ADOVA: Anomaly Detection in Online and Virtual spAces

C. David Emele
RCUK dot.rural DE Hub
University of Aberdeen, UK
c.emele@abdn.ac.uk

Vitalij Spakov
Computer Science Dept.
University of Aberdeen, UK
v.spakov@abdn.ac.uk

Wei Pang
Computer Science Dept.
University of Aberdeen, UK
w.pang@abdn.ac.uk

Jone Bone
Sociology Department
University of Aberdeen, UK
j.bone@abdn.ac.uk

George Coghill
Computer Science Dept.
University of Aberdeen, UK
g.coghil@abdn.ac.uk

ABSTRACT

Online and virtual spaces comprise a myriad of ad-hoc networks and online communities. Such communities are composed of smart devices, agents, systems and people who seek to interact in one way or another. We argue that the task of detecting anomalies in such settings is non-trivial. The complexity is further compounded since there is no clear cut definition/specification of what normal behaviour is, and how far out an outlier should be before it is detected as an anomaly. This is often the case with online and virtual spaces as there is little or no regulation of the interactions between the various players in online communities. Hence, detecting anomalous behaviour in such settings poses a huge challenge. In this paper, we investigate how evolutionary clustering could be exploited to support decision makers, designers and data scientists in the autonomous detection of anomalies in online and virtual spaces. We present preliminary ideas in tackling this issue using a freeform online social media community (Twitter) and explore how emerging patterns and trends could help identify clusters of players (or normal behaviour) and, conversely, anomalies.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Human Factors

Keywords

evolutionary clustering, anomaly, outliers, twitter, social media, online communities, agent technology, virtual organisations

1. INTRODUCTION

Recent advancements in technology and the Internet present new spaces for communities to interact. Such spaces (regarded here as online and virtual spaces) offer new opportunities to engage in pervasive communication between many players in a short time interval. Many social networking platforms have been developed which enable users to engage socially with their environment and communities. Examples of such platforms include Twitter, Facebook, Instagram, and LinkedIn. Through these platforms, users connect with other users to form ad-hoc organisations. Such organisations often operate with little or no rules to guide their interactions with one another. Because of the transient nature of these ad-hoc communities, detecting abnormal behaviours (often referred to as anomaly or outlier) is a difficult task.

Anomaly is a common phenomenon in our everyday lives, and could present grievious repercussions such as potential security threat to lives and property. There is strong evidence that detecting anomaly is a challenge in real-life real-time interactions, particularly in social media. This is largely as a result of the amount of data produced in online social interactions, the time and intelligence required to trawl through massive quantities of data produced every second and the heterogeneity of the network. An even further complication is the fact that communities and organisations formed on social networks are often open, informal and lack rigid rules for conformity - in order words, little or no control. In such scenarios, no-one defines what an acceptable behaviour is and so it is difficult to detect abnormal behaviour when normal behaviour is not clearly defined. Nevertheless, it is important to detect anomalous behaviour and possibly respond to it early enough so as to mitigate imminent threats.

We acknowledge that anomalies could be accidental or deliberate and may be innocuous or harmful. Whatever the case, it is important to develop tools that could help to detect anomalies when they occur, especially in online and virtual spaces, where there is neither regulation nor supervision. To date, many anomaly detection tools and algorithms have been developed, however they are often developed for offline settings and so would struggle to detect anomalies in online and virtual spaces. Some of the work that has been done in detecting anomaly in online settings often fail to deal with scenarios where normal behaviour is not explicitly known, and often do not consider the ad-hoc nature of virtual spaces (where players can join and leave a community as they wish).

In this paper, we propose the use of evolutionary clustering mechanisms to support data analysts and decision makers in detecting anomalies in ad-hoc and unregulated communities. The rest of this paper is formatted as follows:

Section 2 discusses as background some of the existing approaches that have been used in anomaly detection in the literature. Section 3 describes our evolutionary clustering approach and how it can aid anomaly detection in online and virtual communities (e.g. social networks). Section 4 presents preliminary results of our evaluation, while Section 5 concludes with a brief discussion of the results as well as future directions.

2. BACKGROUND

This section presents a brief background on anomaly detection in general. It often involves a set of observations over a time period such that $X = \{x_1, x_2, x_3, \cdots x_n\}$. Social media data is largely noisy and so differentiating noise and anomaly could be a daunting task. Being able to automatically detect anomalies in such noisy, ad-hoc environment presents huge research challenges. Although anomaly detection has been extensively researched in the literature, existing approaches do not adequately detect anomalies in online social networks due to the uniqueness of the environment in terms of noise, little or no regulation, and "ad-hocness".

Anomaly detection is a well-studied problem in the literature (see [11]) and several techniques including statistical ([13]), graph-based ([8]), network learning ([9, 10]) and clustering ([5, 12]) have been applied. However, to the best of our knowledge, evolutionary clustering has not been considered/applied in anomaly detection. Furthermore, it has not been considered in online social networks where network configurations and composition involve heterogenous connections that are constantly changing. The ad-hoc and "unregulated" nature of social interactions in online communities pose interesting challenges which evolutionary clustering could help to resolve (as presented in Section 3). As a background, we discuss related research that has looked into some of the temporal issues involved in anomaly detection including online topic detection, classification, tracking and load balancing. For example, some work in machine learning allow models learned in classification settings to evolve over time but with penalties for errors and shifts introduced at each timestep ([7]). It is, nevertheless, unclear how such algorithms could be employed in ad-hoc, fast changing, heterogenous, unregulated and unsupervised learning settings.

Some online document clustering research has explored temporal aspects, in which a time series of documents are examined sequentially to detect novelty regarding certain features of the network ([2]). Research in machine learning, data mining, statistics, and cloud computing has explored a number of approaches for clustering time-series data. Temporal correlation is perhaps the best-known approach to time-series similarity computation [4, 6]. Statistical approaches have utilised probabilistic models for online document clustering ([13]). Clustering has been applied to automatic discovery and retrieval of topically related material in data streams, and to detect novel events from a temporally ordered collection of news stories ([1]). However, the primary objective of topic detection and tracking is to detect new events in a timeline using approaches such as clustering, and not to produce clusterings that incorporate history into the clustering objective. One clustering mechanism that is known to incorporate history into the clustering objective is evolutionary mechanism. In this light, our research explores evolutionary clustering to detect anomalous behaviours in social network settings.

3. EVOLUTIONARY CLUSTERING IN ONLINE AND VIRTUAL SETTINGS

Evolutionary clustering presents benefits that make it a robust framework for analysing interactions in online and virtual spaces. For example, evolutionary clustering enables clusters to have smoother and more natural transitions over time since new information is added to existing data, which could cause cluster centres to shift. Secondly, evolutionary algorithm enables a new cluster to resemble the cluster in the previous timestep, if at all possible, which we refer to as consistency. On th other hand, if the new data is not representative of the cluster population then it registers the new data as an anomaly. Regarding noise, evolutionary algorithm is robust against noise because previous data points are considered in generating new clusters.

We proceed by presenting our evolutionary clustering algorithm, which has strong roots in the algorithm proposed by Chakrabarti *et al.* [3]. Let t denote a timestep and j be the index of a cluster such that c_j^t denotes a cluster centre j at time t. A matching cluster is represented as $c_{f(j)}^{t-1}$, where $t-1$ depicts a previous timestep and $f(j)$ denotes a cluster matching function that matches a given cluster with the most similar cluster in the previous timestep. We calculate the relative cluster size, denoted by γ, as follows:

$$\gamma = \frac{n_j^t}{(n_j^t + n_{f(j)}^{t-1})} \tag{1}$$

The relative cluster size is an important statistic for calculating a new centre for a given cluster. Another important variable is the change parameter (*cp*, for short), which specifies the trade off between two cluster centres. Change parameter is user defined and ranges from 0 to 1 (that is, $0 < cp < 1$). Putting it all together, we define a centre recalculation function, denoted as $g(c, t, j')$ as follows:

$$g(c, t, j') = (1 - \gamma) \times cp \times c_{f(j)}^{t-1} + \gamma \times (1 - cp) \times c_j^t \tag{2}$$

The centre recalculation function (depicted in Equation 2) calculates a new center location that has to be between a current cluster center retrieved from non-evolutionary clustering, and the one matched from a previous clustering while taking into account the relative sizes of both clusters and change parameter values defined by the user.

Having defined the variables that our evolutionary clustering mechanism depends on, we now turn our attention to describing the algorithm. Firstly, we present the evolutionary clustering algorithm (see Algorithm 1) which would require the centre recalculation algorithm (see Algorithm 2). The centre recalculation function is the most important part of the evolutionary clustering algorithm. It is responsible for relocating cluster centres, and reassigning the current input points to new clusters respectively. In line 8 of Algorithm 2 we use greedy comparison to find a previous cluster instance to match with a given cluster; hence, a given cluster will be matched with the closest cluster from a previous clustering step. Our evolutionary clustering mechanism is robust to handle multidimensional data as long as the cluster centres contain numeric coordinates, and this is typical of social media data (as will be shown in the preliminary results of our evaluation).

Algorithm 1 Evolutionary Clustering Algorithm.

INPUT : *InputData* (that is, time series data)
OUTPUT : *Output* (that is, cluster centre locations and cluster sizes)
Initialise $Output \leftarrow \emptyset$
Split *InputData* by timesteps
for each timestep in *InputData* **do**
  CurrentData ← data in first timestep of *InputData*
  InputData ← Remove *CurrentData*
  Perform k-means clustering on *CurrentData*
  if timestep > 1 **then**
    Recalculate clustering centres by using historical data as reference (see Algorithm 2)
  else
    Do Nothing
  end if
end for
Generate *Output* using modified clustering information
Save *CurrentData* clustering centre locations and cluster sizes for historical reference
return *Output*

Algorithm 2 Centre Recalculation Algorithm.

INPUT : *previousTimeClusters*
INPUT : *currentTimeClusters*
INPUT : *dataPoints*
OUTPUT : *clusterCentres* (that is, a list of cluster centre locations)
Initialise $clusterCentres \leftarrow \emptyset$
dimensionCount ← size of centre dimensions of *currentTimeClusters*
for each *center* in *currentTimeClusters* **do**
  find previous instance of *centre* from *previousTimeClusters*
  apply centre recalculation function (see Equation 2)
  save modified *centre* in *currentTimeClusters*
  add *centre* to the list of cluster centres (that is, *clusterCentres*)
end for
for each *point* in *dataPoints* **do**
  find the shortest distance to a cluster centre in *currentTimeClusters*
  assign *point* to the closest cluster found
end for
return *Output*

4. EVALUATION

In evaluating our work, we developed a social network analysis tool which used Twitter social media platform as an online and virtual space that enabled the creation of ad-hoc online communities around a given topic or event (e.g., the 2014 scottish referendum - #indyref). We downloaded 10,000 timestamped tweets relating to the scottish referendum and applied evolutionary clustering mechanism to the collection of tweets. The tweets gathered covered the period from 9th to 25th March 2015. Though the Scottish Referendum took place in September 2014, "#indyref" is still s fairly popular hashtag.

In our preliminary analysis, we attempted to cluster the interactions in the online community based on the number of retweets of the various tweets gathered (that is, retweetCount). One of our objectives was to see how the clusters evolved over time. Results show that in timestep 3 (illustrated in Figure 1, some patterns were already visible from the data, and by timestep 9 (depicted in Figure 2) 4 clusters had emerged. In both graphs we could spot the anomalies clearly as they do not fit into any of the clusters.

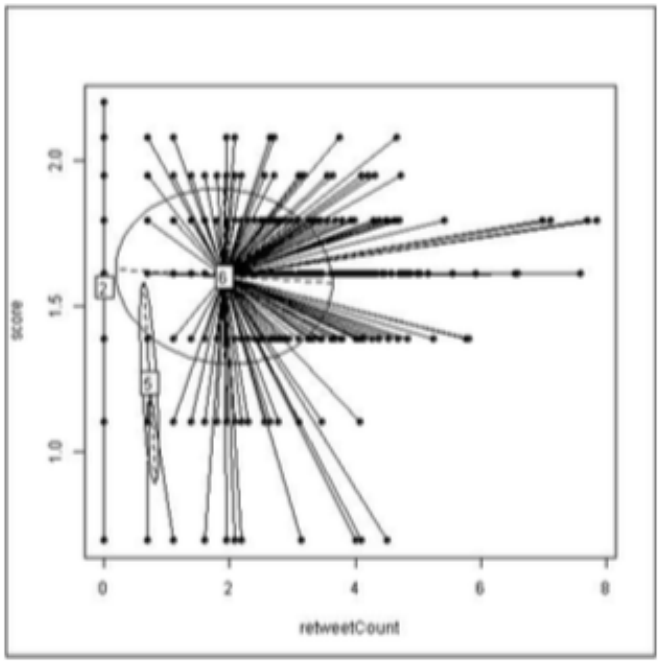

Figure 1: The plot at the third timestep.

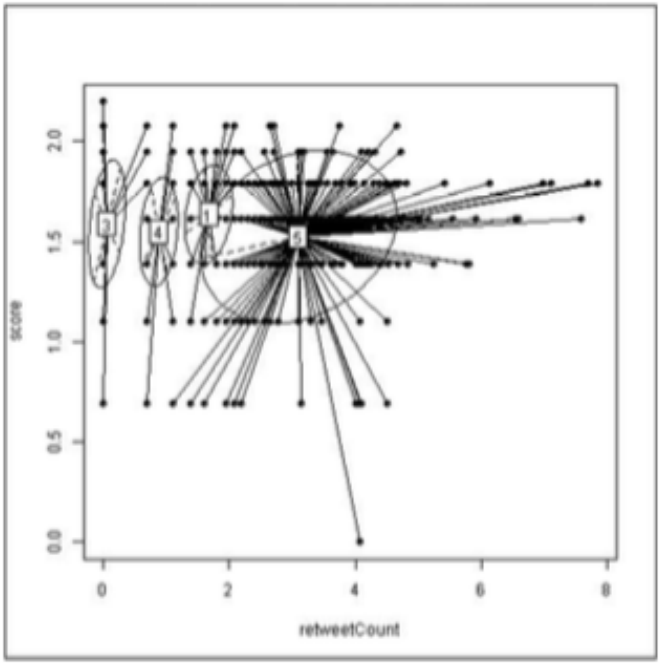

Figure 2: The plot at the ninth timestep.

5. CONCLUSIONS

This paper presented a prototype evolutionary clustering approach for detecting anomaly in online and virtual spaces. We discussed how our approach could support decision makers, designers and data scientists in the detection of anomalies in online and virtual spaces. We used a freeform online social media community (Twitter) to explore how emerging patterns within a dataset could help identify clusters of players (or normal behaviour) and, conversely, anomalies. Preliminary results of our work show that our evolutionary clustering approach is robust to handle multidimensionality in ad-hoc and unregulated settings presented by social media communities. We are currently engaging in more detailed analyses of the results of our experiments. In the future, we plan to investigate additional features that will enable faster anomaly detection in online and virtual spaces.

Acknowledgments

The research described here is supported by the award made by the Research Councils UK Digital Economy programme to the dot.rural digital economy research hub, at University of Aberdeen, UK (award reference: EP/G066051/1).

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[2] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L. Saul, and B. Scholkopf, editors, *The Semantic Web, ISWC 2002*, volume 16 of *Advances in Neural Information Processing Systems*. 2004.

[3] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 554–560. ACM, 2006.

[4] C. Chatfield. *The Analysis of Time Series*. Chapman and Hall, New York, NY, USA, 1984.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, NY, USA, 2000.

[6] M. Gupta, J. Gao, Y. Sun, and J. Han. Community trend outlier detection using soft temporal pattern mining. In P. Flach, T. Bie, and C. Nello, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 16, pages 692–708. Springer, Berlin Heidelberg, 2012.

[7] M. Herbster and M. K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

[8] C. C. Noble and D. J. Cook. Graph-based anomaly detection. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636. ACM, 2003.

[9] S. Pandit, D. Chau, S. Wang, and C. Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th International Conference on World Wide Web*, pages 201–210. ACM, 2007.

[10] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina. Web graph similarity for anomaly detection. *Journal of Internet Service Applications*, 1:19–30, 2001.

[11] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang. Anomaly detection in online social networks. *Social Networks*, 39(0):62 – 70, 2014.

[12] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition, 2005.

[13] J. Zhang, Z. Ghahramani, and Y. Yang. A probabilistic model for online document clustering with applications to novelty detection. In *Proceedings of Advances in Neural Information Processing Systems*, 2005.