Flu Detector - Tracking Epidemics on Twitter

Vasileios Lampos, Tijl De Bie, and Nello Cristianini

Intelligent Systems Laboratory
University of Bristol, UK

Abstract. We present an automated tool with a web interface for tracking the prevalence of Influenza-like Illness (ILI) in several regions of the United Kingdom using the contents of Twitter's microblogging service. Our data is comprised by a daily average of approximately 200,000 geolocated tweets collected by targeting 49 urban centres in the UK for a time period of 40 weeks. Official ILI rates from the Health Protection Agency (HPA) form our ground truth. Bolasso, the bootstrapped version of LASSO, is applied in order to extract a consistent set of features, which are then used for learning a regression model.

1 Introduction

Monitoring the diffusion of an epidemic disease such as seasonal influenza is a very important task. Various methods are deployed by the health sector in order to detect and constrain epidemics, such as counting the consultation rates of general practitioners (GPs) [1], school or workforce absenteeism figures [2], etc. The need of a proper infrastructure and the time delays due to the necessary data processing are the main drawbacks of those methodologies.

We argue that information available on the web can provide an additional means for tackling this problem. In ([3,4]) it has been demonstrated that user queries on web search engines can be used to provide an early warning for an epidemic. Furthermore, recent work ([5,8]) has shown that the social web media have a predictive power on different domains. In particular, article [5] presents a method for inferring ILI rates for several regions in the UK by using data from Twitter[1]. The core of the method performed feature selection and regression with L1 regularisation by applying the LASSO [7]. This paper extends our previous methodology and presents a complete pipelined application of it: the "Flu detector" (`http://geopatterns.enm.bris.ac.uk/epidemics/`).

Short and geolocated messages from users on Twitter, commonly known as 'tweets' on the one side, and weekly ILI reports from the HPA on the other for 3 UK regions, namely Central England & Wales (r_1), South England (r_2) and North England (r_3), are the sources of our information. For the experimental part of this work, we use 40 weeks of the Twitter corpus and the respective ground truth, from 22/06/2009 to 28/03/2010. We apply Bolasso, the bootstrapped version of LASSO [6], for performing feature selection on the vector

---

[1] Twitter micro-blogging social network, `http://twitter.com/`

space representation of the Twitter corpus, *i.e.* for extracting a consistent set of keywords that can be used for inferring HPA's ILI rates optimally. The selected features are then used in a regression model; evaluating the performance of our model on unseen data yields inferences which are very well correlated with the ground truth.

2 Methodology

Our aim is to compute a flu-score from Twitter corpus on a daily basis. For this purpose, we learn a set of weighted keywords (we refer to them as markers or features) via an automated process. Given a set of markers $\mathcal{M} = \{m_i\}$, $i \in [1, n]$, their respective weights $\mathcal{W} = \{w_i\}$, $i \in [1, n]$, and a set of tweets $\mathcal{T} = \{t_j\}$, $j \in [1, k]$, Twitter's flu-score $f_{\mathcal{S}}$ is defined as $f_{\mathcal{S}}(\mathcal{T}, \mathcal{W}, \mathcal{M}) = \sum_j \sum_i w_i \times g(t_j, m_i)/k$, where $g(t_j, m_i) = 1$, if a tweet t_j contains a marker m_i, otherwise $g(t_j, m_i) = 0$.

We start by creating a pool of candidate features by using encyclopedic and informal references related to influenza as well as some flu-related word clusters created by Google Sets. After the necessary preprocessing (tokenisation, stemming, stop-word and name removal), we end up with a set of $\theta = 2675$ candidate features, denoted by $\mathcal{C} = \{c_u\}$, $u \in [1, \theta]$.[2] Given a set $\mathcal{T}$ of k tweets, each candidate feature $c_u \in \mathcal{C}$ has its own normalised and unweighted flu score $f_{\mathcal{C}}(\mathcal{T}, c_u) = \sum_j g(t_j, c_u)/k$ (denoted also as f_{c_u} for keeping the notation short). For a time period of h days, the flu-score time series of each candidate feature (denoted by $f_{c_u}^{(h)}$) forms an $h \times \theta$ array, $X^{(h)} = [f_{c_1}^{(h)} \dots f_{c_\theta}^{(h)}]$. HPA ILI rates for the same time period are denoted by $y^{(h)}$. We use Bolasso method for extracting a consistent set of markers with respect to the ground truth. Internally, Bolasso uses the LASSO method for performing regression with L1-regularisation which provides a sparse solution [7]. In our case, LASSO is formulated as the following optimisation problem:

$$\min_w \|X^{(h)}w - y^{(h)}\|_2^2 \quad \text{s.t. } \|w\|_1 \leq t,$$

where vector w is guaranteed to be a sparse solution and t is the regularisation parameter. Bolasso decides automatically the optimal value for t. A soft version of Bolasso is used, *i.e.* we select the markers that have non zero weights in $s = 65\%$ to 75% of the bootstraps. Then, we perform linear least squares regression for learning the weights of the selected markers. During testing, each inferred daily flu score is smoothed (averaged) with the flu scores of the past 6 days to infer a weekly trend.

The performance of the described method is evaluated by computing the mean absolute error (MAE) between the inferred and the target values. When the ground truth signal is clearly present, we additionally compute its linear correlation with the inferences. The predictability of the selected markers is assessed by testing on the last 3 weeks (training is performed on the preceding weeks); this gives out MAEs equal to 5.27, 4.26 and 2.18 for regions r_1, r_2

---

[2] More information on the candidate features is available at `http://goo.gl/7TZA`

and r_3 respectively. Based on the fact that the actual flu rates are high in the beginning of the investigated time period (when the epidemic was emerging), we also test our method's inferences for the first 3 weeks (training is performed on the succeeding weeks); for regions r_1, r_2 and r_3, the MAEs are equal to 18.34, 9.38 and 27.29, and the corresponding linear correlations are equal to 0.94, 0.84 and 0.87 (with p-values $< 10^{-5}$). As an overall performance quantification, 10-fold cross validation is performed, where each fold is formed by 4 contiguous weeks; the MAE is on average equal to 11.1 with a standard deviation of 10.04.[3]

3   Data Collection and Back-End Operations

We focus our data collection on tweets geolocated within a 10 Km radius from the 49 most populated urban centres in regions r_1, r_2 and r_3. Twitter's Search API is used for querying the social network periodically in order to get the most recent tweets per urban centre. The posts are retrieved in Really Simple Syndication (RSS) format and are being parsed using the ROME Java API. Data collection is a non-stop process which is performed automatically by our crawling software; the retrieved data are stored in a MySQL database.

For the experimental part of this work we are using approximately 50 million tweets, *i.e.* 200K tweets per day. Vector space representations from the corpus are produced on demand using our Java libraries; tokenisation, stop word removal, and stemming by applying Porter's algorithm [9] are embedded in the whole process. Our ground truth is formed by weekly epidemiological reports from the HPA for several regions in the UK[4]. The reports are based on data gathered by the Royal College of General Practitioners (RCGP) and express the number of GP consultations per 10^5 citizens, where the the diagnosis result was ILI. In order to retrieve an equal representation between the weekly HPA rates and the daily Twitter flu-scores, firstly, we expand each value of the former over a 7-day period, and then we smooth the expanded ground truth time series with a 7-point moving average. Feature selection and learning of the weights are performed offline; we use an implementation of Bolasso made available in the Probabilistic Modeling Toolkit (PMTK)[5].

4   Web Interface

The Flu detector website presents the outcomes of our method (see Figure 1). An automated procedure updates the flu-score inferences on a daily basis. Visitors are able to view the inferred regional flu scores (with corresponding error bounds) in comparison with the actual HPA's ILI rates (which become available with a 7 to 10 day time lag). Apart from the aforementioned regions, we display a flu-score inference for the merged region of England & Wales as well as for the entire

---

[3] A detailed reference on evaluation procedures is available at `http://goo.gl/yZkG`
[4] HPA epidemiological reports, `http://goo.gl/wJex`
[5] Probabilistic Modeling Toolkit v.3, `http://code.google.com/p/pmtk3/`
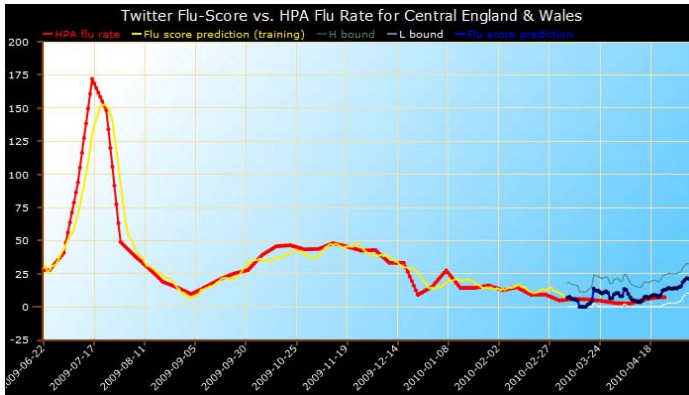
Fig. 1. Flu detector's regional predictions as they appear on the website

UK - for each regional plot, the exact locations which contributed to the score are listed. The periods of training are denoted with distinct colours in every time line and the MAE between the inferred and the actual flu scores is computed and displayed.

References

1. Fleming, D., Elliot, A.: Lessons from 40 years surveillance of influenza in England and Wales. Epidemiology and Infection 136(7), 866–875 (2007)
2. Neuzil, K.M., Hohlbein, C., Zhu, Y.: Illness among schoolchildren during influenza season: effect on school absenteeism, parental absenteeism from work, and secondary illness in families. Arch. Pediatr. Adolesc. Med. 156(10), 986–991 (2002)
3. Ginsberg, J., Mohebbi, M.H., et al.: Detecting influenza epidemics using search engine query data. Nature 457(7232), 1012–1014 (2008)
4. Polgreen, P.M., Chen, Y., et al.: Using internet searches for influenza surveillance. Clinical Infectious Diseases 47, 1443–1448 (2008)
5. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the Social Web. In: 2nd IAPR Workshop on Cognitive Information Processing, pp. 411–416 (2010)
6. Bach, F.R.: Bolasso: model consistent Lasso estimation through the bootstrap. ICML 25, 33–40 (2008)
7. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 58B, 267–288 (1996)
8. Asur, S., Huberman, B.A.: Predicting the Future with Social Media. Arxiv preprint arXiv:1003.5699 (2010)
9. Porter, M.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)