Structured Prediction for Quantification

Andrea Esuli[1] and Fabrizio Sebastiani[2]

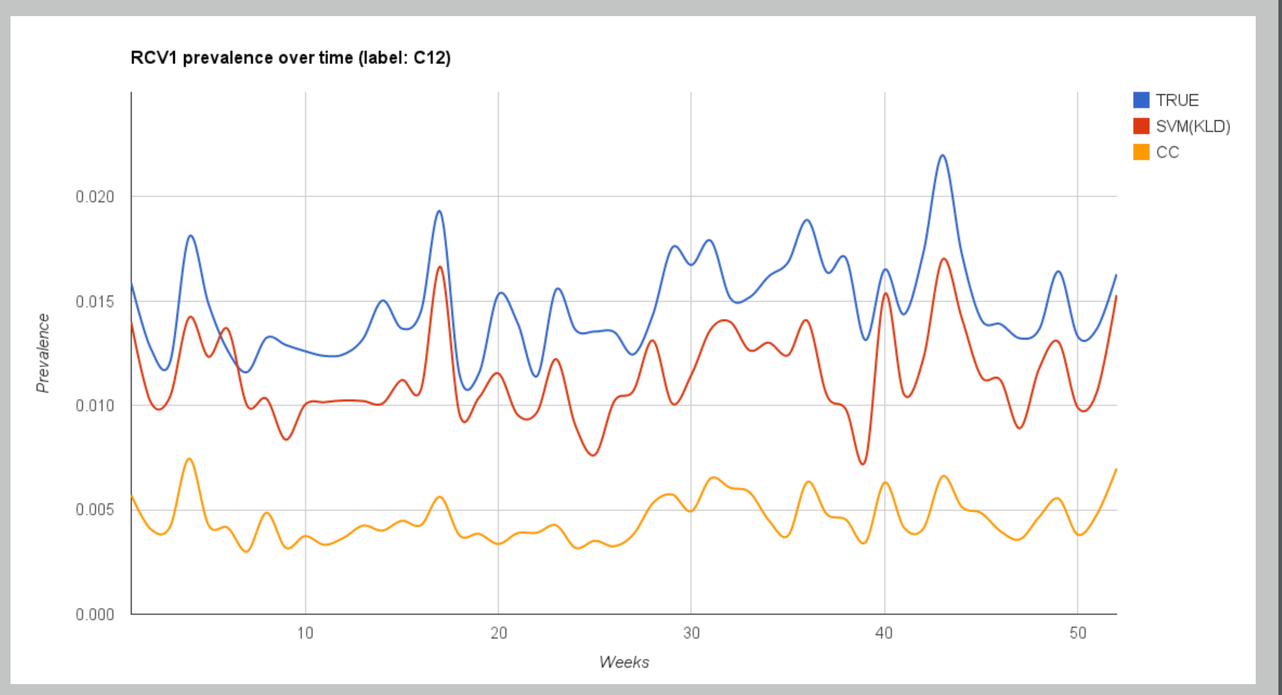
[1]Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy
[2]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar

Introduction

- Quantification: estimating the prevalence $p_S(c_i)$ of each $c_i \in C$ in the unlabelled data S
- Different from classification, since the emphasis is at the aggregate (rather than at the individual) level
- Important in fields such as e.g., market research, social sciences, epidemiology, etc., where interest is "not in the needle, but in the haystack".

Prior work in quantification

- "Classify and count" is a suboptimal quantification method:



- A good classifier need not be a good quantifier, and vice versa

	FP	FN
Classifier A	18	20
Classifier B	20	20

- In most previously proposed methods, a generic classifier h is trained and applied to Te, and the computed prevalences $\hat{p}(c)$ are then corrected according to the estimated (via k-FCV on Tr) *bias* of h; but these latter estimates might not be reliable in the presence of drift ...

Explicit Loss Minimization and Quantification

- We propose an approach based on learners specifically designed for quantification
- This may be done by applying explicit loss minimization: i.e., train a learner directly optimized for the measure to be used in evaluating the results
- Problem: "standard" learning methods can minimize linear loss measures (or proxies thereof) only, while error measures for quantification, such as

$$KLD(\hat{p}, p) = \sum_{c_j \in \mathcal{C}} p(c_j) \log \frac{p(c_j)}{\hat{p}(c_j)} \quad (1)$$

are inherently non-linear.

Structured Prediction for Quantification

- We use SVM$_{perf}$[a], a structured output prediction algorithm that can generate classifiers optimized for any non-linear loss that can be computed from a contingency table.
- Instead of handling hypotheses $h : \mathcal{X} \rightarrow \mathcal{Y}$, SVM$_{perf}$ considers hypotheses $\bar{h} : \bar{\mathcal{X}} \rightarrow \bar{\mathcal{Y}}$, where $\bar{\mathrm{x}} = (\mathrm{x}_1, ..., \mathrm{x}_n)$ and $\bar{\mathrm{y}} = (y_1, ..., y_n)$.
- Instead of learning hypotheses

$$h(\mathrm{x}) = sign(\mathrm{w} \cdot \mathrm{x} + b) \quad (2)$$

SVM$_{perf}$ learns hypotheses

$$\bar{h}(\bar{\mathrm{x}}) = \arg\max_{\bar{y}' \in \mathcal{Y}} (\mathrm{w} \cdot \Psi(\bar{\mathrm{x}}, \bar{y}') + b) \quad (3)$$

where $\Psi(\bar{\mathrm{x}}, \bar{y}')$ (the *joint feature map*) evaluates how compatible $\bar{\mathrm{x}}$ is with $\bar{y}'$.
- We instantiate SVM$_{perf}$ with KLD, and apply SVM(KLD) to (binary) quantification.

[a]Thorsten Joachims. A support vector method for multivariate performance measures. ICML 2005:377-384.

Experiments

- We test SVM(KLD) against 10 baselines on a large textual dataset (RCV1-v2), chopped up into 5,148 test sets

		RCV1-v2
ALL	Total # of docs	804,414
	# of classes (i.e., binary tasks)	99
	Time unit used for split	week
TRAINING	# of docs	12,807
	# of features	53,204
	Min # of positive docs per class	2
	Max # of positive docs per class	5,581
	Min prevalence of the positive class	0.0001
	Max prevalence of the positive class	0.4375
TEST	# of docs	791,607
	# of test sets per class	52
	Total # of test sets	5,148
	Avg # of test docs per set	15,212
	Min # of positive docs per class	0
	Max # of positive docs per class	9,775
	Min prevalence of the positive class	0.0000
	Max prevalence of the positive class	0.5344

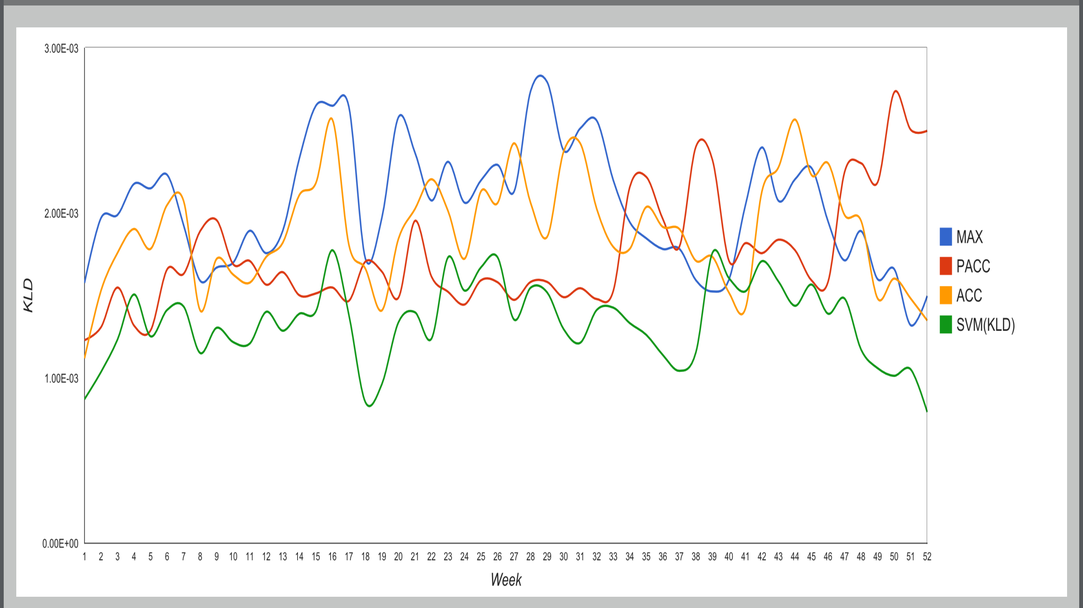
Results according to the "class prevalence" dimension

		VLP	LP	HP	VHP	All
RCV1-v2	SVM(KLD)	2.09E-03	**4.92E-04**	7.19E-04	1.12E-03	**1.32E-03**
	PACC	2.16E-03	1.70E-03	4.24E-04	2.75E-04	1.74E-03
	ACC	2.17E-03	1.98E-03	5.08E-04	6.79E-04	1.87E-03
	MAX	2.16E-03	2.48E-03	6.70E-04	**9.03E-05**	2.03E-03
	CC	2.55E-03	3.39E-03	1.29E-03	1.61E-03	2.71E-03
	X	3.48E-03	8.45E-03	1.32E-03	2.43E-04	4.96E-03
	PCC	1.04E-02	6.49E-03	3.87E-03	1.51E-03	7.86E-03
	MM(PP)	1.76E-02	9.74E-03	2.73E-03	1.33E-03	1.24E-02
	MS	1.98E-02	7.33E-03	3.70E-03	2.38E-03	1.27E-02
	T50	1.35E-02	1.74E-02	7.20E-03	3.17E-03	1.38E-02
	MM(KS)	2.00E-02	1.14E-02	9.56E-04	3.62E-04	1.40E-02

Results according to the "distribution drift" dimension

		VLD	LD	HD	VHD	All
RCV1-v2	SVM(KLD)	**1.17E-03**	**1.10E-03**	**1.38E-03**	1.67E-03	**1.32E-03**
	PACC	1.92E-03	2.11E-03	1.74E-03	1.20E-03	1.74E-03
	ACC	1.70E-03	1.74E-03	1.93E-03	2.14E-03	1.87E-03
	MAX	2.20E-03	2.15E-03	2.25E-03	1.52E-03	2.03E-03
	CC	2.43E-03	2.44E-03	2.79E-03	3.18E-03	2.71E-03
	X	3.89E-03	4.18E-03	4.31E-03	7.46E-03	4.96E-03
	PCC	8.92E-03	8.64E-03	7.75E-03	6.24E-03	7.86E-03
	MM(PP)	1.26E-02	1.41E-02	1.32E-02	1.00E-02	1.24E-02
	MS	1.37E-02	1.67E-02	1.20E-02	8.68E-03	1.27E-02
	T50	1.17E-02	1.38E-02	1.49E-02	1.50E-02	1.38E-02
	MM(KS)	1.41E-02	1.58E-02	1.53E-02	1.10E-02	1.40E-02

Results according to the "temporal" dimension



Get more details in ...

- A. Esuli and F. Sebastiani. Optimizing Text Quantifiers for Multivariate Loss Functions. *ACM Transactions on Knowledge Discovery from Data, 2015, in press.*

http://www.qcr.qa/our-people/fabriziosebastiani fsebastiani@qf.org.qa