ACTIVE ONLINE LEARNING FOR SOCIAL MEDIA ANALYSIS TO SUPPORT CRISIS MANAGEMENT

Mr. Raj Mohammad¹, T. Rishik Chandra², M. S. Srilekha³, N. Dilip Kumar⁴, T. Nikhil⁵

¹, ², ³, ⁴ Department of Computer Science And Engineering, Gitam University, Hyderabad, India.
⁵ Assistant Professor, Department of Computer Science And Engineering, Gitam University, Hyderabad, India.

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.

1. 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. Our previous work on SM in emergency response focused on offline and online clustering of SM messages. The offline clustering approach was applied to identify sub-events (specific hotspots) from SM data of a crisis for an after-the-fact analysis. Online clustering was used to identify sub-events that evolve over time in a dynamic way. 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. 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. The algorithm, named active online multiple prototype classifier (AOMPC), uses unlabeled and labeled data which are tagged through active learning. Data items which fall into ambiguous regions are selected for labeling by the user. The number of queries is controlled by a budget. The requested items help to direct the AOMPC classifier to a better discriminatory capability. While AOMPC can be applied to any streaming data, here we consider in particular SM data.

2. LITERATURE SURVEY

Data Mining and Machine Learning

An introduction is given to the use of prototype-based models in supervised machine learning. The main concept of the framework is to represent previously observed data in terms of so-called prototypes, which reflect typical properties of the data. Together with a suitable, discriminative distance or dissimilarity measure, prototypes can be used for the classification of complex, possibly high-dimensional data. We illustrate the framework in terms of the popular Learning Vector Quantization (LVQ). Most frequently, standard Euclidean distance is employed as a distance measure. We discuss how LVQ can be equipped with more general dissimilarity measures. Moreover, we introduce relevance learning as a tool for the data-driven optimization of parameterized distances. Prototype-based models constitute a very successful family of methodological approaches in machine learning (see e.g. Kohonen 1990, Hastie et al. 2009, Biehl et al. 2016, Biehl et al. 2009). They are appealing for a number of reasons: The extraction of information from previously observed data in terms of typical representatives, so-called prototypes, is particularly transparent and intuitive, in contrast to many, more black-box like systems. The same is true for the working phase, in which novel data are compared with the prototypes by use of a suitable (dis-)similarity or distance measure. Prototype systems are frequently employed for the unsupervised analysis of complex data sets, aiming at the detection of underlying structures, such as clusters or hierarchical relations, see for instance (Biehl et al. 2009). Competitive Vector Quantization or the well-known K-means algorithm are prominent examples for the use of prototypes in the context of unsupervised learning (Duda et al. 2001, Hastie et al. 2009). Potential goals of supervised machine learning are the assignment of data to categories in classification problems, or their characterization by a continuous target value in regression tasks. In both cases, the learning or training process relies on the availability of labeled example data. The aim is to extract relevant information and represent it in terms of a hypothesis for the unknown target function. The obtained hypothesis can then be applied to novel data in a working phase.
Learning Similarity Metrics for Event Identification in Social Media

Social media sites (e.g., Flickr, YouTube, and Facebook) 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). Using this rich context, which includes both textual and non-textual features, we can define appropriate document similarity metrics to enable online clustering of media to events. As a key contribution of this paper, we explore a variety of techniques for learning multi-feature similarity metrics for social media documents in a principled manner. We evaluate our techniques on large-scale, real-world datasets of event images from Flickr. Our evaluation results suggest that our approach identifies events, and their associated social media documents, more effectively than the state-of-the-art strategies on which we build. The ease of publishing content on social media sites brings to the Web an ever increasing amount of content captured during—and associated with—real-world events. Sites like Flickr, YouTube, Facebook and others host user-contributed content for a wide variety of events. These range from widely known events, such as presidential inaugurations, to smaller, community-specific events, such as annual conventions and local gatherings. By automatically identifying these events and their associated user-contributed social media documents, which is the focus of this paper, we can enable powerful local event browsing and search, to complement and improve the local search tools that Web search engines provide. In this paper, we address the problem of how to identify events and their associated user-contributed documents over social media sites. In one scenario, consider a person who is thinking of attending “All Points West,” an annual music festival that takes place in early August in Liberty State Park, New Jersey. Prior to purchasing a ticket, this person could search the Web for relevant information, to make an informed decision. Unfortunately, Web search results are far from revealing for this relatively minor event: the event’s website contains marketing materials, and traditional news coverage is low. Overall, these

Multiple-Prototype Classifier Design

Five methods that generate multiple prototypes from labeled data are reviewed. Then we introduce a new sixth approach, which is a modification of Chang’s method. We compare the six methods with two standard classifier designs: the 1-nearest prototype (1-np) and 1-nearest neighbor (1-nn) rules. The standard of comparison is the resubstitution error rate; the data used are the Iris data. Our modified Chang’s method produces the best consistent (zero errors) design. One of the competitive learning models produces the best minimal prototypes design (five prototypes that yield three resubstitution errors).

3. EXISTING SYSTEM

Our previous work on SM in emergency response focused on offline and online clustering of SM messages. The offline clustering approach was applied to identify sub-events (specific hotspots) from SM data of a crisis for an after-the-fact analysis. Online clustering was used to identify sub-
events that evolve over time in a dynamic way. In particular, online feature selection mechanisms were devised as well, so that SM data streams can be accommodated continuously and incrementally. \textbf{Disadvantages:}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.

4. PROPOSED SYSTEM

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. \textbf{Advantages:} 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. Therefore, we propose an original approach that combines different aspects - such as online learning and active learning - to build a hybrid classifier, AOMPC. AOMPC learns from both, labeled and unlabeled data, in a continuous and evolving way. In this context, AOMPC is designed to distinguish between relevant and irrelevant SM data related to a crisis situation in order to identify the needs of individuals affected by the crisis.

5. METHODOLOGY

5.1 ACTIVE LEARNING

AOMPC relies on active learning. It implies the intervention of a user in some situations to enhance its effectiveness in terms of identifying relevant data and the related event in the SM stream of data. The user is asked to label an item if there is a high uncertainty about the classification as to whether it is relevant or irrelevant. The classifier assigns then the item (be it actively labeled or unlabeled) to the closest cluster or uses it to create a new cluster. A cluster - in this case – represents either relevant (i.e., specific information about the crisis of interest) or irrelevant information (i.e., not related to the crisis).

5.2 AOMPC

Related to the active learning part. The algorithm starts by checking whether the new input item lies in the uncertainty region between the relevant and irrelevant prototypes and whether there is enough budget for labeling this item. More details follow in the next section.

5.3 BUDGET

The idea of active learning is to ask for user feedback instead of labeling the incoming data item automatically. To limit the number of interventions of the user, a so called budget, is defined. Budget can be understood as the maximum number of queries to the user. We adapt the method presented in to implement active learning in the context of online multiple prototype classification.
5.4 DATA ITEMS TO QUERY

In active learning, before querying the label, one has to decide which data points to query. Obviously one has to find those points, for which the classifier is not confident about the assignment decision.

5.5 FEASIBILITY STUDY

Preliminary investigation examine project feasibility, the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running system. All system is feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:

5.6 ECONOMIC FEASIBILITY

A system can be developed technically and that will be used if installed must still be a good investment for the organization. In the economical feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs.

The system is economically feasible. It does not require any addition hardware or software. Since the interface for this system is developed using the existing resources and technologies available at NIC, There is nominal expenditure and economical feasibility for certain.

5.7 OPERATIONAL FEASIBILITY

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization’s operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits. The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

5.8 TECHNICAL FEASIBILITY

The technical issue usually raised during the feasibility stage of the investigation includes the following:
Does the necessary technology exist to do what is suggested? Do the proposed equipments have the technical capacity to hold the data required to use the new system?

- Will the proposed system provide adequate response to inquiries, regardless of the number or location of users?
- Can the system be upgraded if developed create, establish and maintain a workflow among various entities in order to facilitate all concerned users in their various capacities or roles. Permission to the users would be granted based on the roles specified. Therefore, it provides the technical guarantee of accuracy, reliability and security. The software and hard requirements for the development of this project are not many and are already available in-house at NIC or are available as free as open source?
- Are there technical guarantees of accuracy, reliability, ease of access and data security?

Earlier no system existed to cater to the needs of ‘Secure Infrastructure Implementation System’. The current system developed is technically feasible. It is a web based user interface for audit workflow at NIC-CSD. Thus it provides an easy access to the users. The database’s purpose is to

5.9. SOFTWARE REQUIREMENTS
Operating System: Windows 7
User Interface : HTML, CSS
Client-side Scripting: JavaScript
Programming Language : Java
Web Applications: JDBC, Servlets, JSP
IDE/Workbench: My Eclipse 8.6
Database: Oracle 11g
Server Deployment: Tomcat 7.0

5.10. HARDWARE REQUIREMENTS
Processor : Intel core i3 or above
Hard Disk : 500GB or more
6. RESULTS

Fig. 1 Opening web page of the Active Online Learning For Social Media Analysis To Support Crisis Management Website.

Fig. 2 Admin sign up page
Fig. 3 Admin home page. Where admin can accept the requests of the user.

Fig. 4 Bar graph of the crisis.

Part 2: User Operations
Fig. 5 Login through the user credentials

Fig. 6 User searching tweets.
7. CONCLUSIONS

Hence we 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. 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

2. Han-wen Chang, Yu-chin Tai And Jane Yung-jen Hsu, “Context-aware Taxi Demand Hotspots Prediction”, In Int. J. Business Intelligence And Data Mining, Vol. 5, No. 1, 2010