Korelační matice reálných atributů

Pokud vypočteme koeficienty korelace pro všechny dvojice atributů a sestavíme je do symetrické čtvercové matice **R** typu n x n, dostaneme **korelační matici**.

R

	A1	**A2**	**Ai**	**…**	**Aj**	**…**	**An**
A1	1	r12	r1i	…	r1j		r1n
…	…	…			…		
Ai	ri1	ri2	1		**rij**		rán
…	…	…			…		
Aj	rj1				1		rjn
…							
An	rn1	rn2	rni		rnj		1

Vlastní čísla a vlastní vektory matice korelační

Pro výpočet hlavních komponent potřebujeme nejprve vypočítat vlastní čísla a vlastní vektory matice **R** (K). Zopakujeme si proto některé pojmy z maticového počtu.

Charakteristickým mnohočlenem čtvercové matice **R** nazveme determinant matice $|\mathbf{R} - \lambda \mathbf{E}|$, kde **E** je jednotková matice řádu n. Jeho kořeny λ_i nazýváme vlastními (charakteristickými) čísly matice **R**.

Jde o matici symetrickou (protože $r_{ij} = r_{ij}$) s reálnými prvky. Platí, že vlastní čísla takové matice jsou reálná čísla. Protože jde o matici řádu n, je determinant matice $|\mathbf{R} - \lambda \mathbf{E}|$ mnohočlenem stupně n a charakteristická rovnice má n reálných kořenů. Bez újmy na obecnosti si můžeme kořeny uspořádat sestupně podle velikosti.

Stopou čtvercové matice R nazýváme součet diagonálních prvků matice $r_{11} + r_{22} + \ldots + r_{nn}$.

Vlastní čísla jsou tedy řešením charakteristické rovnice.

$$| \mathbf{R} - \lambda \mathbf{E} | = 0$$

kde λ_i, pro i = 1, . . ., n (kořeny této rovnice) jsou vlastní čísla matice R.

E je jednotková matice.

Jde tedy o rovnici

$$\begin{vmatrix} \mathbf{1-\lambda} & r12 & r1i & \ldots & r1j & & r1n \\ \ldots & \ldots & & & \ldots & & \\ ri1 & ri2 & \mathbf{1-\lambda} & & \mathbf{rij} & & rin \\ \ldots & \ldots & & & \ldots & & \\ & & & & \mathbf{1-\lambda} & & \\ rn1 & rn2 & rni & & rnj & & \mathbf{1-\lambda} \end{vmatrix} = \mathbf{0}$$

Jejím řešením jsou (sestupně dle velikosti uspořádaná) vlastní čísla $\boldsymbol{\lambda_1, \lambda_2, \ldots, \lambda_n}$.

Vlastní vektory $\mathbf{u_i}$ matice **R** dostaneme pomocí vlastních čísel korelační matice. Pro každé λi se odpovídající vlastní vektor $\mathbf{u_i}$ (i-tý sloupec **U**) vypočte z rovnice

$$(\mathbf{R} - \lambda_i \mathbf{E})\, \mathbf{u_i} = \mathbf{o}$$

kde **o** je nulový vektor.

Pro n vlastních čísel dostaneme n vlastních vektorů (sloupců) a tedy čtvercovou matici vlastních vektorů **U**.

❑ Výpočet vlastních čísel a vlastních vektorů

Metod výpočtu vlastních čísel a vektorů existuje v numerické matematice mnoho, pro symetrické matice existují speciální metody. Jednou z nich je například Jacobiho metoda, založená na maticové redukci.

Výpočet vlastních čísel a vlastních vektorů matice **R (K)** metodou Jacobiho:

Na vstupu je výchozí matice **R** a pomocná matice **U**, která je jednotková; pak se provádí eliminace **R** do diagonální matice a současně se paralelně transformuje stejnými operacemi pomocná jednotková matice **U.**

Před eliminací:

R

1-λ	…	r1i	…	r1n
ri1		**1-λ**		rán
…				
rn1	…	rni		**1-λ**

E

1	0	…	0	0
0	1			
…		…	…	…
0			1	0
0			0	1

Po eliminaci:

eliminovaná **R**

λ1	…	0	…	0
…				…
0	…	**λi**	…	0
…				…
0	…	0	…	**λn**

eliminovaná E = **U**

$\mathbf{u_{11}}$	…	$\mathbf{u_{1i}}$	…	$\mathbf{u_{n1}}$
…				
$\mathbf{u_{i1}}$		$\mathbf{u_{ii}}$	…	$\mathbf{u_{in}}$
…				…
$\mathbf{u_{n1}}$		$\mathbf{u_{ni}}$	…	$\mathbf{u_i}$

Na výstupu jsou pak v diagonále matice **R** vlastní čísla původní matice **R** a ve sloupcích pomocné matice **U** jsou vlastní vektory původní matice **R**. Barevně vyznačená jsou v 1. matici vypočtená vlastní čísla, ve 2. matici 1. vlastní vektor odpovídající 1. vlastnímu číslu.

Stopa eliminované matice **R** je (**λ1 + λ2 + … + λn**). Významnost jednotlivých vlastních čísel určíme jako procentuální podíl jednotlivých vlastních čísel vůči stopě, tedy

$$\lambda i * 100 / (\lambda 1 + \lambda 2 + \ldots + \lambda n)$$

Pokud významnost vlastního čísla je vůči stopě malá, příslušné hlavní komponenty můžeme zanedbat a snížit tak počet nových atributů.

❑ Hlavní komponenty

Je-li spočítána matice **U** s vlastními vektory korelační matice **R**, můžeme vypočíst hledané skryté proměnné jako hlavní komponenty původních atributů. Transformace do hlavních komponent se provádí pomocí vlastních vektorů matice korelační:

$$\mathbf{Z} = \mathbf{X}\,\mathbf{U}$$

kde **X** je původní matice dat, **Z** je transformovaná matice dat, **U** je matice vlastních vektorů vypočtených z matice korelační **R** pro všechny dvojice atributů x_i, x_j matice dat **X**.

Po prvcích jde o transformační rovnici

$$z_{ij} = \sum_{k=1}^{n} (x_{ik} * u_{kj})$$

Výsledkem je nová tabulka dat, v níž jsou původní (případně závislé) atributy {A1, …, An} nahrazeny vzájemně nezávislými hlavními komponentami {H1, …, Hn}, kde každá komponenta je jistou lineární kombinací původních atributů. Podle významnosti vlastních čísel je možno z n komponent uvažovat jen několik prvních dostatečně významných a snížit tak počet nových atributů.

Problémem někdy může být pojmenování nových komponent.

Algoritmus výpočtu hlavních komponent

Dána datová matice **X** (= tabulka s numerickými - reálnými atributy).

1. výpočet korelační matice **R** (příp. kovarianční K)
2. výpočet vlastních čísel λ_i a vlastních vektorů $\mathbf{u_i}$ této matice **R (K)**
3. uspořádání λ_i sestupně a uložení odpovídajících vektorů $\mathbf{u_i}$ do sloupců matice **U**
4. transformace datové matice X do nové datové matice rovnicí **Z** = **X U**

Příklad 3.3.

V souboru jsou údaje o 104 plavkyních testovaných ve 4 stylech a 2 délkách tratě. Úkolem zjistit, zda „plavecká schopnost" je souhrn více nezávislých vlastností nebo jediná.

Data: plav (50 kraul, 200 kraul, 50 prsa, …,50 znak, … 50 delfín, …)

Výsledek: *Hlavní komponenty*

z1= **0.88** x1+ **0.82** x2 + 0.27 x3 + 0.20 x4 + **0.77** x5 + **0.76** x6 + **0.70** x7+ **0.57**x8
z2= 0.19 x1+0.26 x2 + **0.78** x3 + **0.77** x4 + 0.21 x5 + 0.35 x6 + 0.40 x7+ 0.48x8

Zjednodušené vyjádření hlavních komponent pomocí zaokrouhlení spočítaných složek vlastních vektorů na celá čísla - (zvýrazněných) koeficientů s vyšší absolutní hodnotou na hodnotu 1 a ostatních na hodnotu 0:

z1= x1+ x2 + x5 + x6 + x7+x8
z2= x3 + x4

Nyní můžeme říct, že existují 2 hlavní faktory, ovlivňující plaveckou schopnost, případně je pojmenovat.

Trenér plavkyň pojmenoval první dvě hlavní komponenty:

z1 *...schopnost plavat převážně pomocí paží*
z2 *...schopnost plavat převážně pomocí nohou*

Závěr: *existují 2 typy plavkyň, jedentyp je nadaný pro styl prsa, druhý typ je nadaný pro ostatní styly.*

♦

❑ Využití výsledků hlavních komponent

Výsledky výpočtu hlavních komponent se mohou využít

1. seskupením atributů spojenou s definováním skrytých proměnných,
2. k redukci počtu původních proměnných a tak ke zjednodušení popisu objektů,
3. ke zprostředkovanému měření nepřímo měřitelných proměnných a jejich odhadu,
4. k transformaci původních proměnných do výhodnějšího tvaru, spojené s jejich ortogonalizací.

Příklad 3.4. **[13]**

Na Pedagogické fakultě v Ostravě v 80. (komunistických) letech hodnotili úroveň výchovně vzdělávacího procesu u 32 absolventů - učitelů chemie po 1 roce praxe. Každý učitel byl hodnocen svým patronem, vedením školy a inspektorem bodováním 0 - 4 bodů v následujících 28 ukazatelích:

1. Splnění povinností daných učebním plánem.
2. Vědecká úroveň přednášek.
3. Dodržení kontinuity přednášeného učiva a využití mezipředmětových vztahů.
4. Spojení teorie s praxí v probíraném učivu.
5. Výchovné využití učiva.
6. Tvůrčí schopnosti a využívání nových metod ve výuce.
7. Osvětlování a využívání dialektických souvislostí při upevňování a opakování látky.
8. Názornost ve výuce a používání didaktické techniky.
9. Funkční začlenění chemických pokusů do výuky.
10. Kontakt učitele s žákem a učební klima při výuce.
11. Podnícení zájmu žáků o chemii.
12. Aktivita žáků při vyučování.
13. Smysl pro pořádek a estetiku.
14. Osobnost učitele, jeho vystupování, znalosti, dovednosti, osobní kladné vlastnosti.
15. Veřejně politická angažovanost učitele.
16. Spolupráce učitele s PO SSM.
17. Spolupráce učitele s rodiči.
18. Sebevzdělávání a sebevýchova učitele.
19. Pedagogická dokumentace.
20. Spolupráce učitele s kolektivem školy.
21. Výchovné využití politického vzdělávání.
22. Diferencovaný přístup k žákům.
23. Agitační práce ve škole.
24. Všeobecná iniciativa a smysl pro pokrok.
25. Věcné znalosti.
26. Metodické znalosti.
27. Úroveň společenskopolitické praxe.
28. Zhodnocení fyzických podmínek pro výkon učitelského povolání.

Data není nutno předzpracovávat ani transformovat. Pro hlavní komponenty byly spočítány:

Stopa kovarianční matice = 19.37

Vlastní čísla : $\lambda = \{7.60, 2.34, 0.65, \ldots\}$

Procentuální významnost vlastních čísel: {39.2%, 12.1%, 3.3%, …}

Vlastní vektory :

u1 = (0.02, 0.13, 0.12, 0.10, 0.10, **0.28**, 0.11, 0.11, 0.17, **0.27**, 0.10, **0.25**, 0.15, 0.15, **0.29**, 0.02, -0.03, **0.26**, 0.06, 0.12, 0.12, 0.13, 0.21, **0.29**, 0.18, 0.15, 0.11, -0.02)
u2 = (-0.10, -0.12, -0.14, -0.17, -0.09, -0.12, -0.08, -0.14, -0.09, -0.08, -0.21, -0.14, -0.10, -0.17, **0.63, 0.46**, -0.05, 0.07, 0.05, 0.05, 0.09, 0.08, 0.17, -0.04, 0.01, -0.08, **0.24**, 0.19)
...

*První vlastní vektor má nejvyšší hodnoty u atributů číslo 6, 10, 12, 15 a 24. Jejich společný rys bychom mohli nazvat „míra **zápalu pro učitelství**“. Hodnota prvního vlastního čísla je 39% stopy, tedy první nejvýraznější faktor variability je dosti výrazný.*

*Druhý vlastní vektor má nejvyšší hodnoty s čísly 15, 16 a 27, jejich společný rys můžeme shrnout jako „míra **společenské aktivity**“. Hodnota vlastního čísla je 12 % stopy, tedy druhý faktor je výrazný mnohem méně a znamená větší vyrovnanost absolventů.*

Další vlastní čísla vyšla relativně velmi malá, proto nemá smysl uvažovat o dalších faktorech.

***Závěr**: existují 2 hlavní faktory výše pojmenované, rozlišující mezi sebou učitele; pokud by bylo možnohodnotit přímo jen tyto 2 vlastnosti, stačily by hodnotitelům 2 atributy, pokud není možno přímo najít pravidlo pro jejich hodnocení, stačilo by místo původních 28 otázek položit hodnotitelům jen otázky 6, 10, 12, 15, 16, 24 a 27 (tedy jen 7) a z nich zjednodušenými transformačními rovnicemi spočítat tyto 2 vlastnosti:*

$z1 = 28.x6 + 27.x10 + 25.x12 + 29.x15 + 26.x18 + 29.x24$

$z2 = 63.x15 + 46.x16 + 24.x27$

V obou případech dojde k výraznému zmenšení počtu atributů bez výrazného omezení informace o objektech. Menší počet nezávislých atributů v dalších metodách DM

♦

Příklad 3.5.

Poctivý génius, který lže a krade

Člověk bývá charakterizován množstvím vlastnost. Otázkou je, zda se všechny vlastnosti dají popsat menším množstvím „reprezentativních“ vlastností. Výzkum na toto téma prováděla skupina analytiků po světě. Z původních mnoha atributů, které byly sbírány o lidech:

Člověk (ustaranost, hněvivost, sebevědomí, impulsivnost, zranitelnost, vřelost, společenskost, aktivita, citlivost, důvěřivost, podrobivost, umírněnost, něžnost, poslušnost, cílevědomost, disciplinovanost, uvážlivost, ...)

⇓ *faktorová analýza našla 5 nezávislých komponent:*

Člověk (emoční stálost, otevřenost, extroverze, přátelskost, svědomitost)

emoční stálost = míra odolnosti vůči záporným pocitům
vysoce reaktivní ⇔ kliďas
otevřenost = míra zvědavosti na vnitřní a vnější svět
badatel ⇔ konzervativní „udržovatel“
extroverze = míra udržování aktivních vztahů k jiným lidem
extrovert ⇔ introvert, do sebe zahleděný
přátelskost = měřítko altruismu
altruista ⇔ egocentrista
svědomitost = míra sebekontroly vůle něčeho dosáhnout
soustředěnost na osobní cíle ⇔ pružný, požitkářský

⇓

průzkum mezi lidmi, názory na sebe, známé osobnosti, politiky

Výsledek:

- sebe a známé osobnosti (z kultury, filmu, sportu, ...) hodnotí dotazovaní ve všech 5 rozměrech
- politiky hodnotí pouze ve 2 rozměrech, a to ještě značně závislých:

  1. jak politik dovede přesvědčit okolí o tom, že je poctivý, pravdomluvný, odpovědný, spolehlivý, precizní, vytrvalý, umírněný, velkorysý a citově vyrovnaný

  2. jak politik dovede přesvědčit okolí o tom, že je aktivní, přiměřeně sebejistý, energický, tvořivý, chytrý, moderní, výkonný, vynalézavý a srdečný.

Bez ohledu na skutečné vlastnosti hodnotí jen herecké schopnosti. Politikovi stačí hrát roli + pomluvit protivníka (efekt spáče).

♦

Shrnutí pojmů 3.

Korelace dvou veličin.

Hlavní komponenty a jejich konstrukce.

Reprezentace dat pomocí lineárních kombinací původních veličin.

Využití výsledků hlavních komponent

Otázky 3.

1. Definujte korelaci dvou veličin – atributů.
2. Popište podstatu skrytých faktorů (hlavních komponent) pomocí známých veličin.
3. Popište metodu výpočtu hlavních komponent.
4. K čemu slouží výsledek výpočtu hlavních komponent?

Úlohy k řešení 3.

1. Jsou dána data jednoho ostravského fitcentra. V něm nabízejí jako službu svým zákazníkům tzv. „zjištění celkové kondice těla“. K výsledku je zapotřebí znát některé vlastnosti daných osob, proto si fitcentrum vytvořilo dotazník pro jejich zjištění, další data naměřili a vypočetli.

   U jednotlivých osob byly sledovány tyto vlastnosti:

   Základní údaje

   - id – identifikace osoby
   - věk – celé číslo udávající počet roků
   - pohlaví – muž/žena
   - klidová TF – tepová frekvence/minutu
   - krevní tlak – nízký/normální/vysoký

   Tělesné proporce

- výška – v centimetrech
- váha – v kilogramech
- obvod krku – v centimetrech
- obvod bicepsu – v centimetrech
- obvod hrudi – v centimetrech
- obvod pasu – v centimetrech
- obvod boku – v centimetrech
- obvod hýždí – v centimetrech
- obvod stehna – v centimetrech
- obvod lýtka – v centimetrech

Aerobní zdatnost

- ergometr - watt/kilogram
- VO_2MAX - mililitr kyslíku/kilogram/minuta
- step test - tepová frekvence/minuta

Test síly

- sedy-lehy – počet sedů lehů za 1 minutu
- kliky – počet kliků lehů za 1 minutu
- benchPress1 – v kilogramech
- benchPress2 – v kilogramech
- legPress – v kilogramech

Antropologické měření

- PT – procento tuku v těle - v procentech
- A.T.H – hmotnost bez tuku – v kilogramech
- H.T – hmotnost tuku – v kilogramech
- BMI – Body Mass Index = index tělesné zdatnosti (výška [cm]/hmotnost [kg]) -
- WHR – Waist to Hip Ratio = Poměr pasu a boků (obvod pasu/obvod boků)

Pro tato data navrhněte použití metody hlavních komponent.

Semestrální projekt Analýza 3

Pro svá data poveďte analýzu použitelnosti metody hlavních komponent. Pokud je metoda použitelná, navrhněte její využití. Pokud není, zdůvodněte to.