Chapter

Introductory Chapter: Data Streams and Online Learning in Social Media

Alberto Cano

1. Introduction

Since the establishment of the World Wide Web and online social media networks, people have changed the way they communicate, share experiences, and connect with each other, both in their professional and personal lives [1]. Billions of users exchange digital information on popular sites such as Facebook, Twitter, and LinkedIn but also in smaller and topic-specific networks [2, 3]. The ever-increasing number of users and content shared makes it challenging for information systems to process all the information, especially if we consider the increasing speed at which content is generated [4, 5]. Consequently, new open issues have risen regarding the effective and efficient processing of such high-speed large-scale volumes of data in online social media. How can we build machine learning systems that can handle and scale to the impressive volume of data? How can we keep a low latency in the response to classifying new real-time data? How can we classify users and their behavior? How can we early detect changes in the user's behavior and emerging trends? These are open questions to the data science scientific community [6–8].

In recent years, the design of machine learning systems to detect bot networks [9], fake content [10], or hate speech in social media, among many others, has gained increasing popularity. One may think of fake reviews on Amazon, fake news on user forums, bots on Twitter following/retweeting certain politicians to promote political campaigns, or hate campaigns aimed at systematically attacking certain underprivileged groups with messages full of hate [11, 12]. All of these are growing challenges in online social media networks which demand new machine learning solutions.

Analyzing temporal and contextual patterns in this data is important to discover emerging topics, trends, correlations, causations, and periodic occurrences, happening on real-time data. Data stream mining is the machine learning area devoted to analyzing real-time high-speed online data. This chapter will present some advances on research and applications of data stream mining to problems in online social media.

2. Data stream mining for online learning

A data stream is an ordered and potentially unbounded sequence of data instances arriving continuously to a machine learning system [13]. It is unknown when the volume and speed at which data will arrive to the system. However, it is required to provide a fast prediction, as a delay in the prediction or bottlenecks are not permitted. Moreover, machine learning models need to be continuously updated to make sure they reflect the most up-to-data state of the stream, following

up with any changes that data may experience with time. Data may evolve with time and experience the appearance or fading of data classes, features, and data distributions. The changes that data may experience with time are known as concept drift [14], and it may be analyzed from multiple perspectives.

Decision boundaries: real vs. virtual drift. Real concept drift has an impact in the classification boundaries, increasing the error when new instances are misclassified. Virtual concept drift observes a change in the distribution of data with time but does not affect the decision boundaries.

Scope of the changes: global vs. local. Global concept drift affects the entire stream, while local affects only certain regions of the feature space or a subset of features.

Speed of drift: incremental vs. gradual. Incremental concept drift is a steady progression from one concept to another. Therefore, it comprises multiple intermediate concepts in between. On the other hand, gradual concept drift reflects a change in a probability distribution in which there is a decreasing probability of observing the old concept and an increasing probability of the new concept to occur.

Concept drift may also suffer from recurrent patterns which happen periodically (e.g., seasonal trends) or blips (noise or random changes that should be ignored and not to be confused with a true drift).

Detecting concept drift is a challenging task itself. There are two types of detectors: explicit and implicit. Explicit concept drift detectors explicitly monitor the characteristics of the stream including statistical distribution variations, density changes, etc. They emit an alert whenever a drift is detected, informing the classifier to update the classification model. Implicit concept drift detectors assume the classifier inherently adapts itself to changes, e.g. by using a dynamic sliding window or by using online learners. How can we detect the emerging of new topics and the fading of others on Twitter? Detecting and anticipating to concept drift remains an open challenge to the machine learning community [15].

Ensemble learning combines multiple classifiers to jointly provide an improved performance compared to single classifiers [16–18]. Ensembles must be composed of mutually complementary and individually competent classifiers, advocating for diversity in its components. Ensembles are natural solvers for stream mining problems with concept drift, as new concepts may be modeled by new components added to the ensemble, whereas older concepts no longer present in the stream may be simply seen their classifiers deleted from the ensemble. Moreover, in the case of recurrent drifts, components may just be disabled (not deleted) so that by the time we anticipate the concept will reoccur, then we may preemptively reenable, avoiding the cost of relearning the classifier, both in terms of lost time and accuracy. One may think about the recommendation systems on Amazon to show the most likely purchased product to users in recurrent seasons (Mother's Day, Christmas, etc.).

Class imbalance is another recurrent problem in data stream mining. Data class distributions may not be evenly represented, plus their proportions may change with time. The majority class may become the minority or reversely. In such a situation, ensembles also help to balance the representativeness of the data and the classification metrics performance as one may want not to bias the algorithms to learn the majority class only. To resolve these issues, several authors have proposed ensembles for drifting, imbalanced streams.

The Kappa Updated Ensemble [16] for drifting data stream mining proposes a hybrid online and batch-based architecture that uses the Kappa statistic for dynamic weighting and selection of classifier components. To achieve ensemble diversity, it proposes to employ different subsets of features on each classifier, along with online bagging. Thanks to the Kappa statistic, it abstains predictions from models that negatively impact the performance of the classifier, increasing the

robustness of the ensemble. Abstaining components has also shown to improve the classification in other non-imbalanced streaming problems.

Some real-world problems are characterized for having instances simultaneously categorized into multiple labels. This problem is known as multi-label learning [19–20]. The complexity of correctly classifying the instance increases with the size of the output space. Moreover, concept drift may simultaneously happen to some or many of the labels. Therefore, it is more difficult to detect and adapt to concept drift. Authors have proposed solutions for multi-label data streams, including self-adjusting windows to identify the more accurate and most recent subset of instances in a sliding window [19]. Moreover, punitive systems have shown that penalizing instances leading to erroneous label predictions and early removing them from the window increase the overall accuracy of the classifier [21].

Algorithmic solutions to these open issues in data stream mining come at the expense of an increased computational cost. It would not be possible to provide both an accurate and fast classification and fast update of the classification model if one wants to adapt to concept drift quickly. Therefore, high-performance computing architectures are needed to speed up algorithms in order to meet the real-time constraints of stream learning.

GPUs and MapReduce distributed computing frameworks have become increasingly popular to speed up large-scale data mining problems. They offer higher scalability to big data problems for a fraction of the cost of a traditional mainframe solution. GPUs are particularly efficient for streaming environments and provide a very fast decision with minimum label latency [22–27]. However, they are often associated with a more difficult code implementation and limited memory, which makes it difficult to scale to true big data problems. Distributed GPU solutions may partially alleviate but not solve this problem.

While Apache Hadoop was one of the first and most popular frameworks for MapReduce publicly available, it does not provide the tools nor the speed to work for real-time streams. In such a scenario, there are other solutions much more efficient for real-time streams. Apache Spark Streaming, Apache Flink, and Apache Storm are MapReduce-based frameworks for streaming data [28–32]. However, they lack efficient implementations of effective machine learning algorithms. Therefore, there is a need to implement publicly available methods for stream learning in such frameworks. There are some works on distributed nearest neighbor search and feature selection. However, there is a whole area of asynchronous deep learning models for data streams on MapReduce that is yet to be addressed. While deep learning-based methods may provide the best accuracy, there is also a need to provide interpretable models and demand explanations of the prediction system, particularly for domains requiring accountability, such as medical diagnosis.

3. Conclusions

The popularity of online social media demands new transformative solutions to the emerging problems in social media content and networks, including community detection, bot detection, fake reviews, user behavior prediction, etc. Machine learning provides solutions to these problems, but there are many unresolved open issues. Data stream mining focuses on the analysis of the real-time high-speed streams of data that continuously arrive to a classifier. Data stream mining can detect changes in the property of the stream data and adapt the classification model accordingly. However, there are still too may open issues both from the basic research and application perspectives [32–36] which call for the scientific community to propose new efficient and effective solutions, particularly using high-performance computing architectures.

Acknowledgements

This research was partially supported by the 2018 VCU Presidential Research Quest Fund and an Amazon AWS Machine Learning Research award.

Author details

Alberto Cano
Virginia Commonwealth University, Richmond, VA, USA

*Address all correspondence to: acano@vcu.edu

References

[1] Stieglitz S, Mirbabaie M, Ross B, Neuberger C. Social media analytics–challenges in topic discovery, data collection, and data preparation. International Journal of Information Management. 2018;**39**:156-168. DOI: 10.1016/j.ijinfomgt.2017.12.002

[2] Batrinca B, Treleaven PC. Social media analytics: A survey of techniques, tools and platforms. AI & Society. 2015;**30**(1):89-116. DOI: 10.1007/s00146-014-0549-4

[3] Emmert-Streib F, Yli-Harja O, Dehmer M. Data analytics applications for streaming data from social media: What to predict? Frontiers in Big Data. 2018;**1**:2. DOI: 10.3389/fdata.2018.00002

[4] Injadat M, Salo F, Nassif AB. Data mining techniques in social media: A survey. Neurocomputing. 2016;**214**:654-670. DOI: 10.1016/j.neucom.2016.06.045

[5] Zatari T. Data mining in social media. International Journal of Scientific and Engineering Research. 2015;**6**(7):152-154

[6] Barbier G, Liu H. Data mining in social media. In: Aggarwal C editor. Social Network Data Analytics. Boston, MA: Springer; 2011:327-352. DOI: 10.1007/978-1-4419-8462-3_12

[7] Feng J, Barbosa LD, Torres V. Systems and methods for social media data mining. United States patent US 9,262,517; 2016

[8] Felt M. Social media and the social sciences: How researchers employ big data analytics. Big Data & Society. 2016;**3**(1):205. DOI: 10.1177/2053951716645828

[9] Flammini A. The rise of social bots. Communications of the ACM. 2016;**59**(7):96-104. DOI: 10.1145/2818717

[10] Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter. 2017;**19**(1):22-36. DOI: 10.1145/3137597.3137600

[11] Jain A, Katkar V. Sentiments analysis of twitter data using data mining. In: International Conference on Information Processing. 2015. pp. 807-810

[12] Grossniklaus M, Scholl MH, Weiler A. Towards adaptive event detection techniques for the twitter social media data stream. IEEE Computer Society Technical Committee on Data Engineering. 2015;**38**(4):116-123

[13] Gaber MM, Zaslavsky A, Krishnaswamy S. Mining data streams: A review. ACM Sigmod Record. 2005;**34**(2):18-26. DOI: 10.1145/1083784.1083789

[14] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Computing Surveys (CSUR). 2014;**46**(4):44. DOI: 10.1145/2523813

[15] Nguyen DT, Jung JJ. Real-time event detection on social data stream. Mobile Networks and Applications. 2015;**20**(4):475-486. DOI: 10.1007/s11036-014-0557-0

[16] Cano A, Krawczyk B. Kappa updated ensemble for drifting data stream mining. Machine Learning. 2019. DOI: 10.1007/s10994-019-05840-z. (In Press)

[17] Krawczyk B, Cano A. Adaptive ensemble active learning for drifting data stream mining. In: Proceedings of the International Joint Conference on Artificial Intelligence; 10-16 August 2019. Macao; 2019. pp. 2763-2771

[18] Cano A. An ensemble approach to multi-view multi-instance learning. Knowledge-Based Systems. 2017;**136**:46-57. DOI: 10.1016/j.knosys.2017.08.022

[19] Roseberry CA. Multi-label kNN classifier with self adjusting memory for drifting data streams. In: Second International Workshop on Learning with Imbalanced Domains: Theory and Applications, LIDTA@PKDD/ECML; 10-14 September 2018. Dublin; 2018. pp. 23-37

[20] Gonzalez-Lopez J, Ventura S, Cano A. Distributed nearest neighbor classification for large-scale multi-label data on spark. Future Generation Computer Systems. 2018;**87**:66-82. DOI: 10.1016/j.future.2018.04.094

[21] Roseberry M, Krawczyk B, Cano A. Multi-label punitive kNN with self-adjusting memory for drifting data streams. ACM Transactions on Knowledge Discovery from Data. 2019;**13**(6):60. DOI: 10.1145/3363573

[22] Cano A. A survey on graphic processing unit computing for large-scale data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018;**8**(1):e1232. DOI: 10.1002/widm.1232

[23] Cano A, Krawczyk B. Evolving rule-based classifiers with genetic programming on GPUs for drifting data streams. Pattern Recognition. 2019;**87**:248-268. DOI: 10.1016/j.patcog.2018.10.024

[24] Cano A, Krawczyk B. Learning classification rules with differential evolution for high-speed data stream mining on GPUs. In: Proceedings of the IEEE Congress on Evolutionary Computation; 8-13 July 2018. Rio de Janeiro, New York: IEEE; 2018. pp. 197-204

[25] Cano A, Zafra A, Ventura S. Parallel evaluation of Pittsburgh rule-based classifiers on GPUs. Neurocomputing. 2014;**126**:45-57. DOI: 10.1016/j.neucom.2013.01.049

[26] Cano A, Ventura S, Cios K. Scalable CAIM discretization on multiple GPUs using concurrent kernels. The Journal of Supercomputing. 2014;**69**(1):273-292. DOI: 10.1007/s11227-014-1151-8

[27] Cano A, Zafra A, Ventura S. Solving classification problems using genetic programming algorithms on GPUs. In: 5th International Conference on Hybrid Artificial Intelligent Systems (HAIS); 23-25 May 2010. Wroclaw; 2010. pp. 17-26

[28] Cano A, Garcia C, Ventura S. Extremely high-dimensional optimization with MapReduce: Scaling functions and algorithm. Information Sciences. 2017;**415-416**:110-127. DOI: 10.1016/j.ins.2017.06.024

[29] Gonzalez-Lopez J, Ventura S, Cano A. Distributed selection of continuous features in multi-label classification using mutual information. IEEE Transactions on Neural Networks and Learning Systems. 2019. DOI: 10.1109/TNNLS.2019.2944298. (In Press)

[30] Gonzalez-Lopez J, Ventura S, Cano A. Distributed multi-label feature selection using individual mutual information measures. Knowledge-Based Systems. 2019. DOI: 10.1016/j.knosys.2019.105052. (In Press)

[31] Krawczyk B, Cano A. Online ensemble learning with abstaining classifiers for drifting and noisy data streams. Applied Soft Computing. 2018;**68**:677-692

[32] Korycki, Cano A, Krawczyk B. Active learning with abstaining classifiers for imbalanced drifting data streams. In: Proceedings of the IEEE International Conference on BigData; 9-12 December. Los Angeles, New York: IEEE; 2019. p. 2019

[33] Wu Y, Cao N, Gotz D, Tan YP, Keim DA. A survey on visual analytics of social media data. IEEE Transactions on Multimedia. 2016;**18**(11):2135-2148. DOI: 10.1109/TMM.2016.2614220

[34] Grimmer J. We are all social scientists now: How big data, machine learning, and causal inference work together. Political Science & Politics. 2015;**48**(1):80-83. DOI: 10.1017/S1049096514001784

[35] Tsou M. Research challenges and opportunities in mapping social media and big data. Cartography and Geographic Information Science. 2015;**42**(sup 1):70-74. DOI: 10.1080/15230406.2015.1059251

[36] Bello-Orgaz G, Jung JJ, Camacho D. Social big data: Recent achievements and new challenges. Information Fusion. 2016;**28**:45-59. DOI: 10.1016/j.inffus.2015.08.005