VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Data mining
4. projekt

16. dubna 2017

Martin Omacht

1 Co je to data mining?

Dolování dat, je definováno jako proces objevování různých vzorů v datech. Tento proces musí být automatický nebo (většinou) poloautomatický. Vyhledané vzory musí být smysluplné, tak že vedou k nějaké výhodě, většinou ekonomické. Data jsou vždy přítomné v podstatném množství. [5]

2 Užití data miningu

> *„Where there are data, there are data mining applications“* [6]

Data mining se užívá téměř všude, kde jsou data. Dá se rozdělit do dvou hlavních skupin – predikce a deskripce. [7] Příklad predikce je software MineCarTM, který využívá nejen aktuálně naměřené data o vozidlu, ale také historické data, informace o zájmech majitele vozidla a detaily o opravě. Pomocí data miningu poté předem upozorní na potřebnou opravu nebo prohlídku vozidla. [1]

Společnost Apple si například nechala patentovat metodu analýzy výkonu softwaru pomocí data miningu. [2]

3 Dataminingové úlohy

Úlohy data miningu můžeme rozdělit do několika skupin:

- Klasifikace
- Shlukování/Segmentace
- Predikce
- Regrese
- Asociační pravidla
- Text Mining [9]

4 Algoritmy data miningu

Existuje mnoho algoritmů pro dolování dat, proto zde vyjmenuju jenom pár z nich.

4.1 Rozhodovací stromy

Rozhodovací stromy jsou jednoduchou, ale užitečnou formou analýzy vícero proměnných. Nabízejí unikátní možnosti, jak doplnit nebo nahradit:

- Tradiční statistické formy analýzy (např. násobná lineární regrese)
- Různé druhy nástrojů dolování z dat (např. neuronové sítě)
- Nedávno vyvinuté vícerozměrné formy reportů a analýz v oboru business intelligence 9 [3]

4.2 *k*-means algoritmus

Tento algoritmus je jednoduchá iterativní metoda rozdělení daných dat na uživatesky specifikovaný počet shluků. Byl objeven několika vědci z různých oborů, zejména Lloydem (1957, 1982), Forgeyem (1965), Friedmanem a Rubinem (1967) a McQueenem (1967). [10]

4.3 Support vector machines

Support vector machines je metoda, která hledá nejlepší rozhodovací linii mezi třídami dat. Lineární jádrovou funkci pro tento typ učení nalezl Vladimir Vapnik v roce 1963. Třídy by se neměli být lineárně separovatelné a nepřekrývající se. [4]

5 Zrychlení data miningu

Některé algoritmy pro data mining jsou výpočetně velice náročné a analýza dat tak může trvat několik hodin. Pro zrychlení analýzy existují různé alternativy. Zatímco někteří uživatelé spoléhají na cloudové řešení, heterogení prostředí založené na GPU architekturách se jeví jako cenné řešení pro zlepšní výkonu s významnou úsporou nákladů. [8]

Reference

[1] Data Mining; Sprint and Agnik Team on Advanced Data Mining for Enhanced Vehicle Performance Applications. *Telecommunications Weekly*, 2012: str. 653, ISSN 1945-841x.

[2] Apple Assigned Patent for Software Performance Analysis Using Data Mining. *Targeted News Service*, 2013.

[3] FABIAN, J.: *Využití technik data miningu v různých odvětvích.* Diplomová práce, Vysoké učení technické v Brně, Fakulta podnikatelská, 2013.

[4] Hricko, J.: *Rozpoznávání ručně psaných číslic pomocí support vecotr machines*. Bakalářská práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2010.

[5] Ian H. WITTEN, E. F.: *Data mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann Publishers, druhé vydání, 2005, ISBN 0-12-088407-0.

[6] Jiawei HAN, J. P., Micheline KAMBER: *Data mining: concepts and techniques*. Waltham: Morgan Kaufmann, tøetí vydání, 2012, ISBN 978-0-12-381479-1.

[7] Procházka, M.: Data mining: jiný pohled na problém. [online], [cit. 2017-04-16].
URL <http://vtm.e15.cz/aktuality/data-mining-jiny-pohled-na-problem>

[8] ScienceDirect: *Performance Improvement of Data Mining in Weka through GPU Acceleration*, Elsevier B.V., 2014, ISSN 1877-0509.

[9] StatSoft: Úvod do data miningu. [online], [cit. 2017-04-15].
URL <http://www.statsoft.cz/file1/PDF/newsletter/2014_02_26_StatSoft_Uvod_do_data_miningu.pdf>

[10] Xindong WU, J. R. Q., V. Kumar: Top 10 algorithms in data mining. 4. prosinec 2007, [online], [cit. 2017-04-15].
URL <http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf>