Juhaňák, Libor

Metody a techniky analýzy dat

In: Juhaňák, Libor. *Analytika učení a data mining ve vzdělávání v kontextu systémů pro řízení výuky.* Vydání první Brno: Masarykova univerzita, 2023, pp. 49-58

ISBN 978-80-280-0184-1; ISBN 978-80-280-0185-8 (online ; pdf)

Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.77693
Access Date: 22. 02. 2024
Version: 20230228

5 METODY A TECHNIKY ANALÝZY DAT

Podat systematický a zároveň dostatečně podrobný přehled metod používaných v oblasti data miningu ve vzdělávání či analytiky učení je poměrně komplikovaný úkol. Jednak proto, že se výčty používaných metod mezi jednotlivými autory více či méně liší, ale i proto, že různí autoři seskupují metody do obecnějších kategorií odlišným způsobem. Zatímco například Baker a Inventado (2014) nabízejí čtyři základní kategorie data miningových metod ve vzdělávání, Bakhshinategh, Zaiane, ElAtia a Ipperciel (2018) nabízí výčet devíti základních metod, které neseskupují do žádných souvisejících kategorií. Níže proto představuji vlastní kategorizaci a výčet základních metod a technik analýzy dat z oblasti data miningu ve vzdělávání a analytiky učení. Vycházím přitom z kategorizace Bakera a Inventada (2014), kterou však upravuji a doplňuji s ohledem na další důležité zdroje v této oblasti (především Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018; Papamitsiou & Economides, 2014; Romero & Ventura, 2010, 2013), jakož i s ohledem na zdroje týkající se data miningu obecně (především Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Gorunescu, 2011; Witten, Frank, & Hall, 2011).

5.1 Prediktivní metody

První skupinu metod lze označit jako prediktivní metody. Cílem predikce a prediktivních metod je vytvoření modelu, který je schopen odvodit či odhadnout jeden konkrétní atribut či aspekt dat (tj. predikovanou, výstupní či závisle proměnnou) na základě určité kombinace dalších aspektů dat – tj. prediktorů, resp. vysvětlujících či nezávisle proměnných (Baker & Inventado, 2014; Romero & Ventura, 2013). Prediktivní metody vyžadují, abychom předem měli určitou množinu dat,

kde známe správnou hodnotu predikované proměnné. Pokud bychom využili terminologie ze související oblasti strojového učení (*machine learning*), pak lze kategorii prediktivních metod vesměs ztotožnit s tím, co se v rámci strojového učení označuje jako metody učení s učitelem (*supervised learning*). Ty se od metod učení bez učitele (*unsupervised learning*) liší právě tím, že máme k dispozici tzv. trénovací data s předem známou hodnotou predikované proměnné. Na těchto trénovacích datech se pak použitý algoritmus může „naučit" správnou souvislost mezi prediktory a predikovanou proměnnou. Tuto naučenou „znalost" (v podobě určitého modelu) lze pak použít na nových datech, kde již hodnotu predikované proměnné neznáme.

Obecně patří prediktivní metody mezi jedny z nejpoužívanějších. V kontextu data miningu ve vzdělávání se pak používají např. k predikci úspěšnosti studentů, k detekci různých forem chování studentů v online prostředí, k predikci úrovně znalostí studentů či k odhadování afektivních stavů studentů při plnění určitých studijních úkolů. Mezi dva základní typy metod v této kategorii pak patří metody klasifikace a regrese.

V případě regrese (*regression*) má predikovaná proměnná obvykle podobu číselné, resp. spojité proměnné. Mezi prediktory pak mohou být jak číselné, tak i kategorické proměnné. Cílem regrese je na základě dostupných prediktorů odhadnout konkrétní hodnotu (číslo) výstupní proměnné. Mezi nejčastěji používané metody patří klasická lineární regresní analýza, pro účely predikce spojité proměnné však lze použít například i regresní stromy (*regression trees*), podpůrné vektory (*support vector machines*) či umělé neuronové sítě (*neural networks*). V kontextu data miningu ve vzdělávání je však použití těchto „pokročilejších" metod výrazně méně časté než v jiných oblastech využívajících obecné data miningové metody. Dle Bakera a Inventada (2014) je hlavním důvodem to, že v oblasti vzdělávání hraje obvykle důležitou roli velké množství faktorů (tj. vysvětlujících proměnných) a zároveň data ze vzdělávacího kontextu často obsahují velké množství šumu (tj. chybějících dat, nepravidelností, chyb v datech apod.). V souvislosti s regresní analýzou je třeba upozornit, že ačkoli např. klasická lineární regresní analýza je běžně využívaná i v kontextu statistického testování hypotéz standardně aplikovaného i v jiných oblastech výzkumu, způsob využití regresní analýzy v data miningu je mírně odlišný (např. co se týče validace modelu).

V případě klasifikace (*classification*) má naopak predikovaná proměnná kategorický charakter a cílem klasifikace je tak správné zařazení případu do jedné ze dvou či více možných kategorií. Mezi nejčastěji používané metody klasifikace v oblasti data miningu ve vzdělávání přitom patří logistická regrese (*logistic regression*), rozhodovací stromy (*decision trees*) a náhodný les (*random forest*). Pro účely klasifikace lze však využít i řadu jiných technik, resp. klasifikačních algoritmů, jako jsou například již zmíněné podpůrné vektory a umělé neuronové sítě, ale také algoritmus k-nejbližších sousedů (*k-nearest neighbor algorithm*), naivní Bayesův

klasifikátor (*naive Bayes*) či diskriminační analýzu (*discriminant analysis*). Jako specifický typ metody klasifikace v oblasti vzdělávání uvádějí Baker a Inventado (2014) odhad latentních znalostí (*latent knowledge estimation*), což je metoda využívaná především v kontextu inteligentních tutorských systémů a doporučovacích systémů. Dodejme, že podle Peña-Ayaly (2014b) patří metody klasifikace v kontextu data miningu ve vzdělávání mezi zdaleka nejčastěji používané (cca 42 % ze všech analyzovaných studií).

5.2 Shlukování

Za druhou skupinu metod lze považovat tzv. shlukování či klastrování (*clustering*, *cluster analysis*). Základním cílem shlukování je seskupení jednotlivých případů do určitého počtu skupin (shluků) na základě jejich podobnosti. Co znamená ona „podobnost", je pak samozřejmě různé dle analyzovaných dat a použitých shlukovacích algoritmů. Obecně lze však podobnost chápat tak, že případy v rámci určitého shluku mají napříč jednotlivými proměnnými „blízké" hodnoty, a naopak případy z různých shluků mají hodnoty v jednotlivých proměnných spíše odlišné (Dutt, Ismail, & Herawan, 2017). Shlukování lze přitom považovat za jednu ze základních metod učení bez učitele (*unsupervised learning*), tzn. takovou metodu, kdy se pracuje s daty, u nichž není předem známo jejich zařazení do odpovídajících kategorií (oproti klasifikaci) a nelze tedy použít trénovací data, na kterých by se algoritmus mohl „naučit" správné zařazení případu do odpovídající kategorie.

Lze přitom souhlasit s Hebákem et al. (2013), že pojem shlukování či shluková analýza ve skutečnosti zahrnuje větší množství poměrně různorodých specifických metod a přístupů. Obecně lze metody shlukování rozdělit na hierarchické a nehierarchické. Hierarchické metody shlukování spočívají v tom, že je vytvářena hierarchická posloupnost postupného spojování, resp. rozdělování jednotlivých případů v datech, kterou lze znázornit pomocí tzv. dendrogramu (tj. diagramu znázorňujícího onou stromovou či hierarchickou strukturu). Jednotlivé techniky hierarchického shlukování lze pak dále rozdělit na aglomerativní a divizivní. Aglomerativní algoritmy hierarchického shlukování postupují od jednotlivých případů, které postupně spojují do stále menšího počtu shluků tak, že nakonec vznikne jediný shluk tvořený všemi případy. Samotné spojování do shluků se provádí na základě vypočtené vzdálenosti, resp. blízkosti jednotlivých případů a shluků. Divizivní algoritmy pak postupují opačně, tzn. místo spojování jednotlivých případů a následně shluků fungují tak, že jeden výchozí shluk (tj. kompletní data) postupně rozdělují na menší a menší shluky, dokud se nedopracují ke stavu, kdy každý z případů v datech tvoří svůj vlastní shluk. Za častěji využívané metody lze však považovat spíše metody aglomerativního hierarchického shlukování.

Druhým základním typem shlukovacích metod jsou metody nehierarchického shlukování. Jak napovídá samotné označení těchto metod, základním rozdílem oproti předchozí kategorii je to, že nepracují s hierarchickou posloupností spojování či rozdělování shluků. V rámci nehierarchických metod lze dále rozlišovat různé dílčí typy shlukovacích technik. Za jedno ze základních rozdělení můžeme považovat to, zda jde o disjunktivní či kategorické shlukování, kdy může být každý případ zařazen vždy jen do jednoho shluku, anebo jde o tzv. fuzzy či probabilistické shlukování, kdy je pro každý případ vypočtena míra jeho příslušnosti do každého z existujících shluků (srov. Witten, Frank, & Hall, 2011). Za nejčastěji využívané metody nehierarchického shlukování lze považovat algoritmy k-shlukování, a zvláště metodu k-průměrů (*k-means clustering*). Oproti hierarchickým metodám je u k-shlukování předem stanoven počet shluků, do kterých se mají data rozdělit. Samotný algoritmus pak funguje iterativně, kdy se opakovaně vypočítává tzv. centroid[26] jednotlivých shluků a následně se upravuje zařazení jednotlivých případů do shluků na základě jejich nejbližšího centroidu. Metoda k-průměrů patří v oblasti data miningu ve vzdělávání mezi celkově nejčastěji používané metody shlukování (viz Peña-Ayala, 2014b). Jak ale upozorňuje například Hebák et al. (2013), problémem k-shlukování je to, že dosahuje pouze lokálně optimálního řešení, které je navíc závislé na výchozí pozici centroidů, resp. na pořadí případů v datech.

5.3 Metody redukce dimenzí

Třetí skupinou metod jsou metody redukce dimenzí či metody redukce dimenzionality. Tyto metody lze v určitém pohledu považovat za podobné metodám shlukování. Např. Baker a Inventado (2014) dokonce obě tyto skupiny metod zahrnují pod jednu společnou kategorii metod zabývajících se odhalováním struktury (*structure discovery*) v analyzovaných datech. Přesto považuji za vhodnější rozdělit obě skupiny metod do dvou různých kategorií, jelikož je mezi nimi jeden zásadní rozdíl. Ten lze v principu vysvětlit následovně. Zatímco v případě shlukování nám jde o jednotlivé případy a jejich seskupení do skupin dle podobnosti, v případě redukce dimenzionality se zaměřujeme na proměnné a možnosti redukce jejich počtu (ať již pomocí sloučení několika proměnných do jedné, anebo výběrem jen těch nejpřínosnějších proměnných). Metody redukce dimenzí se hojně používají i mimo oblast data miningu či strojového učení, ačkoli se obvykle v různých oblastech poněkud liší používaná terminologie a zároveň jsou tyto metody používány pro mírně odlišné účely.

V kontextu psychometriky, ale i obecně pedagogického výzkumu využívajícího dotazníková šetření, je relativně běžně využívanou metodou spadající do oblasti

26 Pro jednodušší pochopení si jej lze představit jako střed či těžiště daného shluku.

redukce dimenzionality tzv. faktorová analýza (*factor analysis*). Cílem faktorové analýzy je redukce většího množství naměřených proměnných na menší množství latentních (přímo neměřených) proměnných, označovaných jako faktory. Řada disciplín v sociálních vědách totiž pracuje s poměrně komplexními konstrukty, které lze jen obtížně (pokud vůbec) měřit přímo. Obvykle se tak pomocí určitého množství jednodušších přímo měřených položek zjišťuje (resp. odhaduje) hodnota nějakého složitějšího přímo neměřitelného konstruktu. Tím může být např. určitá schopnost, motivace, postoj apod. Faktorová analýza pak v tomto kontextu obvykle slouží k ověření toho, zda použitý měřicí nástroj skutečně měří ty konstrukty, které by měl dle předpokladu měřit. Obecně lze přitom rozlišit konfirmační faktorovu analýzu (*confirmatory factor analysis – CFA*), která předem předpokládá určitou strukturu faktorů a jednotlivých položek, a explorační faktorovu analýzu (*exploratory factor analysis – EFA*), u které není předem dán počet faktorů a jejich vztah k jednotlivým položkám.

V kontextu data miningu ve vzdělávání je častěji používanou metodou analýza hlavních komponent (*principal component analysis*), jež je však svým charakterem velmi podobná explorační faktorové analýze. Stejně jako v případě faktorové analýzy je cílem analýzy hlavních komponent redukce výchozího počtu proměnných na menší počet „umělých" či skrytých proměnných, které by však obsahovaly co možná největší míru informace obsaženou v původní sadě proměnných. Určitý rozdíl mezi oběma metodami je pak jednak v konkrétním způsobu výpočtu[27], jednak částečně také v cílech obou metod a v interpretaci jednotlivých komponent, resp. faktorů. V případě metody hlavních komponent je obvykle hlavním cílem redukce původně velkého množství proměnných pro účely zjednodušení následných analýz. Primárním požadavkem je tak to, aby nové proměnné (komponenty) co nejvíce vysvětlovaly variabilitu původních proměnných, přičemž ne vždy je nutné, aby jednotlivé komponenty zároveň měly nějakou věcnou interpretaci. Naopak v rámci faktorové analýzy je věcná interpretace jednotlivých faktorů obvykle zcela zásadní, jelikož jednotlivé faktory mají odpovídat teoretickým konstruktům, kterým je v daném výzkumu věnována pozornost. Podobně je při využívání faktorové analýzy obvykle požadováno, aby vztahy mezi jednotlivými faktory (jakož i původními položkami) odpovídaly výchozím teoretickým předpokladům.

Mezi další metody redukce dimenzí, jež jsou používané v oblasti data miningu ve vzdělávání, lze zařadit například lineární diskriminační analýzu (*linear discriminant analysis*) či metodu označovanou jako non-negativní faktorizace matic (*non-negative matrix factorization*), viz Romero a Ventura (2013).

27 V případě analýzy hlavních komponent se vychází z kovarianční matice původních proměnných, zatímco u faktorové analýzy se pracuje spíše s korelační maticí (viz Hebák et al., 2013).

5.4 Dolování vztahů

Další skupinu tvoří metody zaměřující se na zkoumání vztahů mezi jednotlivými proměnnými. Baker a Inventado (2014) pro označení těchto metod používají zastřešující termín dolování vztahů (*relationship mining*). Základním cílem dolování vztahů je nalezení důležitých souvislostí mezi proměnnými v datech. A to obvykle v situacích, kdy data obsahují velmi velké množství proměnných. Za jednu z nejjednodušších technik lze v této oblasti považovat klasickou korelační analýzu běžně používanou i v tradičně orientovaném pedagogickém výzkumu. Pro tu se v kontextu data miningu někdy používá označení dolování korelací (*correlation mining*). Mimo jednoduché hledání významných korelací se však lze v oblasti dolování vztahů setkat s řadou dalších metod. Níže zmiňuji tři vybrané příklady, které patří v data miningu ve vzdělávání a v analytice učení k častěji používaným.

Asi nejčastěji využívanou metodou dolování vztahů jsou tzv. asociační pravidla, resp. dolování asociačních pravidel (*association rule mining*). Cílem této metody je nalézt v datech sadu často se vyskytujících pravidel v podobě jestliže-pak (*if-then*). To lze přiblížit tak, že se automaticky hledá společný výskyt určitých hodnot v určitých proměnných napříč jednotlivými případy. Pokud bychom chtěli dát příklad z kontextu využívání LMS, pak si lze takové asociační pravidlo představit např. v následující podobě:

- *JESTLIŽE student otevřel požadovaný studijní materiál A ZÁROVEŇ se studijním materiálem intenzivně pracoval, PAK student úspěšně dokončil průběžný test.*

Takové pravidlo přitom neukazuje automaticky na kauzální vztah, ale pouze na to, že v datech existuje mnoho případů (studentů), u kterých existuje společný výskyt daných hodnot všech tří proměnných zmíněných ve výše naznačeném asociačním pravidle. Dodejme, že ačkoli se metoda asociačních pravidel výrazně rozšířila zvláště po roce 1993 v souvislosti s prací autorů Agrawal, Imieliński a Swami (1993), již v roce 1966 přišli s návrhem obdobné metody automatického získávání pravidel či vztahů čeští autoři Hájek, Havel a Chytil (1966) a tato jejich metoda označovaná zkratkou GUHA (*General Unary Hypotheses Automaton*) je dodnes v oblasti obecného data miningu využívána a rozvíjena.

Tradičním způsobem využití asociačních pravidel a zároveň první oblastí, v níž byla asociační pravidla využita, je tzv. analýza nákupního košíku (*market basket analysis*). Tu si lze jednoduše představit jako tabulku nákupů obsahující informaci o tom, jaké produkty byly zakoupeny společně v rámci jednoho nákupu. Pomocí metody asociačních pravidel lze pak získat informaci o tom, jaké produkty jsou obvykle nakupovány společně, což následně prodejce využívá např. pro automatické doporučování určitých produktů v průběhu nákupu (v e-shopech), případně pro umisťování společně nakupovaných produktů do vzájemné blízkosti (v tradičních kamenných obchodech). Jedním z klasických algoritmů používaných pro dolování asociačních pravidel je pak algoritmus *Apriori* (viz Agrawal, Imieliński, & Swami,

1993). V současnosti je však již využívána i řada dalších algoritmů. V kontextu data miningu ve vzdělávání se obdobný přístup využívá např. v doporučovacích systémech (ať už při doporučování kurzů, či konkrétních studijních materiálů), ale také při řešení specifických problémů, jako je např. identifikace testových úloh, u kterých nastává společný výskyt chybných odpovědí studentů (srov. Baker & Inventado, 2014; Romero & Ventura, 2007).

Do značné míry blízkou metodou k asociačním pravidlům je metoda dolování sekvencí (*sequence mining* či *sequential pattern mining*). I v případě této metody je cílem odhalování důležitých asociací či vztahů. Navíc je však zohledňováno pořadí či časová následnost mezi jednotlivými prvky. Zatímco tedy v případě asociačních pravidel a analýzy nákupního košíku sehrával roli pouze společný výskyt (tj. např. které produkty jsou společně v nákupním košíku), v případě dolování sekvencí je důležité i to, jakým způsobem za sebou prvky následují. Jednou z oblastí využití této metody je např. bioinformatika, kde se dolování sekvencí využívá při práci s daty obsahujícími sekvence písmen A, G, C a T tvořících DNA. V oblasti data miningu ve vzdělávání lze dolování sekvencí využít zvláště v situacích, kdy nás zajímá způsob průchodu studenta kurzem či způsoby chování studentů v rámci nějaké výukové aktivity.

Vedle dolování sekvencí, které je využíváno jednak v data miningu obecně, jednak v rámci specifických disciplín (např. již zmiňovaná bioinformatika), je v kontextu data miningu ve vzdělávání a analytiky učení pro účely analýzy posloupností či sekvencí aktivit uplatňován ještě i jiný metodologický přístup. Jde o tzv. dolování procesů (*process mining*), které vychází primárně z oblasti managementu a řízení podnikových procesů, kde se zaměřuje na analýzu procesních dat z různých typů podnikových a manažerských systémů (srov. van der Aalst, 2011, 2016). Dolování procesů je však uvedeno i mezi základními analytickými technikami data miningu ve vzdělávání ve stěžejní publikaci *Handbook of Educational Data Mining* (Romero, Ventura, Pechenizkiy, & Baker, 2010). Zároveň v této oblasti existuje několik studií, jež dolování procesů využívají (např. Schoor & Bannert, 2012; Romero, Cerezo, Bogarín, & Sánchez-Santillán, 2016; Papamitsiou & Economides, 2016; Juhaňák, Zounek, & Rohlíková, 2019). Základním úkolem dolování procesů je získávání užitečných informací z dat, která mají procesní charakter (tj. ve své podstatě zachycují nějaký proces). V kontextu systémů typu LMS se přitom ovykle jedná o tzv. logy.

5.5 Analýza sociálních sítí

Do výše nastíněné kategorie dolování vztahů by z určitého hlediska bylo možné zařadit i metodu analýzy sociálních sítí (*social network analysis*), jejíž stěžejní součástí je právě zkoumání vztahů mezi jednotlivými aktéry či prvky v rámci dané sítě. Přesto se jeví jako vhodnější hovořit o samostatné kategorii, a to hned z několika

důvodů. Předně analýzu sociálních sítí lze v současné době jen obtížně považovat za jednu dílčí konkrétní metodu. Mnohem spíše lze souhlasit s Wassermanem a Faustovou (1994), že jde o obecnější výzkumnou perspektivu či výzkumný přístup, který zahrnuje relativně širokou škálu metod zaměřujících se na zkoumání nejrůznějších charakteristik (sociálních) sítí. V současnosti je zároveň analýza sociálních sítí považována za mezioborovou oblast výzkumu, jež zahrnuje výzkumníky z široké škály disciplín, jako je sociologie, antropologie, informatika, pedagogika, kriminologie, ekonomie, medicína, epidemiologie, politologie a řada dalších (Scott & Carrington, 2011). Navíc historické kořeny analýzy sociálních sítí lze nalézt spíše v sociologii a teorii grafů (součást matematiky) než v oblasti data miningu. Ve zdrojích týkajících se metod data miningu tak není analýza sociálních sítí příliš často uváděna. Podobně pak analýzu sociálních sítí v souvislosti s data miningem ve vzdělávání neuvádí mezi základními metodami například ani Baker a Inventado (2014), přičemž vysvětlují, že z jejich pohledu je analýza sociálních sítí využívána spíše v rámci analytiky učení než v oblasti data miningu ve vzdělávání. Na druhou stranu řada dalších autorů a výzkumníků v této oblasti považuje analýzu sociálních sítí za součást metodologické výbavy i v oblasti data miningu ve vzdělávání (viz např. Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018; Papamitsiou & Economides, 2014; Peña-Ayala, 2014a; Romero & Ventura, 2010).

Analýzu sociálních sítí lze v určitém ohledu chápat jako podobnou dolování procesů či dolování sekvencí. A to v tom smyslu, že oproti ostatním výše zmiňovaným metodám se metody jako analýza sociálních sítí a dolování procesů zaměřují na analýzu specifických typů dat. A zatímco dolování procesů se zaměřuje na data zachycující určitý proces, analýza sociálních sítí je schopna analyzovat data, která mají charakter sítě – tj. zachycují určité vztahy mezi určitými prvky. Zároveň je třeba dodat, že tyto metody obvykle nejsou schopny analyzovat data tradičního typu. Za „tradiční" data lze přitom považovat tabulku tvořenou sloupci v podobě proměnných a řádky reprezentujícími jednotlivé případy, tj. taková data, která jsou běžně analyzována různými metodami jako regrese, klasifikace, shlukování, redukce dimenzí apod. Metody analýzy sociálních sítí či dolování procesů je tak třeba chápat (oproti jiným zmiňovaným metodám) vždy jen jako metody zaměřující se na analýzu specifického typu dat.

Jedním z relativně častých způsobů využití analýzy sociálních sítí v kontextu vzdělávání a online systémů typu LMS je analýza komunikace v online diskuzních fórech (viz např. Hernández-García, González-González, Jiménez-Zarco, & Chaparro-Peláez, 2015; Rabbany, ElAtia, Takaffoli, & Zaïane, 2014; Wise & Cui, 2018). Najdou se ale i jiné způsoby využití. Např. Saqr, Fors a Nouri (2018) využívají analýzu sociálních sítí v kontextu predikce úspěšnosti studentů. Analýza sociálních sítí se rovněž využívá v kontextu skupinové spolupráce a skupinového učení, kde může být příkladem studie autorů Xie, Di Tosto, Lu & Cho (2018), kteří se zaměřili na mapování sociální dynamiky a identifikaci lídrů v rámci skupinového

učení. Přitom použili kombinaci analýzy sociálních sítí a metody dolování textů (viz následující kapitolu). Nutno však podotknout, že analýza sociálních sítí bývá mnohem častěji využívána v kontextu analytiky učení spíše než v kontextu data miningu ve vzdělávání.

5.6 Dolování textů a zpracování přirozeného jazyka

Dolování textů (*text mining*)[28] lze považovat za další velkou skupinu metod nejen data miningu ve vzdělávání, ale i analytiky učení. Obecným cílem dolování textu je automatizovaná extrakce relevantních informací z textových dokumentů. Bousbia a Belamri (2014) považují text mining za rozšíření oblasti data miningu o takové metody a techniky, které jsou schopny pracovat s daty v podobě textu. K rozvoji těchto metod pak přispívá nejen oblast data miningu jako takového (příp. oblast strojového učení), ale také např. počítačová lingvistika či oblast označovaná jako zpracování přirozeného jazyka (*Natural Language Processing – NLP*).

Data je v základu možné rozdělit na strukturovaná a nestrukturovaná. Zatímco strukturovaná data mají stanoven určitý systém, na jehož základě jsou ukládána, a tudíž mají ve výsledku jasnou a jednotnou strukturu (typicky v podobě tabulky, resp. databáze), nestrukturovaná data takový pevný systém ukládání nemají, tudíž ve výsledku nemají jednotnou strukturu. Typickým příkladem nestrukturovaných dat mohou být nejrůznější textové dokumenty v přirozeném jazyce (např. e-maily, příspěvky lidí na sociálních sítích, literární díla apod.), ale spadají sem také různé vizuální dokumenty (obrázky, fotografie apod.) či zvukové a audiovizuální dokumenty. Dolování textu a zpracování přirozeného jazyka se pak zaměřuje specificky na zpracování a analýzu těchto nestrukturovaných dat v podobě textu v přirozeném jazyce[29]. Tím se zároveň oblast text miningu odlišuje od obecného data miningu a všech výše zmiňovaných metod, jelikož ty se zaměřují na analýzu strukturovaných dat.

Konkrétní metody a techniky text miningu lze seskupovat do obecnějších skupin podle toho, k čemu slouží, resp. k řešení jakého úkolu se obvykle využívají. Mezi běžné úkoly v rámci text miningu patří například:

- kategorizace textů podle tématu či druhu textu,
- shlukování textů, při kterém je cílem identifikovat podobné texty,

28 Někdy též „text data mining“ či „text analytics“, viz Romero a Ventura (2013).

29 To však nevylučuje kombinaci text miningových metod s dalšími uvedenými metodami. Naopak velmi častým způsobem využití metod dolování textu v kontextu konkrétních studií a výzkumů je v první fázi aplikace metod pro zpracování přirozeného jazyka tak, aby z nestrukturovaných dat vznikla data strukturovaná, a v druhé fázi pak aplikace některé z výše popsaných metod pro práci se strukturovanými daty.

- extrakce jmenných entit, kdy jsou z textů automaticky extrahována jména lidí, organizací, míst, časových údajů apod.,
- analýza sentimentu, zaměřující se na názory či pocity vyjadřované v rámci analyzovaných textů, či
- sumarizace textů, v jejímž rámci dochází k automatické tvorbě krátkých textových shrnutí původních analyzovaných dokumentů (Bousbia & Belamri, 2014; Lang, Siemens, Wise, & Gašević, 2017; Romero & Ventura, 2013).

V kontextu data miningu ve vzdělávání se text mining používá především v souvislosti s analýzou obsahu diskuzních fór či chatů, analýzou textových prací či odpovědí studentů (obvykle za účelem automatického hodnocení či poskytování zpětné vazby) nebo analýzou obsahu studijních materiálů a zdrojů (obvykle za účelem doporučování vhodného studijního obsahu studentům, srov. Lang, Siemens, Wise, & Gašević, 2017).

5.7 Objevování pomocí modelů

Zatímco výše uvedené metody a techniky jsou rozšířeny i mimo oblast data miningu ve vzdělávání, metodu označovanou jako objevování pomocí modelů (*discovery with models*) lze považovat za do značné míry specifickou právě pro data mining ve vzdělávacím kontextu, jelikož v jiných oblastech aplikace data miningových metod se příliš neuplatňuje (viz Baker & Yacef, 2009; Baker & Inventado, 2014). V základu lze postup objevování pomocí modelů rozdělit do dvou fází. V první fázi je pomocí určitých data miningových metod vytvořen model nějakého fenoménu či konstruktu. V druhé fázi je pak tento model využit jakožto součást jiné analýzy na jiných datech. Jedním z poměrně častých způsobů využití tohoto metodologického přístupu je situace, kdy je nejprve vytvořen úvodní model a následně jsou predikované hodnoty tohoto výchozího modelu využity jakožto prediktory v nějakém dalším modelu v rámci navazující analýzy. Jiným příkladem je využití počátečního modelu k detekci různých typů studentů či jejich chování v online prostředí, přičemž v následné analýze je věnována pozornost rozdílům mezi dříve identifikovanými skupinami studentů v rámci studovaného problému (např. motivace, úspěšnost apod.). Za do značné míry ukázkový příklad využití této metody lze považovat studii Hershkovitze et al. (2013). Relativně často je tento přístup využíván při modelování chování studentů označovaného jako *gaming the system*[30] (viz např. Baker & Gowda, 2010).

30 Do češtiny by bylo možné termín přeložit jako „obcházení systému“, tj. situaci, kdy student využívá možností systému pro to, aby si např. ulehčil práci, obešel stanovená pravidla, splnil stanovenou aktivitu pouze formálně bez skutečného zapojení apod.