ACTIVE ONLINE LEARNING FOR SOCIAL MEDIA ANALYSIS TO SUPPORT CRISIS MANAGEMENT

S.Devarajan[1], K.Pushpavathi[2]

[1]*Dept of Computer Science, School of Arts and Science, Vinayaka Mission's Research Foundation, Chennai, India.*
Email:devaraansengeni98@gmail.com

[2]*Associate Professor, Dept of Computer Science, School of Arts and Science, Vinayaka Mission's Research Foundation, Chennai, India.*
Email: pushpavathi@avit.ac.in

ABSTRACT

People use social media (SM) to describe and discuss different situations they are involved in, like crises. It is therefore worthwhile to exploit SM contents to support crisis management, in particular by revealing useful and unknown information about the crises in real-time. Hence, we propose a novel active online multiple-prototype classifier, called AOMPC. It identifies relevant data related to a crisis. AOMPC is an online learning algorithm that operates on data streams and which is equipped with active learning mechanisms to actively query the label of ambiguous unlabeled data. The number of queries is controlled by a fixed budget strategy. Typically, AOMPC accommodates partly labeled data streams. AOMPC was evaluated using two types of data: (1) synthetic data and (2) SM data from Twitter related to two crises, Colorado Floods and Australia Bushfires. To provide a thorough evaluation, a whole set of known metrics was used to study the quality of the results. Moreover, a sensitivity analysis was conducted to show the effect of AOMPC's parameters on the accuracy of the results. A comparative study of AOMPC against other available online learning algorithms was performed. The experiments showed very good behavior of AOMPC for dealing with evolving, partly-labeled data streams.

Keywords: Online Learning, Multiple Prototype Classification, Active Learning, Social Media, Crisis Management

INTRODUCTION

The primary task of crisis management is to identify specific actions that need to be carried out before (prevention, preparedness), during (response), and after (recovery and mitigation) a crisis occurred. In order to execute these tasks efficiently, it is helpful to use data from various sources including the public as witnesses of emergency events. Such data would enable emergency operations centers to act and organize the rescue and response. In recent years, a number of research studies have investigated the use of social media as a source of information for efficient crisis management. A selection of such studies, among others, encompasses Norway Attacks, Minneapolis Bridge Collapse, California Wildfire , Colorado Floods , and Australia Bushfires. The extensive use of SM by people forces (re)thinking the public engagement in crisis management regarding the new available technologies and resulting opportunities.

Related Works

"Incremental Learning Based on Growing Gaussian Mixture Models,"

Incremental learning aims at equipping data-driven systems with self-monitoring and self-adaptation mechanisms to accommodate new data in an online setting. The resulting model underlying the system can be adjusted whenever data become available. The present paper proposes a new incremental learning algorithm, called 2G2M, to learn Growing Gaussian Mixture Models. The algorithm is furnished with abilities (1) to accommodate data

online, (2) to maintain low complexity of the model, and (3) to reconcile labeled and unlabeled data. To discuss the efficiency of the proposed incremental learning algorithm, an empirical evaluation is provided.

"Incremental Learning with Multi-Level Adaptation,"

Self-adaptation is an inherent part of any natural and intelligent system. Specifically, it is about the ability of a system to reconcile its requirements or goal of existence with the environment it is interacting with, by adopting an optimal behavior. Self-adaptation becomes crucial when the environment changes dynamically over time. In this paper, we investigate self-adaptation of classification systems at three levels: (1) natural adaptation of the base learners to change in the environment, (2) contributive adaptation when combining the base learners in an ensemble, and (3) structural adaptation of the combination as a form of dynamic ensemble. The present study focuses on neural network classification systems to handle a special facet of self-adaptation, that is, incremental learning (IL). With IL, the system self-adjusts to accommodate new and possibly non-stationary data samples arriving over time. The paper discusses various IL algorithms and shows how the three adaptation levels are inherent in the system's architecture proposed and how this architecture is efficient in dealing with dynamic change in the presence of various types of data drift when applying these IL algorithms.

"Learning with Incrementality,"

Learning with adaptivity is a key issue in many nowadays applications. The most important aspect of such an issue is incremental learning (IL). This latter seeks to equip learning algorithms with the ability to deal with data arriving over long periods of time. Once used during the learning process, old data is never used in subsequent learning stages. This paper suggests a new IL algorithm which generates categories. Each is associated with one class. To show the efficiency of the algorithm, several experiments are carried out.

"Prototype-based Models in Machine Learning,"

An overview is given of prototype-based models in machine learning. In this framework, observations, i.e., data, are stored in terms of typical representatives. Together with a suitable measure of similarity, the systems can be employed in the context of unsupervised and supervised analysis of potentially high-dimensional, complex datasets. We discuss basic schemes of competitive vector quantization as well as the so-called neural gas approach and Kohonen's topology-preserving self-organizing map. Supervised learning in prototype systems is exemplified in terms of learning vector quantization. Most frequently, the familiar Euclidean distance serves as a dissimilarity measure. We present extensions of the framework to nonstandard measures and give an introduction to the use of adaptive distances in relevance learning.

"Learning Similarity Metrics for Event Identification in Social Media,"

Social media sites (e.g., Flickr, YouTube, and Face book) are a popular distribution outlet for users looking to share their experiences and interests on the Web. These sites host substantial amounts of user-contributed materials (e.g., photographs, videos, and textual content) for a wide variety of real-world events of different type and scale. By automatically identifying these events and their associated user-contributed social media documents, which is the focus of this paper, we can

enable event browsing and search in state-of-the-art search engines. To address this problem, we exploit the rich "context" associated with social media content, including user-provided annotations (e.g., title, tags) and automatically generated information (e.g., content creation time).

II. EXISTING SYSTEM

In particular, online feature selection mechanisms were devised as well, so that SM data streams can be accommodated continuously and incrementally. It is interesting to note that people from emergency departments (e.g., police forces) already use SM to gather, monitor, and to disseminate information to inform the public . Hence, we propose a learning algorithm, AOMPC, that relies on active learning to accommodate the user's feedback upon querying the item being processed. Since AOMPC is a classifier, the query is related to labeling that item. The primary goal in using user-generated contents of SM is to discriminate valuable information from irrelevant one.

DISADVANTAGES

- Describe an analysis approach based on visual analytics for combining information from different sources with a specific focus on multilingual issues.
- If the social media items consist of two parts, the body of the message and the geo-location that indicates where the message was issued in terms of coordinates, then we apply a combined distance measure.

III. PROPOSED SYSTEM

We propose classification as the discrimination method. The classifier plays the role of a filtering machinery. With the help of the user, it recognizes the important SM items (e.g., tweets), that are related to the event of interest. The selected items are used as cues to identify sub-events. Note that an event is the crisis as such, while sub-events are the topics commonly discussed (i.e., hotspots like flooding, collapsing of bridges, etc. in a specific area of a city) during a crisis. These sub-events can be identified by aggregating the messages posted on SM networks describing the same specific topic. We propose a Learning Vector Quantization (LVQ)- like approach based on multiple prototype classification. The classifier operates online to deal with the evolving stream of data.

ADVANTAGES

- The advantage of AOMPC compared to the other algorithms is the continuous processing of data streams and incremental update of knowledge, where the existing prototypes act as memory for the future.
- The advantage of our algorithm compared to the others is the transferred knowledge from one batch to the next creating a continuous view on the arriving data.
- In particular, online feature selection mechanisms were devised as well, so that SM data streams can be accommodated continuously and incrementally.

IV Module Description

Crisis management:

The primary task of crisis management is to identify specific actions that need to be carried out before (prevention, preparedness), during (response), and after (recovery and mitigation) a crisis occurred. In order to execute these tasks efficiently, it is helpful to use data from various sources including the public as witnesses of emergency events. Such data would enable emergency operations centers to act and organize the rescue and response. In recent years, a number

of research studies have investigated the use of social media as a source of information for efficient crisis management. A selection of such studies, among others, encompasses Norway Attacks, Minneapolis Bridge Collapse, California Wildfire , Colorado Floods, and Australia Bushfires.The extensive use of SM by people forces (re)thinking the public engagement in crisis management regarding the new available technologies and resulting opportunities.

Multiple Prototype Classification and LVQ Classification

A prototype-based classification approach operates on data items mapped to a vector representation (e.g., vector space model for text data). Data points are classified via prototypes considering similarity measures. Prototypes are adapted based on items related/similar to them. A Rocchio classifier is an example of a single prototype-based classifier. It distinguishes between two classes, e.g., “relevant” and “irrelevant”. In real world-scenarios, due to the nature of the data, it is often not possible to describe the data with a single prototype-based classifier. Multiple prototype classifiers (i.e., several prototypes) are needed.

Online Learning and Active Learning (with Budget Planning)

Online learning receives data items in a continuous sequence and processes them once to classify them accordingly ,Use Growing Gaussian Mixture Models for online classification. Compared to the algorithm proposed in this work, there is a difference in adapting the learning rate and representing the prototypes. Use multiple prototypes representing an event. New incoming items are assigned to the most similar events (by using an offline-trained SVM) or otherwise new events are created. Another important topic in streaming analysis is active learning to improve results of classification with an amount of labeled data actively asked by the system.

Active online multiple prototype classifier (aompc)

Due to the fact that SM data is noisy, it is important to identify relevant SM items for the crisis situation at hand. The idea is to find an algorithm that performs this classification and also handles ambiguous items in a reasonable way. Ambiguous denotes items where a clear classification is not possible based on the current knowledge of the classifier. The knowledge should be gained by asking an expert for feedback. The algorithm should be highly self-dependent, by asking the expert only labels for a limited number of items.

V. CONCLUSION

This project presents a streaming analysis framework for distinguishing between relevant and irrelevant data items. It integrates the user into the learning process by considering the active learning mechanism. We evaluated the framework for different datasets, with different parameters and active learning strategies. We considered synthetic datasets to understand the behavior of the algorithm and real-world social media datasets related to crises. We compared the proposed algorithm, AOMPC, against many existing algorithms to illustrate the good performance under different parameter settings.

FUTURE WORK

The algorithm can be extended to overcome many issues, for instance by considering: dynamic budget, dynamic deletion of stale clusters, and generalization to handle non-contiguous class distribution.

REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Semantics + Filtering + Search = Twitcident. Exploring Information in SocialWeb Streams," in Proc. of the 23rdACMConf. on Hypertext and Social Media. ACM, 2012, pp. 285–294.

[2] U. Ahmad, A. Zahid, M. Shoaib, and A. AlAmri, "Harvis: An integrated social media content analysis framework for youtube platform," Information Systems, vol. 69, pp. 25 – 39, 2017.

[3] G. Backfried, J. Gollner, G. Qirchmayr, K. Rainer, G. Kienast, G. Thallinger, C. Schmidt, and A. Peer, "Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project," in Eur. Intel. and Security Informatics Conf., Aug 2013, pp. 143–146.

[4] BBC News Europe. (2012, Aug.) England Riots: Maps and Timeline. [Online]. Available: http://www.bbc.co.uk/news/ uk-14436499

[5] H. Becker, M. Naaman, and L. Gravano, "Learning Similarity Metrics for Event Identification in Social Media," in Proc. of the Third ACM Int'l Conf. on Web Search and Data Mining, ser. WSDM '10. NY, USA: ACM, 2010, pp. 291–300.

[6] J. Bezdek, T. Reichherzer, G. Lim, and Y. Attikiouzel, "Multiple- Prototype Classifier Design," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 28, no. 1, pp. 67– 79, Feb 1998.

[7] M. Biehl, B. Hammer, and T. Villmann, "Prototype-based Models in Machine Learning," Wiley Interdisciplinary Reviews: Cognitive Science, vol. 7, no. 2, pp. 92–111, 2016.

[8] A. Bouchachia, "Learning with Incrementality," in Proc. of the Int'l Conf. on Neural Information Processing, 2006, pp. 137–146.

[9] ——, "Incremental Learning with Multi-Level Adaptation," Neurocomputing, vol. 74, no. 11, pp. 1785–1799, 2011.

[10] A. Bouchachia and C. Vanaret, "Incremental Learning Based on Growing Gaussian Mixture Models," in 10th Int'l Conf. on Machine Learning and Applications and Workshops (ICMLA), vol. 2, Dec 2011, pp. 47–52.