Efektivní analýza velkých dat pomocí Apache Spark a samoučících neuronových sítí na jediném počítači

David Andrešič, Petr Šaloun

Katedra informatiky, FEI VŠB-TUO v Ostravě
17. listopadu 15/2172, 708 33 Ostrava - Poruba

{david.andresic.st, petr.saloun}@vsb.cz

Abstrakt. Apache Spark je běžně používaná platforma pro analýzu velkých dat na velkých počítačových clusterech, kde pro svou práci využívá především hlavní paměť počítače. Pokusili jsme se přidat softwarovou knihovnu samoučící se neuronové sítě do jednoho takového analytického celku pro big data. Výsledek je efektivní a rychlý dokonce na jediném běžném počítači.

Tento přístup je přínosem pro výzkumníky s omezenými zdroji, kterým přináší možnost analýzy velkých dat. Náš nápad byl experimentálně ověřen a je popsán zde. Jako případovou studii pro naši metodu jsme použili dostupná data ze sociální sítě Twitter, konkrétně tweety pro hashtag #Brexit a jejich analýzu sentimentu, přičemž jsme hledali korelace s burzovními daty.

Klíčová slova: Apache Spark, samoučící neuronové sítě, big data, Twitter, brexit, burza

1 Úvod

Korelace tweetů a burzy statistickými metodami je častým předmětem výzkumu i publikací. Náš přístup pro sociální síť Twitter a vývoj burzy zahrnoval agregování veřejných dat a shlukovou analýzu metodami strojového učení (samo-organizujících mapy – SOM) pomocí Apache Spark[1] určeného pro zpracování velkých dat na počítačových clusterech a to na jediném, standardním počítači pro pre-processing velkých dat. Jelikož se Spark snaží zpracovávat data primárně v paměti RAM, čelili jsme omezeným prostředkům. S podobným problémem jsme se potýkali v analýze shluků pomocí SOM, kde byla efektivita implementací velmi rozdílná (viz část 3.2).

2 Metodologie a datové sady

Celý postup vypadal takto:

– *Analýza sentimentu* za účelem zjištění jak pozitivní, či negativní daný tweet byl.

[1] Apache Spark homepage: http://spark.apache.org/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 67-71.

- *Spojení a agregování dat,* zahrnující výpočet korelačních koeficientů a transformaci dat do podoby vhodné pro SOM, vše s použitím Apache Spark.
- *Shluková analýza* pomocí metody SOM ve snaze najít shluky podobných společností, které korelovaly nejvíce.

2.1 Analýza sentimentu dat z Twitteru

Tweety pro hashtag *#brexit* byly k dispozici pro několik dní z období 29.4.2016 až 2.7.2016. Analýza sentimentu ukázala, že před referendem (23.6.2016) byly tweety spíše kritikou EU, zatímco po něm k Brexitu samotnému. Celkový počet analyzovaných tweetů byl 21137 (10799 před referendem a 10338 po referendu).

2.2 Burzovní data a data společností

Zdrojem věrohodných burzovních dat bylo Yahoo Finance[2]. Poskytuje historická data z mnoha burz, včetně námi zvolené London Stock Exchange (1292 společností). Yahoo Finance umožňuje stažení těchto burzovních dat: datum, hodnota akcie při otevření a zavření burzy v daný den, maximum a minimum hodnoty daný den a objem zobchodovaných akcií. K těmto jsme přidali další atributy z Google Finance (adresa, oblasti působnosti, město, země).

3 Transformace a agregace dat

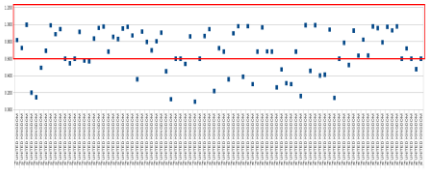
Nejprve jsme použili data podrobněji popsaná v části 2.1 s interpolací chybějících hodnot. SOM vyžadovala, aby pro každou společnost byly informace o střední hodnotě ceny akcie při uzavření burzy a počtu zobchodovaných akcií, jejich korelační koeficient se sentimentem, město, zemi a oboru působnosti. Výpočty byly rozděleny do tří period: před a po referendu a celé období. Spojením obou burzovních datových sad jsme získali informace o 962 společnostech. Následně byl spočítán korelační koeficient pro hodnoty burzovních ukazatelů a sentiment v daný den pro každou periodu pomocí Sparku v módu clusteru o jediném uzlu (počítač Intel Core i3 se dvěma jádry/čtyřmi vlákny, 8GB RAM). Jde o současný komoditní hardware, pro který je Spark navržen (škálování do šířky). Pokusili jsme se všechny operace spojení datových sad a selekce provést najednou v RAM. Bohužel, každý výběrový dotaz z důvodu velikosti datového rámce (stovky řádků pro každou společnost, které byly dále spojeny s jejími dodatečnými atributy) zabral několik minut a výpočty ani po několika dnech nedoběhly. Klasický přístup relačních databází s indexem na atribut burzovního symbolu (tickeru) ve světě Spark SQL nelze použít (pouze transformací do RDD). Vyladili jsme proto části kódu tak, aby se výpočty prováděly iterativně na menších blocích. Pomocí této optimalizace jsme se vyhli drahým výběrovým operacím v rozsáhlém datovém rámci a dokončili výpočty za necelé dvě hodiny. Jelikož se při výpočtu pro každou společnost používají pouze její data, je možné tento postup

[2] Yahoo Finance: https://finance.yahoo.com/

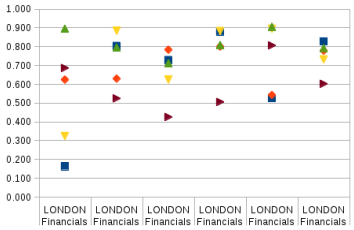
paralelizací dále zefektivnit. Výsledná tabulka tak navíc obsahovala střední hodnoty akcií a jejich zobchodovaného objemu na konci dne pro všechny periody normalizované v rozsahu 0-1. Kategorické (nominální) atributy byly „binarizovány“ (např. „London“ v atributu „town“ vytvořil nový atribut „is_town_London“ nabývající hodnot 0, nebo 1).

3.1 Analýza shluků

Zde jsme zvolili SOM [3], které jsou na rozdíl od dále uvažovaných k-means méně náchylné k lokálním optimům [2]. Jednou z oblíbených implementací je Weka[3]. V té nicméně ani po několika hodinách pro 100 vzorků z naší datové sady výpočty nedoběhly. Vyzkoušeli jsme tedy implementaci *java-ml*[4] a skončili s obdélníkovou mřížkou 5x5, 1000 iteracemi, koef. učení 0.5 a počátečním radiusem 8, která vygenerovala 24 nejprůkaznějších shluků během několika minut. Pro vyhodnocení jsme využili bodové diagramy (BD; pomocí *java-ml* nelze sestavit U-matici). Shluky ukazovaly korelaci finančního sektoru (FS) v Londýně se sentimentem (viz Obr. 1, 2 a 3). Toto není překvapující, Londýn je brán za fin. centrum Velké Británie (VB) s mladší a vzdělanější populací aktivní na sociálních sítích [4]. Lze také vidět korelaci společností z FS VB po referendu (Obr. 1). To naznačuje, že FS reflektoval sentiment na Twitteru. Naopak technologický sektor nekoreloval vůbec (pouze společnosti přímo z VB vykazovaly mírně větší korelační koeficienty - KK), viz Obr. 4.

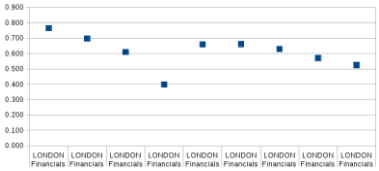


Obr. 1. BD shluku 3. Většina bodů (společnosti ve VB a KK pro hodnoty akcií na konci dne po referendu) má KK větší, než 0.6.

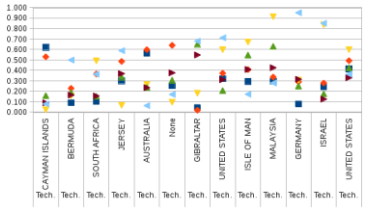


Obr. 2. BD shluku 8. Většina bodů (pro téměř všechny KK finančních společností ve VB) má KK větší, než 0.6.

[3] Weka homepage: http://www.cs.waikato.ac.nz/ml/weka/
[4] java-ml homepage: http://java-ml.sourceforge.net/



Obr. 3. BD shluku 9. Většina bodů (společnosti ve VB a KK pro hodnoty akcií na konci dne před referendem) mají KK větší, než 0.6.



Obr. 4. BD shluku 11. Většina bodů (tech. společnosti mimo VB a jejich KK) mají KK menší, než 0.5.

4 Závěr

Popsaná případová studie Londýnské burzy ukázala největší korelace ve finančním sektoru VB, zjevné u společností sídlících v Londýně. Naopak, technologický sektor nevykazoval korelaci žádnou. Jak je popsáno v [4] a [5], data z Twitteru (a sociálních sítí obecně) často postrádají demografické, věkové, či vzdělanostní údaje. Pro další výzkum bychom měli uvažovat doplnění těchto informací. Je také nutné uvažovat, že uživatelé sociální sítě představují specifickou skupinu obyvatelstva [5].

Je-li Apache Spark v clusteru s jedním uzlem, můžou omezené zdroje (především RAM) způsobit pomalý výběr z datových rámců. Přepsáním kódu spojení datových rámců jsme se vyhnuli paměťově náročným operacím, což umožnilo zredukovat výpočetní čas z několika dní na dvě hodiny. Tento postup byl dále paralelizovatelný.

Použití SOM umožnilo analýzu fenoménu Brexitu se stabilními a prokazatelnými výsledky napříč iteracemi. Lišila se efektivita knihoven, (implementace *java-ml* byla mnohem rychlejší, než Weka, což může být rovněž přínosné pro další výzkumníky).

Výzkum byl v plné verzi publikován anglicky na mezinárodní konferenci.

Literatura

1. A. J. Awan, M. Brorsson, V. Vlassov and E. Ayguade, "Performance Characterization of In-Memory Data Analytics on a Modern Cloud Server," *Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on*, pp. 1-8, 2015.
2. F. Bação, V. Lobo and M. Painho, "Self-organizing Maps as Substitutes for K-Means Clustering," *Computational Science -- ICCS 2005: 5th International Conference, Atlanta, GA, USA, May 22-25, 2005, Proceedings, Part III*, pp. 476-483, 2005

3. T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52-65, January 2013. [Online]. Available: http://aisii.azc.uam.mx/mcbc/Cursos/IntCompt/Lectura 8. SOM.pdf. [Accessed March 10, 2017].
4. L. Vasiliu, R. McDermott, M. Zarrouk, M. Hürlimann, B. Davis, T. Daudert, M. B. Khaled, D. Byrne, S. Fernández, A. Freitas, F. Caroli, S. Handschuh and Angelo Cavallini, "In or Out? Real-Time Monitoring of BREXIT sentiment on Twitter," *CEUR Workshop Proceedings*, 2016.
5. M. Hürlimann, B. Davis, K. Cortis, A. Freitas, S. Handschuh and S. Fernández, "A Twitter Sentiment Gold Standard for the Brexit Referendum," *Proceedings of the 12th International Conference on Semantic Systems*, pp. 193-196, 2016.

Poděkování: Tento článek vznikl díky finanční podpoře grantů: Grantová agentura ČR- GACR P103/15/06700S, Grant SGS č. SGS 2017/134, VŠB-Technická univerzita Ostrava, MŠMT Národní program udržitelnosti (NPU II) projekt ”IT4Innovations excellence in science - LQ1602”.

Annotation:

Apache Spark is a common big data analysis platform on large computer clusters. It uses primarily the main memory. We added a SOM library to such big data analytical stack. The result is effective and fast enough even on a single computer. This approach brings the possibility of big data analysis even for researchers with limited resources. Our idea was experimentaly tested and is described here. We used the Twitter data (tweets for #brexit hashtag) and their sentiment analysis for finding correlations with stock exchange data as a case study for our approach.