Optimizing Text Quantifiers for Multivariate Loss Functions

by Andrea Esuli and Fabrizio Sebastiani

Quantification - also known as class prior estimation – is the task of estimating the relative frequencies of classes in application scenarios in which such frequencies may change over time. This task is becoming increasingly important for the analysis of large and complex datasets. Researchers from ISTI-CNR, Pisa, are working with supervised learning methods explicitly devised with quantification in mind.

In some applications involving classification the final goal is not determining which class(es) individual unlabelled data items belong to, but determining the prevalence (or 'relative frequency') of each class in the unlabelled data. This task has come to be known as 'quantification'.

For instance, a company may want to find out how many tweets that mention product X express a favourable view of X. On the surface, this seems a standard instance of query-biased tweet sentiment classification. However, the company is likely not interested in whether a specific individual has a positive view of X but in knowing how many of those who tweet about X have a positive view of X; that is, the company is actually interested in knowing the relative frequency of the positive class.

Quantification (also known as 'class prior estimation', or 'prevalence estimation') has several applications, in fields as diverse as machine learning, sentiment analysis, natural language processing [1], data mining, social science [3], epidemiology, and resource allocation. The research community has recently shown a growing interest in tackling quantification as a task in its own right, instead of a mere byproduct of classification. One reason for this is that quantification requires evaluation measures that are different from those used for classification. Second, using a classifier optimized for classification accuracy is suboptimal when quantification accuracy is the real goal, since a classifier may optimize classification accuracy at the expense of bias. Third, quantification is predicted to be increasingly important in tomorrow's applications; the advent of big data will result in more application contexts in which analysis of data at the aggregate rather than the individual level will be the only available option.

The obvious method for dealing with quantification is to classify each unlabelled document and estimate class prevalence by counting the documents that have been attributed the class. However, when a standard learning algorithm is used, this strategy is suboptimal since, as observed above, classifier A may be more accurate than classifier B but may also exhibit more bias than B, which means that B would be a better quantifier than A.

In this work (see [2] for details) we take an 'explicit loss minimization' approach, based upon the use of classifiers explicitly optimized for the evaluation function that we use for assessing quantification accuracy. Following this route for solving quantification is non-trivial, because the measures used for evaluating quantification accuracy are inherently non-linear and multivariate, and the assumption that the evaluation measure is instead linear and univariate underlies most existing discriminative learners, which are thus suboptimal for tackling quantification.

In order to sidestep this problem we adopt the 'SVM for Multivariate Performance Measures' (SVMperf) learning algorithm proposed by Joachims, and instantiate it to optimize Kullback-Leibler Divergence, the standard measure for evaluating quantification accuracy; we dub the resulting system SVM(KLD). SVMperf is a learning algorithm of the Support Vector Machine family that can generate classifiers optimized for any non-linear, multivariate loss function that can be computed from a contingency table, such as KLD. SVMperf is a learning algorithm for 'structured prediction', i.e., an algorithm designed for predicting multivariate, structured objects. It is fundamentally different from conventional algorithms for learning classifiers: while the latter learn univariate classifiers (i.e., functions that classify individual instances independently of each other), SVMperf learns multivariate classifiers (i.e., functions that jointly label all the instances belonging to a set S). By doing so, SVMperf can optimize properties of entire sets of instances, properties (such as KLD) that cannot be expressed as linear functions of the properties of the individual instances.

Experiments conducted on 5,500 binary text quantification test sets, averaging 14,000+ documents each, have shown that SVM(KLD) outperforms existing state-of-the-art quantification algorithms both in terms of accuracy and sheer stability, and is computationally more efficient than all but the most trivial algorithms.

Link:
http://nmis.isti.cnr.it/sebastiani/Publications/TKDD15.pdf

References:
[1] Y. S. Chan, H. T. Ng: "Word sense disambiguation with distribution estimation", in proc. of IJCAI 2005, Edinburgh, UK, 1010–1015, 2005.
[2] A. Esuli, , F. Sebastiani: "Optimizing Text Quantifiers for Multivariate Loss Functions", ACM Transactions for Knowledge Discovery and Data, forthcoming.
[3] D. J. Hopkins, G.King: "A method of automated non-parametric content analysis for social science", American Journal of Political Science 54, 1, 229–247, 2010.

Please contact:
Andrea Esuli - ISTI-CNR
Tel: +39 050 3152 878
E-mail: andrea.esuli@isti.cnr.it