3rd Ph.D. Review

Tracking trends on the web using novel Machine Learning methods

Vasileios Lampos

advised by

Professor Nello Cristianini

Intelligent Systems Laboratory, University of Bristol

May 25, 2010

Aims

The **general aims** of our research project can be summarised in the following points:

1. Track **trends** on the Web by applying Machine Learning methods (track expresses the notions of infer or predict as well)

2. Extend current or invent new **methodologies** (where and if needed) for accomplishing our primary aim

3. Build **tools** that apply the experimental/theoretical results in real and large-scale applications (featured research)

Being more specific

1. **Trends** about what? Examples?
   - Predict flu rates (*epidemics*)
   - Infer vote intensions (*politics*)
   - Infer traffic/weather conditions (*toy problems*)

Being more specific

1. **Trends** about what? Examples?
   - Predict flu rates (*epidemics*)
   - Infer vote intensions (*politics*)
   - Infer traffic/weather conditions (*toy problems*)

2. **Methodologies**?
   - Feature extraction/selection
   - Exploit probabilistic relationships (PGMs)
   - Regression/classification/ranking scenarios
   - Active learning

Being more specific

1. **Trends** about what? Examples?
   - Predict flu rates (*epidemics*)
   - Infer vote intensions (*politics*)
   - Infer traffic/weather conditions (*toy problems*)

2. **Methodologies**?
   - Feature extraction/selection
   - Exploit probabilistic relationships (PGMs)
   - Regression/classification/ranking scenarios
   - Active learning

3. **Applications**?
   - Back-end infrastructure for data collection/retrieval/mining
   - Real time online tools for making and displaying predictions
     (like the **Flu detector**)

P1 - Summary (1 of 3)

Title:	**Tracking the flu pandemic by monitoring the Social Web**
Authors:	V. Lampos and N. Cristianini
Submitted to:	IAPR Cognitive Information Processing 2010 (accepted)

- Twitter and Health Protection Agency data for weeks 26-49, 2009 (on average 160,000 tweets collected per day geolocated in 54 urban centres in the UK)
- Frequency of **41 flu related words** (markers) in Twitter corpus had a correlation of $>$**80%** with the HPA flu rates in all UK regions
- Learn a better list of weighted markers **automatically**:
    - Generate a list of candidate markers (1560 words taken from flu related web pages)
    - Use **LASSO** for feature selection

P1 - Summary (2 of 3)

Validation schemes:

1. Train on one region, validate regularisation parameter on another, test on the remaining regions (for all possible combinations)

Train/Validate (regions)	A	B	C	D	E
A	-	0.9594	0.9375	0.9348	0.9297
B	0.9455	-	0.9476	0.9267	0.9003
C	0.9154	0.9513	-	0.8188	0.908
D	0.9463	0.9459	0.9424	-	0.9337
E	0.8798	0.9506	0.9455	0.8935	-
				Total Avg.	**0.9256**

97 selected words: lung, unwel, temperatur, like, headach, season, unusu, chronic, child, dai, appetit, stai, symptom, spread, diarrhoea, start, muscl, weaken, immun, feel, liver, plenti, antivir, follow, sore, peopl, nation, small, pandem, pregnant, thermomet, bed, loss, heart, mention, condit, ...
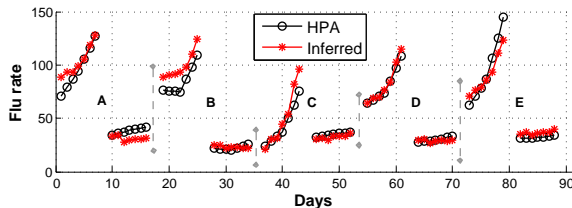
2. Aggregate data from all regions, test on weeks 28 and 41 (2009) and train using the rest of the data set

P1 - Summary (3 of 3)

1. Inferred vs Official flu rate in North England
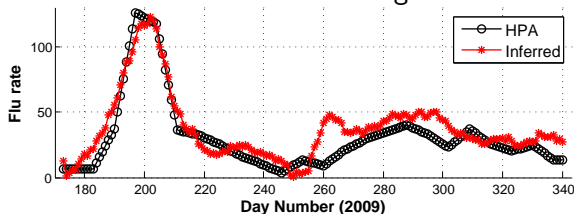
2. Inferred vs Official rates in all regions (aggregated data set)

P2 - Summary

Title:	**Flu detector - Tracking epidemics on Twitter**
Authors:	V. Lampos, T. De Bie, and N. Cristianini
Submitted to:	ECML PKDD 2010 Demos (under review)

- Extending and making more robust the methodology of P1

- Larger data sets (bigger time series) and more (2675) candidate features

- Select a list of features (markers) using BoLASSO (bootstrap version of LASSO)

- Then learn weights of those markers via linear least squares regression

- Stricter evaluation of the methodology - **Available online**

- Put all this into practice and come up with the **Flu detector**

Other activities

- Studied/implemented the necessary statistical tools and algorithms (in MATLAB or Java)
- Extended further the infrastructure for conducting large scale experiments and data retrieval on demand
- TA for Intro to AI, Data Analysis and Pattern Analysis & Statistical Learning
- Attended some of the ISL meetings and seminars

General goals & activities

1. ... (content omitted)
2. ... (content omitted)
3. ... (content omitted)
4. ... (content omitted)

A more time specific tentative plan

- In **June**: ... (content omitted)
- In **July**: ... (content omitted)
- In **August**: ... (content omitted)
- In **September** - **November**: ... (content omitted)

This is the last slide.

This is the last slide.

Any questions?