Machine learning a data mining 1

Jan Outrata

KATEDRA INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI

přednášky

Literatura

- Tan P.-N., Steinbach M., Kumar V.: Introduction to Data Mining. Pearson Education, 2005.
- Marsland S.: Machine Learning: An Algorithmic Perspective, 2nd ed. Chapman and Hall/CRC, 2014.
- Zaki M. J., Meira W. Jr: Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014.
- Poole D. L., Mackworth A. K.: Artificial Intelligence: Foundations of Computational Agents, 2nd ed. Cambridge University Press, 2017.

Úvod

Úvod

- narůstající objem generovaných a sbíraných dat (výzkum, senzory, obchod, osobní; Zettabyte Era) – jednoduchost sběru a uložení $\rightsquigarrow$ jakákoliv, kdekoliv, kdykoliv, často jen s vírou užitečnosti
- omezení lidí a tradičních (statistických) metod analýzy a využívání dat: **velikost a komplexnost dat („surovost", nový typ)**, potřeba **jiné zpracování (výstup)**

$\rightarrow$ **data mining** – extrakce „informace, znalosti" z dat (nových typů, novým způsobem), tradiční základem (motivace stejné), potřeba **porozumění** datům a oblasti

$\rightarrow$ **machine learning** – využití „informace, znalosti" z dat pro zlepšení nějaké činnosti, potřeba schopnosti **adaptace a generalizace** na data a z dat

- metody často těžko zařaditelné – jen extrakce „informace, znalosti" nebo už i její využití k něčemu?? $\rightsquigarrow$ splývání

Úvod

- aplikace (a také motivace/důvody vzniku):
  - **výzkum (a průmysl)**: sběr dat o lidech a pro nové objevy a potřeby, např. o pacientech (testy), genech a proteinech (sekvenování genomu, exprese), klimatu (senzory satelitů), vesmíru (teleskopy), materiálech (výsledky pokusů), systémech a strojích (monitoring, asistence) – velikost dat, nový typ (časoprostornost), šumovost apod.
  - **obchod**: info o nákupech a zákaznících (z prodejen, e-shopů, supportu aj., profilace), typicky pro vylepšování a cílení nabídky (personalizace, doporučování, predikce), cenové politiky (konkurence), optimalizaci zásob a distribuce zboží, detekce podvodů (služby)
  - **IT**: komunikační data (síťový provoz, management, obchod), multimedia – velikost, stream, zpracování, rozpoznávání (obraz, zvuk)
  - ! otázky narušení soukromí (obchod, zdravotnictví) → privacy-preserving data mining
- konference na data mining: IEEE ICDM, ACM SIGKDD, SDM, PKDD, PAKDD
- konference na machine learning: ICML, AAAI

Data mining (DM, „dolování dat“)

= proces extrakce nových a užitečných (jinak skrytých) informací z (rozsáhlých) dat

- ne získávání (partikulárních) informací, např. databázový nebo vyhledávací dotaz = **information retrieval** – datové struktury (indexy), vyhledávací algoritmy, DM vylepšují
- součást celého procesu získávání (objevování) znalosti z dat, „transformace“ dat na informace = **knowledge discovery in databases (KDD)** (více např. Fayyad):
  1. **data preprocessing** – sběr a předzpracování vstupních dat, např. získávání z více zdrojů, formátování, čištění (od šumu), výběr a úpravy atributů (feature selection, dimensionality reduction, normalization) a záznamů aj., časově náročné
  2. data mining – vybraná metoda se zvolenými parametry
  3. **postprocessing** – úprava výstupní informace pro využití (např. v systémech pro rozhodování, v metodách **machine learning**), např. filtrace validní a užitečné (statisticky, testování hypotéz), vizualizace, interpretace
  4. **opakování** – s jinou metodou nebo parametry

Machine learning (ML, strojové učení)

= proces adaptace na a generalizace z dat o (obecně) „činnosti“ pro její správnější „vykonávání“, na základě informací z dat, např. „naučení se“ kroků k vítězství ve hře na základě průběhů her, ale třeba i jen „vytvoření“ predikce hodnoty nějaké veličiny na základě hodnot jiných veličin

- člověk se učí ze zkušenosti: zapamatování, adaptace na a generalizace z ní pro správnější rozhodování → modelování informacemi z dat a procesu učení metodou ML
- části/fáze:
  1. data preprocessing – stejné jako u KDD
  2. **feature selection** – výběr jen některých „proměnných“ (atributů dat), někdy sloučené se sběrem a předzpracováním dat (např. data rozsáhlá, těžko k získání všech, chybovost při získávání apod.)
  3. výběr metody ML a volba parametrů – experimentování
  4. **training** = proces učení – aplikace metody ML na případná tzv. trénovací data (training set) pro naučení se **modelu dat**, může být výpočetně náročnější (probíhá málo často) než následné použití modelu na další data
  5. **evaluation** – testování a vyhodnocení „správnosti“ naučeného modelu na tzv. testovacích datech (testing/validation set), před jeho použitím na další data, často porovnáním s informacemi od expertů v oblasti

Machine learning (ML, strojové učení)

- část oblasti/základ **umělé inteligence (artificial intelligence, AI)** – „inteligentní chování“ na základě zjištěných informací a (strojového) učení, ale také uvažování a usuzování (reasoning, logical deduction) aj. – symbolické zpracování dat a informací (DM a ML = subsymbolic)

DM a ML

Důvody pro DM a ML:

- **škálovatelnost metod** – pro rozsáhlá data (mnoho záznamů) $\rightarrow$ vyhledávací problémy a strategie, efektivní datové struktury, paralelní a distribuované algoritmy
- **dimenze dat** = mnoho atributů (člověk $\sim$ tři dimenze?), navíc časoprostorových – výpočetní složitost algoritmů narůstá $\rightarrow$ metody redukce dimenzionality (projekce, může skrýt informaci)
- **komplexnost dat** – heterogenní (symbolické i číselné) atributy, semi-strukturované záznamy (text, časové aj.) – vztahy (např. časoprostorové korelace, souvislosti, hierarchické)
- **více zdrojů dat** – problémy komunikace při výpočtu, konsolidace výsledků, bezpečnosti aj. $\rightarrow$ distribuované algoritmy
- **netradiční analýza** – automatizace procesu generování a testování hypotéz, jiný typ výstupu (interpretace)

DM a ML

Vznik z a využití metod:

- **statistiky/matematiky** (DM): vzory (sampling, **pattern recognition**), odhady (estimation), testování hypotéz (více k DM vs. statistika např. Hand)
- **biologie** (ML): modelování/simulace procesu učení (adaptace, generalizace) a biologických procesů (neuronové sítě, evoluční algoritmy)
- **informatiky**: algoritmizace, optimalizace, složitosti výpočtů, evolučních výpočtů, teorie informace, vizualizace, information retrieval, ...
- **IT**: databázových a informačních systémů, paralelních a distribuovaných systémů

DM a ML

Úlohy (typicky):

- **prediktivní** (ML) = predikce hodnoty určitého atributu (**cílový (target), závislá proměnná**) na základě hodnot jiných atributů (**explanatory, nezávislá proměnná**)
- **deskriptivní** (DM) = explorativní extrakce vzorů/informace (korelace, trendy, shluky, anomálie) popisujících vztahy v datech

Typy ML:

- **supervised learning** (učení s učitelem, z příkladů) = naučení se správných akcí (v rámci činnosti) na základě známých správných akcí pro nějaká, trénovací data (o akcích/činnosti), nejběžnější
- **unsupervised learning** (učení bez učitele) = naučení se vzorů (správných) akcí vyhledáním v datech a kategorizací vzorů, pro data neznámé správné akce
- **reinforcement learning** (zpětnovazební učení, s kritikou) = naučení se správných akcí prohledáváním a zkoušením možností na základě dat a sdělení o (míře) ne/správnosti akce, ne přímo správné akce
- **evolutionary learning** (evoluční učení) = naučení se správnější akce modelováním biologické evoluce (adaptace pro přežití) v prostředí daném daty

Úlohy DM a ML

Prediktivní modelování (ML)

= tvorba/naučení **modelu dat**/cílového atributu (závislé proměné) jako funkce jiných atributů (nezávislých proměnných) – často ve formě pravidel „jestliže hodnoty, pak hodnota“

- **klasifikace** – pro diskrétní cílový atribut, např. bude-nebude pršet, zákazník koupí-nekoupí, pacient má-nemá nemoc, ...
- **regrese** – pro spojitý cílový atribut, např. předpověď počasí (teploty), pravděpodobnost zemětřesení, ...

→ minimalizace chyby mezi modelem predikovanou a skutečnou hodnotou cílového atributu

Úlohy DM a ML

Prediktivní modelování / klasifikace (ML)

název	hymenofor	prsten	pochva	jedovatá
hřib borový	rourky	ne	ne	ne
kozák	rourky	ne	ne	ne
klouzek	rourky	ano	ne	ne
čirůvka	lupeny	ne	ne	ne
bedla	lupeny	ano	ne	ne
žampión	lupeny	ano	ne	ne
muchomůrka zelená	lupeny	ano	ano	ano
muchomůrka červená	lupeny	ano	ne	ano
závojenka	lupeny	ne	ne	ano
hřib satan	rourky	ne	ne	ano

část modelu dat?:

jestliže hymenofor = rourky a prsten = ano, pak jednovatá = ne

jestliže hymenofor = lupeny a pochva = ano, pak jedovatá = ano

Úlohy DM a ML

Asociační analýza (DM)

= extrakce vzorů („zajímavých“ vztahů, asociací, skrytých v datech) popisujících související (asociované) atributy – typicky ve formě implikací mezi atributy (items) nebo jejich podmnožin (itemsets)

- např. geny se související funkcí, webové stránky přistupované společně, spolu se vyskytující lidé, ...

→ efektivně jen nejzajímavější (často se v datech vyskytující) vzory, kvůli (exponenciální) velikosti prohledávaného prostoru vzorů

Úlohy DM a ML

Asociační analýza (DM)

tID	obsah košíku
1	chleba, máslo, mléko
2	mléko, káva, kuře, ovoce, sušenky
3	máslo, vejce, ovoce, zelenina
4	chleba, ryba, ovoce, zelenina
5	chleba, vejce, kuře, ovoce
6	máslo, mléko, káva, sušenky

vzory?:

$\{\text{káva}\} \rightarrow \{\text{mléko, sušenky}\}$

$\{\text{zelenina}\} \rightarrow \{\text{ovoce}\}$

Úlohy DM a ML

Shlukování (shluková analýza) (DM)

- nalezení skupin (shluků) záznamů navzájem podobnějších v rámci shluku než mimo něj (v jiných shlucích)
- např. podobné objekty, příbuzné oblasti, ...

→ na základě (definované) podobnosti nad záznamy

Úlohy DM a ML

Shlukování (shluková analýza) (DM)

- nalezení skupin (shluků) záznamů navzájem podobnějších v rámci shluku než mimo něj (v jiných shlucích)
- např. podobné objekty, příbuzné oblasti, ...

→ na základě (definované) podobnosti nad záznamy

článek	klíčová slova
1	škola, rodiče, ekonomika, kvalita, prestiž, hodnocení, technologie, vzdělání
2	škola, rodiče, kvalita, individualita, hra, hodnocení, vzdělání
3	metodika, škola, hodnocení, inovace, program, technologie, vzdělání
4	emoce, strach, věk, terapie, rodina, porucha
5	věk, vývoj, emoce, porucha, vztahy
6	produktivita, bezpečnost, vzdělání, emoce, výsledky, vztahy

shluky?:

$\{1,2,3\}$, $\{4,5,6\}$

Úlohy DM a ML

Detekce anomálií (DM)

- identifikace záznamů, tzv. **anomálií (outliers)**, výrazně odlišných od ostatních – problém s falešným označením běžných záznamů
- např. detekce podvodů, (síťových) incidentů, neobvyklých vzorů, ...

→ tvorba „profilů“ (vzorů) běžných záznamů a porovnání záznamů s nimi

Úlohy DM a ML

Detekce anomálií (DM)

- identifikace záznamů, tzv. **anomálií (outliers)**, výrazně odlišných od ostatních – problém s falešným označením běžných záznamů
- např. detekce podvodů, (síťových) incidentů, neobvyklých vzorů, ...

→ tvorba „profilů“ (vzorů) běžných záznamů a porovnání záznamů s nimi

odchozí platby z účtu (po měsících):
650, 730, 50, 580, 6800, 880
490, 920, 660, 6800, 390
570, 410, 50000, 770, 6800, 840

anomálie (outliers)?:
50, 50000