Active, Semi-Supervised Learning for Textual Information Access

Anastasia Krithara, Cyril Goutte, Massih-Reza Amini and Jean-Michel Renders

I. Introduction

MACHINE learning techniques have been used for various tasks of document management and textual information access, such as categorisation, information extraction, or automatic organization of large document collections. Acquiring the annotated data necessary to apply supervised learning techniques is a major challenge for text applications, especially in very large collections. Annotating textual data usually requires humans who can read and understand the texts, and is therefore very costly, especially in technical domains. In this contribution, we address the problem or reducing this annotation burden.

II. Method

We describe a method which combines semi-supervised and active learning for document categorisation. We use a semi-supervised version of the Probabilistic Latent Semantic Analysis (PLSA) model [3], which we combine with a pool-based active learning method. In our study we adopt the semi-supervised learning approach of [2]. On top of this, we perform active learning by selecting the most ambiguous non-annotated examples and annotating them. As the set of annotated data grows, the performance of the model increases. Semi-supervised learning and active learning both address the problem of reducing the annotation effort by 1) learning from partially annotated data and 2) selecting for annotation the examples that are believed to be the most useful for improving the model.

A. Semi-supervised Latent Semantic Analysis

Our semi-supervised learning technique follows [2]. Let us assume a dataset consisting of ℓ labelled examples $\mathcal{L} = \{(x_1, y_1), \dots (x_\ell, y_\ell)\}$ and u unlabelled examples $\mathcal{U} = \{x_{\ell+1}, \dots x_{\ell+u}\}$, hence $N = \ell + u$ examples in all. We introduce an additional "fake label" variable z such that $\forall i, 1 \leq i \leq \ell, z_i = y_i$ and $\forall i, \ell < i \leq N, z_i = 0$.

In the case of textual data modelled by PLSA, each example x is the co-occurrence of a word w with a document d, which we denote by $x = (w, d)$. The probabilistic model is a mixture of multinomial distributions over (w, d, z):

$$P(w, d, z) = P(d) \sum_c P(w|c)P(c|d)P(z|c)$$

where $c = 1 \dots K$ is the index over K latent components. Note that the modelling assumption is that w, d and z are conditionally independent given c.

We then use a variant of the Expectation-Maximisation (EM) algorithm to train the multinomial mixture model, with appropriate constraints on the components to ensure good separation of the labelled examples from different classes. The (log)likelihood of the data is:

$$L = \sum_{i \in \mathcal{L} \cup \mathcal{U}} \ln P(w_i, d_i, z_i) = \sum_d \sum_w n(w, d) \ln P(w, d, z(d))$$

where $z(d)$ is the (unique) label of document d (assuming documents may not be partially annotated), and $n(w, d)$ the number of occurrences of word w in document d. The EM algorithm alternates the following Expectation and Maximisation steps. In the E-step, we calculate

$$\pi_c(w, d) = P(c|w, d, z(d)) = \frac{P(c|d)P(w|c)P(z(d)|c)}{\sum_{c'} P(c'|d)P(w|c')P(z(d)|c')},$$

and in the M-step we update the model parameters by:

$$P(w|c) \propto \sum_d n(w, d)\pi_c(w, d) \quad (1)$$

$$P(c|d) \propto \sum_w n(w, d)\pi_c(w, d) \quad (2)$$

$$P(z|c) \propto \sum_{d, z(d)=z} \sum_w n(w, d)\pi_c(w, d). \quad (3)$$

In addition, in order to avoid mixing examples with different labels in the same component, we assume that each component c may only generate examples from one class ($z = z_c$) or unlabelled examples ($z = 0$). This means that $\forall c, z, P(z|c) = 0$ if $z \notin \{0; z_c\}$ (it turns out that this condition is a fixed point of EM and therefore easy to enforce).

Once the model parameters are obtained (as a fixed point of EM), we may "decode" them to obtain probabilities of assignment to the "true" labels:

$$P(y|x) \propto P(z = y|x) + \lambda P(y|z = 0)P(z = 0|x).$$

In our experiments, $P(y|z = y) = P(y|z = 0) = 1/2$ and $\lambda = 0.02$.

A. Krithara and J.-M. Renders are with the Xerox Research Centre Europe, 6 chemin de Maupertuis, F-38240 Meylan, France. Email: Anastasia.Krithara@xrce.xerox.com, Jean-Michel.Renders@xrce.xerox.com

C. Goutte was with XRCE and is now with the National Research Council Canada, Institute for Information Technology, Interactive Language Technology Group, 101 rue St-Jean-Bosco, Gatineau, QC K1A 0R6, Canada. Email: Cyril.Goutte@nrc-cnrc.gc.ca

M.-R. Amini is with the Department of Computer Science, University Pierre et Marie Curie, 8 rue du Capitaine Scott, F-75015 Paris, France. Email: Massih-Reza.Amini@lip6.fr

B. Active learning

In [2], it was shown that this model seems to provide a better estimate of the labelling uncertainty than the standard semi-supervised EM [4]. This uncertainty is used as a guide to choose the queried examples in the active learning step. In order to gain the maximum information from the annotation, we choose the *most ambiguous* example, ie the example for which the category assignment probability has the largest entropy. For binary categorisation this reduces to picking the example for which the assignment probability is closest to 1/2.

We then query the annotation of that example from the user, and update the model parameters accordingly. We will now illustrate the use of the active semi-supervised PLSA algorithm on some small binary categorisation tasks.

III. Experiments

In order to illustrate this approach we perform some categorisation experiments on the 20 newsgroups dataset [1]. We consider three binary categorisation tasks between pairs of newsgroups (number of examples in brackets):

1) rec.sport.baseball (994) vs. rec.sport.hockey (999)
2) comp.sys.ibm.pc.hardware (982) vs. comp.sys.mac.hardware (961)
3) talk.religion.misc (628) vs. alt.atheism (799)

In each case, we set aside 20% of the data as test set to provide an unbiased estimate of the categorisation accuracy. The remaining 80% is the training set, both labelled and unlabelled. In each experiment, we start with 1 random annotated example from each class, and all remaining examples are unlabelled. Then, at each step of the active learning procedure, we query the label of one example, retrain the model, etc. In our experiments, we query a total of 50 examples.

We test the method described above (active, semi-supervised PLSA) and compare it to two contenders. The first is the same semi-supervised PLSA model, but at each step, we query a random example instead of the most ambiguous one. This allows us to evaluate the influence of the active learning step in the overall performance gain. The second contender is a SVM model, where the active learning step is done by querying the annotation of the example closest to the separating hyperplane [5].

Each active learning session is repeated 10 times, with different randomly chosen annotation seeds (the 2 initial examples), for each model. Figure 1 displays the average results obtained on the first task. The active semi-supervised PLSA displays an average gain of around 10 points in accuracy. Interestingly, most of the gain is obtained with the first 5-10 examples, and slows down afterwards. Note that 10 examples, on this task, corresponds to only 0.6% of the unlabelled data. Querying random examples does not work as well, especially after the first few examples. With the SVM, the performance starts much lower. This is due to the fact that the SVM uses only the annotated examples, and, starting off with too few examples, is unable to learn any useful separation. The rate of progress of the SVM, however, is much larger and more regular than for the other two methods. It eventually reaches comparable performance, but not before around 50 additional examples have been queried and labelled.

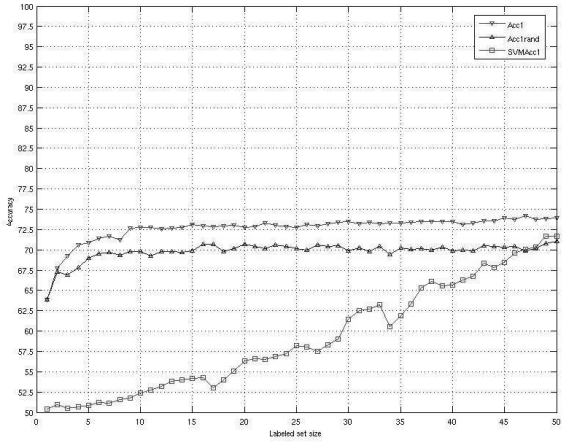

Fig. 1. Example of result on the first binary task. Active semi-supervised PLSA (top) compared to PLSA querying random examples (middle) and SVM querying examples closest to the margin (bottom).

On the second and third tasks, the results are qualitatively similar, although the amplitude of the gain is lower. As a consequence, the advantage of the active semi-supervised PLSA seems smaller.

IV. Conclusion

We have reported on our investigation of the use of a combination of active and semi-supervised learning in order to reduce the annotation burden for a supervised text classification task. For many textual information access problems, obtaining the amount of annotated data required for applying supervised learning methods is a key issue. We argue that the proposed technique addresses this issue in two ways: semi-supervised learning helps leverage unlabelled examples to improve categorisation accuracy, while active learning helps annotators concentrate on the potentially most informative examples. Our algorithm can easily be extended in order to take into account different labelling costs. In that way, we will be able to choose a document for labelling by its trade-off of ambiguity and cost of annotation (for example, it is harder to annotate a long document than a short one). Also we plan to continue investigating different active learning techniques in addition to the current proposition.

References

[1] http:// people.csail.mit.edu/jrennie/20Newsgroups/
[2] Eric Gaussier and Cyril Goutte, "Learning from partially labelled data – with confidence,", *Proc. of Learning with Partially Classified Training Data - ICML 2005 workshop, Bonn, Germany*, 2005.
[3] Thomas Hofmann, "Probabilistic latent semantic indexing," *Proc. SIGIR-99*, pages 35–44, 1999.
[4] David J. Miller and Hasan S. Uyar, "A mixture of experts classifier with learning based on both labelled and unlabelled data," *In Michael Mozer, Michael Jordan, and Thomas Petsche, editors, Advances in Neural Information Processing Systems 9*, pp. 571–577, 1997.
[5] Simon Tong and Daphne Koller, "Support Vector Machine Active Learning with Applications to Text Classification," *Journal of Machine Learning Research*, 2, pp. 45–66, 2001.