METODY ANALÝZY VELKÝCH DAT

Jan Hendl

Anotace

Analýza velkých dat souvisí s ukládáním dat do cloudových úložišť a snahou je využít. Velká data se charakterizují třemi vlastnostmi 3V: objem (volume), různorodost (variety), rychlost změn (velocity). Ukládat a organizovat se mohou před zpracováním v platfomách jako Hadoop a Mapreduce. Analýza těchto dat je procesem zkoumání s cílem získat vhodné znalosti, jejichž typ je vázán na aplikační oblast. Uvažujeme analýzu textů, zvuků a obrazů a kvantitativních strukturovaných údajů. Popisujeme několik tříd metod pro analýzu velkých dat. Věnujeme se dvěma softwarovým systémům pro dolování znalostí a analýzu velkých dat (Weka, Rapidminer).

Klíčová slova

big data, analýza dat, znalosti

1 Úvod

Velká data (big data) mohou být strukturovaná, semistrukturovaná a nestrukturovaná, přičemž se jedná o velká množství v rozmezí peta- až exa-bytů dat. Taková data jsou často vhodná pro dolování s cílem získat informace. Velká data se charakterizují třemi vlastnostmi 3V: objem volume), různorodost (variety), rychlost změn (velocity). Ukládat a organizovat se mohou před zpracováním v platfomách jako Hadoop a Mapreduce. Analýza těchto dat je procesem zkoumání s cílem získat nějaké vhodné znalosti, jejichž typ je vázán na aplikační oblast. Uvažujeme analýzu textů, zvuků a obrazů a kvantitativních strukturovaných údajů. Některé z nich mají vztah k umělé inteligenci.

Význam procedur pro analýzu velkých dat vzrostl v poslední době s pokrokem informačních technologií a prostředků pro sběr a ukládání dat [1, 4, 6, 9, 11]. Výsledkem je situace, že se osamostatnila celá oblast metod, kterým se říká datové analytiky. Při analýze jde o strukturovací proces, v kterém se snažíme zjistit nápadné konfigurace, korelace a trendy. Tato oblast představuje dnes důležitou část informačních technologií ve zdravotnictví [3, 7, 8].

2 Oblasti analýzy velkých dat

Mezi obecné situace, kdy je zapotřebí uplatnit metody získávání znalosti z velkých dat, patří:

Hledání pravidel asociace
Jedná se o oblast vyhledávání klasifikačních pravidel. Zahrnuje analýzu a vyhledání vztahů mezi daty. Výsledkem je kategorizace dat, které mají určité vlastnosti společné. Využívá se v různých oblastech života. Využívá se při kategorizaci textů. Význam webových stránek často závisí na počtu jejich čtenářů. V této oblasti se zjišťují počty uživatelů stránky a význam stránky.

Genetické algoritmy
Genetické algoritmy se používají pro identifikaci nejčastěji sledovaných videí, televizních pořadů a jiných forem médií. Pomocí genetických algoritmů se odhaluje evoluční konfigurace.

Strojové učení
Jde o oblast, kde jde o kategorizaci a určení pravděpodobných výsledků na základě specifické množiny dat. Používá se v prediktivní analytice. Příkladem je získání nějakého právního sporu nebo úspěch s nějakým výrobkem.

Analýza sociálních sítí
Sociální média patří mezi nejdůležitější komunikační média. Jde o identifikaci jedinců s určitými vlastnostmi danými jeho interakčními postupy. Pomocí ní se charakterizují vlastnosti vztahů mezi členy skupiny.

Postupy obecně klasifikujeme do dvou kategorií: učení s učitelem (supervised learning) a učení bez učitele (unsupervised learning).

- **Učení s učitelem**: predikce známé proměnné (kategoriální, spojité). Hledáme funkci, která nejlépe predikuje cílovou proměnnou.
- **Učení bez učitele**: nemáme k dispozici cílovou proměnnou, chceme porozumět přirozeně daným strukturám.

Mezi těmito extrémy se nalézají další metody.

3 Učení z dat bez učitele

Shlukování pomocí k – průměrů
Tento algoritmus slouží pro rozdělení datové množiny podle jejich podobnosti. Nedefinujeme shluky předem, ale postupně je vytvoříme. Cílem je vytvoření shluků, kdy uvnitř jsou podobné a mezi shluky jsou nepodobné. Potřebují nějakou metriku (euklidovskou, Minkowského, Manhatten). Většina potřebuje definovat počet shluků. Hodnota k je počet vzniklých shluků.

Algoritmus:

1. Vyber k.
2. Vyber centroidy.
3. Přiřaď ke každému centroidu bod metodou nejmenší vzdálenosti.
4. Přepočítej centroidy.
5. Opakuj 3. a 4. dokud nenastane konvergence.

Hierarchické shlukování (ze zdola nahoru)

Vytváří se dendrogram, který popisuje postupné shlukování od objektů po větší shluky.
Algoritmus:

1. Zjisti matici vzdáleností.
2. Urči sloučení prvků.
3. Přepočítej matici vzdáleností.
4. Přejdi k 2.

Další metody z oblasti analýzy dat bez učitele

Spektrální shlukování vytvoří i nekonvexní shluky. Pomocí grafů hrany mezi uzly mají malou váhu a uvnitř shluku mají velkou váhu. Pracujeme **s vlastními vektory**.

Shlukování pomocí hlavních komponent slouží k redukci dimenzionality.

Metoda zjišťování **asociačních pravidel** (metoda GUHA). Vymezení je následující. Asociační pravidlo X1, X2, X3 => Y s podporou S a spolehlivostí C. S je počet případů, kde nastanou předpoklady a C je relativní počet, kdy to nastane. Ještě závisí na P(x, y)/P(x)P(y)

4 Učení s učitelem

Rozlišujeme často šest metod učení s učitelem

- Lineární a logistická regrese
- Vektorové stroje pro klasifikace (SVM)
- Rozhodovací stromy
- Umělé neuronové sítě
- Naivní Bayesova metoda

Lineární regrese

Nejdříve technika vyvinutá ve statistice ke zkoumání vztahu mezi výstupní proměnnou a vstupními proměnnými.

Rovnice v jednoduché lineární regresi má tvar

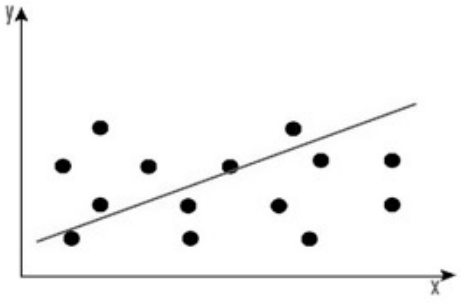
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Pokud máme pouze jednu nezávisle proměnnou, pak má tvar

$$y = \beta_0 + \beta_1 x_1$$

To je ukázáno na grafu. Hodí se pro predikci jedné závislé proměnné v závislosti na několika nezávisle proměnných (viz obrázek 1).

Logistická regrese



Obrázek 1 – Lineární regrese

V logistické regresi predikujeme pravděpodobnost nějakého jevu pomocí několika nezávislých proměnných.

Logistická regrese je matematicky reprezentována rovnicí:

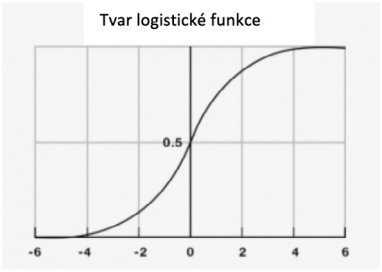
V logistické regresi určujeme váhy pro výpočet váženého

$$P(y = 1 \mid x) = h_\theta(x) = 1 / (1 + \exp(-\theta^T x)) = \sigma(\theta^T x)$$

$$P(y = 0 \mid x) = 1 - P(y = 1 \mid x) = 1 - h_\theta(x)$$

součtu hodnot nezávisle proměnných. Tento součet využijeme k výpočtu pravděpodobnosti nelineární funkce. Sigmoidní / logistická funkce je daná rovnicí:

$$y = 1 / (1 + e)$$ (viz obrázek 2).

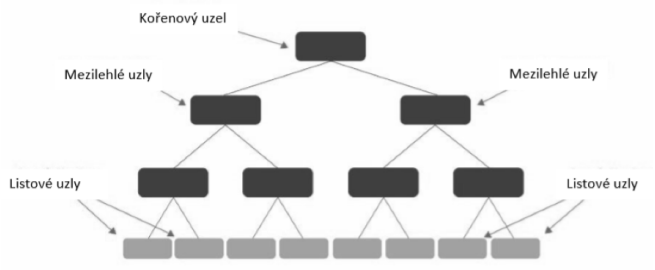


Obrázek 2 – Logistická regrese

Rozhodovací stromy

Rozhodovací strom je podpůrný prostředek. Výsledky jsou ve formě stromu, který může reprezentovat vztah příčiny a jejího efektu. Rozhodovací stromy jsou jednou z nejoblíbenějších DM technik. Hlavními důvody jejich oblíbenosti je zejména jejich přehlednost a snadná interpretace.

Můžeme vytvořit rozhodovací strom tak, že organizujeme vstupní data a prediktory pomocí jednoduchých kritérií, které nějak specifikujeme (viz obrázek 3).



Obrázek 3 – Klasifikační strom

Hlavní kroky jsou následující:

1. Získáme data z dané oblasti.
2. Určíme proměnné pro predikci (prediktory).
3. Určíme cílovou proměnnou.
4. Rozdělíme data na cvičební a zkušební.
5. Generujeme rozhodovací strom.
6. Testujeme a analyzujeme model.

Nevýhodou je, že se model příliš přizpůsobuje datům.

Náhodné stromy

Tato metoda byla navržena, abychom se vyrovnali s některými nevýhodami rozhodovacích stromů.

Náhodný strom zahrnuje rozhodovací stromy, což jsou rozhodovací grafy reprezentující nějaký průběh rozhodování nebo jeho pravděpodobný průběh. Tyto stromy se pak integrují do klasifikačního a regresního modelu.

Při klasifikaci objektu pomocí jeho atributů, každý strom dá určitou klasifikaci, která je jedním "hlasem" pro danou třídu. Nakonec se vybere třída, která má nejvíce hlasů. Při regresi, využíváme průměr predikcí různými stromy.

Náhodné stromy fungují takto:

1. Předpokládáme N objektů. Jde o cvičební soubor.
2. Uvažujeme M vstupních proměnných. Vybereme takové m, které je menší než M. Vytvoříme strom.
3. Každý strom se nechá vyrůst.
4. V další fázi agregujeme predikce z n stromů.

Vektorový stroj

Vektorový stroj (SVM – support vector machine) se nejdříve používal pro analýzu dat (viz obrázek 4). Na začátku se tréninková množina zpracuje SVM algoritmem, který může patřit do různých kategorií.

Nejdříve se ve cvičební fázi vytvoří jednoduchý model dat pomocí lineární diskriminační analýzy.

Algoritmus přiřadí nová data do kategorií, které se naučil rozeznávat v cvičební fázi. K oddělení tříd se využívají hyperroviny.

Naivní metoda Bayese

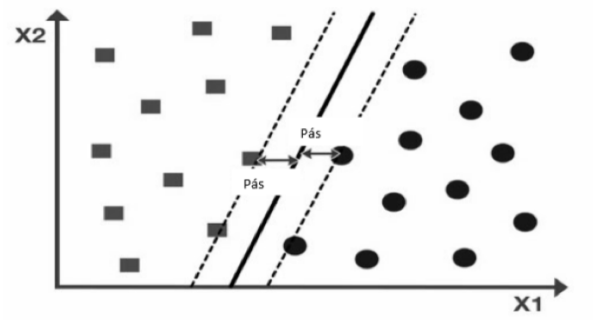
Je jednoduchá, ale přitom efektivní metoda klasifikace na základě maximalizace aposteriorní pravděpodobnosti. Vychází z poznatku, že základní pravděpodobnosti třídy, jsou modifikovány na základě nových dat.

Základní vzorec pro aposteriorní pravděpodobnost:

$$P(A|B) = P(A)\, P(B|A)/P(B)$$

Příkladem je situace, že přijdete pozdě do práce za předpokladu, že je hustá doprava.

Naivní Bayes je klasifikačním algoritmem, který předpokládá, že dvě události jsou nezávislé na sobě navzájem. Tím se



Obrázek 4 – Vektorový stroj

zjednoduší výpočet. Zpočátku šlo o spíše akademické uvažování. Ukázalo se, že pracuje dobře v reálných situacích. Naivní Bayes může být využit k nalezení jednoduchého vztahu mezi různými parametry, aniž bychom měli všechna data.

Umělé neurální sítě

Neurální sítě jsou množinou klasifikátorů uspořádaných do vrstev, kde výstup jedné vrstvy je vstupem do další vrstvy. Vrstvy mezi vstupní vrstvou a výstupní vrstvou jsou skryté vrstvy (viz obrázek 5), čím více je těchto vrstev, tím komplexnější může být klasifikace. Byly populární v 80. a 90. letech. Ale byly pomalé a drahé. Od roku 2006 byly navrženy další techniky, které umožnily další rozvoj.

Také se postupuje tak, že nejdříve se predikuje celá síť X ve fázi bez učitele. Teprve pak nastává fáze s učitelem. Trénování vyžaduje hodně dat a výpočetní kapacity.

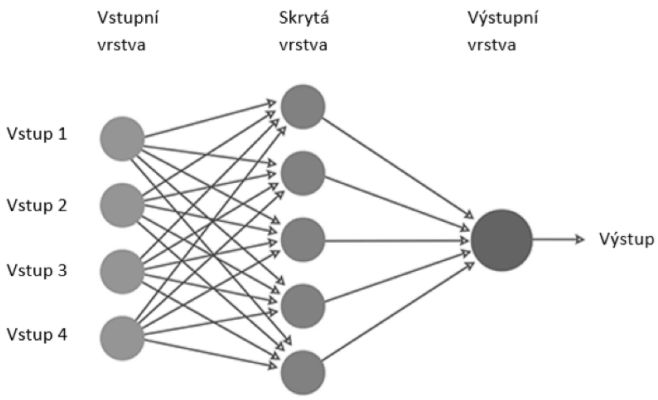
Vytvoříme uzle, které jsou spojené mezi sebou a mají napodobit propojené neurony v mozku. Jednoduše řečeno, každý neuron přijímá informace od ostatních neuronů, vykoná na nich nějakou práci a výstup předá jinému neuronu.

Každý kroužek reprezentuje umělý neuron a šipky reprezentují spojení mezi neurony.

Neurální sítě mohou být užitečné, pokud využijeme závislosti mezi různými třídami.

5 Software pro analýzu velkých dat

Popíšeme dva novější softwarové systémy pro dolování znalostí a analýzu velkých dat. Půjde o systémy Weka a Rapidminer. Staršího data je systém TANAGRA, kterému se v našem sdělení nebudeme věnovat. V podstatě jsou všechny jmenované systémy volně přístupné. Rapidminer má v tomto směru větší omezení.



Obrázek 5 – Neurální síť

WEKA

Weka je volně šiřitelný program vyvinutý na univerzitě Waikato na Novém Zélandě. Tento systém pracuje na principu knihoven programů v Javě. Weka nabízí mnoho algoritmů. Weka obsahuje kolekci vizualizačních nástrojů a algoritmů pro datovou analýzu a prediktivní modelování s grafickým rozhraním pro snadný přístup k těmto funkcím. Weka podporuje několik standardních data miningových úloh, konkrétně preprocessing dat, shlukování, klasifikaci, regresi, vizualizaci a analýzu příznaků.

Nástroj Weka má širokou základnu aktivních uživatelů, kteří do nástroje přispívají svými algoritmy a řešeními v podobě balíčků, které lze jednoduše do nástroje importovat. Tyto balíčky umožňují přidat do nástroje Weka implementace algoritmů, které v základní verzi nejsou. Nástroj Weka spravuje balíčky v Package manageru. Mezi užitečné balíčky patří například WekaHadoop , který umožňuje jistou míru spolupráce mezi nástrojem Weka a technologií Hadoop. Dalším zajímavým balíčkem je Weka-spectral-clusterer, který implementuje do nástroje Weka možnosti spektrálního shlukování.

Nástroj Weka umí zpracovávat textová data a to za předpokladu, že se data nachází v jediném souboru nebo jako relace, kde každý záznam obsahuje fixní počet atributů.

Rapidminer

Rapid Miner německého původu nabízí softwarová řešení v oblasti prediktivní analýzy dat a data miningu. Tento nástroj je zaměřen na sofistikovanou analýzu velkého objemu dat ve velkých databázových systémech, na nestrukturovaná data a texty. Rapid Miner nabízí mnoho nástrojů na zpracování dataminingového modelu a pro jeho vyhodnocení. Následně nabízí také nástroje pro vizualizaci dat, modelů a dalších výsledků. Další významnou oblastí tohoto programu je i hodnocení a odhadování výkonnosti.

Rapid Miner je nastroj pro zpracování, modelování a vizualizaci dat. Integruje v sobě velké množství algoritmů z oblasti statistiky, databází a umělé inteligence podobně jako systém WEKA. Nepostradatelnou součástí programu jsou vizualizační nástroje. Rapid Miner disponuje grafickým designerem pro návrh schémat zpracování dat.

6 Závěr

„Svět je dnes prostoupen daty, jako kyslíkem." Ohromné množství dat, které spotřebováváme a které na nás útočí, je vlastností digitalizovaného světa. Velké datové množiny známé jako big data nejsou zvládnutelné klasickými databázemi. Organizace všech typů potřebují organizovat a analyzovat tato data, aby dělaly lepší rozhodnutí. Náš přehled popisuje různé algoritmy, které využívají analytiky velkých dat.

Analytiky velkých dat patří mezi nejdůležitější technologie informačního průmyslu. Všudypřítomnost velkých dat se přenáší do oblasti komunikačního průmyslu. Ten pomáhá spolu s ukládáním dat v cloudech zvládnout množství dat v internetu a ostatních informačních systémech. Každá technika vede k určité kompresi dat. Existují různé techniky určené k obecné kompresi dat i mimo procedury analytik, které jsme představili.

Dobře přístupné jsou popsané techniky v programových systémech WEKA a Rapidminer. Popisy těchto systému a algoritmů nalezneme v mnoha studentských pracích ekonomických vysokých škol a vysokých škol z oblasti informatiky [např. 5, 8, 12, 13].

Literatura

[1.] BERKA, P.: 2003. Dobývání znalostí z databází. Praha: Academia. ISBN 80-200-1062-9. Dostupné také z: http://sorry.vse.cz/~berka /4IZ450/

[2.] Davisson, L. D., & Gray, R. M. (1976). Data compression. Stroudsburg, Pa: Dowden, Hutchinson & Ross.

[3.] Ebenezer G.E.J. and Durga S: Big data analytics in healthcare: a survey. ARPN Journal of Engineering and Applied Sciences. 10 (8), 2015.

[4.] Foster, I. et al.: Big data and Social Science. A practical guide to methods and tools. London : CRC Press, 2017.

[5.] Karafiát, M. Big Data – Metody zpracování a analýzy velkých dat. Diplomová práce. Univerzita Tomáše Bati 2017.

[6.] Larrose, D. T., Larose, Ch. D. : Discovering knowledge in data. An introduction to data mining. (2. ed.) Wiley 2014

[7.] Masoo, A., Al-Jumaily , A.A. Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms. International Journal of Biomedical Imaging, 22 pages, 2013.

[8.] Nováková, M. Analýza Big Data v oblasti zdravotnictví. Diplomová práce. VŠE 2015.

[9.] Perera, S., & Gunarathne, T. (2013). Hadoop MapReduce cookbook: Recipes for analyzing large and complex datasets with Hadoop MapReduce. Birmingham: Packt Pub.

[10.] Prajapati, V. (2013). Big Data analytics with R and Hadoop: Set up an integrated infrastructure of R and Hadoop to turn your data analytics into Big Data analytics. Birmingham: Packt Publishing.

[11.] Štablová, L. Algoritmy v dataminingu. Bakalářská práce. Univerzita Tomáše Bati 2010.

[12.] Vozába, M. Tools and Methods for Big Data Analysis. Master Thesis. Západočeská univerzita 2016.

Kontakt:

Prof. Jan Hendl
FSV UK – katedra sociologie
U Kříže 8 a 10
158 00 Praha 5 – Jinonice
e-mail: *jan.hendl@fsv.cuni.cz*