Multi-view Learning and Link Farm Discovery

Tobias Scheffer

Humboldt-Universität zu Berlin, Department of Computer Science
Unter den Linden 6, 10099 Berlin, Germany
scheffer@informatik.hu-berlin.de

Abstract. The first part of this abstract focuses on estimation of mixture models for problems in which multiple views of the instances are available. Examples of this setting include clustering web pages or research papers that have intrinsic (text) and extrinsic (references) attributes. Mixture model estimation is a key problem for both semi-supervised and unsupervised learning. An appropriate optimization criterion quantifies the likelihood and the consensus among models in the individual views; maximizing this consensus minimizes a bound on the risk of assigning an instance to an incorrect mixture component. An EM algorithm maximizes this criterion. The second part of this abstract focuses on the problem of identifying link spam. Search engine optimizers inflate the page rank of a target site by spinning an artificial web for the sole purpose of providing inbound links to the target. Discriminating "natural" from "artificial" web sites is a difficult multi-view problem.

1 Introduction

In many application domains, instances can be represented in two or more distinct, redundant views. For instance, web pages can be represented by their text, or by their context in the hyperlink graph, and research papers can be represented by their references from and to other papers, in addition to their content. In this case, multi-view methods such as co-training [1] can learn two initially independent hypotheses. These hypotheses bootstrap by providing each other with conjectured class labels for unlabeled data. Multi-view learning has often proven to utilize unlabeled data effectively, increase the accuracy of classifiers [2,1] and improve the quality of clusterings [3].

Nigam and Ghani [4] have proposed the co-EM procedure that resembles semi-supervised learning with EM [5], using two views that alternate after each iteration. The EM algorithm [6] is very well understood. In each iteration, it maximizes the joint log-likelihood of visible and invisible parameters given the visibles and the parameter estimates of the previous iteration – the Q function. This procedure is known to greedily maximize the likelihood of the data. By contrast, the primary justification of the co-EM algorithm is that it often works very well; it is not known which criterion the method maximizes.

We take a top down approach on the problem of mixture model estimation in a multi-view setting. A result of Dasgupta et al. [7] motivates our work by

Dagstuhl Seminar Proceedings 05051
Probabilistic, Logical and Relational Learning - Towards a Synthesis
http://drops.dagstuhl.de/opus/volltexte/2006/414

showing that a high consensus of independent hypotheses implies a low error rate. We derive a criterion that quantifies likelihood and consensus and derive an EM procedure that maximizes it. We contribute to an understanding of EM for multiple views by showing that the co-EM algorithm [4] is a special case of the resulting procedure. Our solution naturally generalizes co-EM because it operates on more than two views.

Web pages that exist for the sole purpose of inflating the page rank of a target site deteriorates the performance of search engines. Identifying such *link spam* pages has been called one of the most important research problems in information retrieval [8]. Both the content of a web page and the context in the hyperlink graph contribute relevant information. Identifying link spam therefore appears to ba a natural application of multi-view learning.

2 Multi-View EM

The *multi-view* setting that we consider is characterized by available attributes X which are decomposed into views $X^{(1)}, \ldots, X^{(s)}$. An instance $x = (x^{(1)}, \ldots, x^{(s)})$ has representations $x^{(v)}$ that are vectors over $X^{(v)}$. We focus on the problem of estimating parameters of a generative mixture model in which data are generated as follows.

The *data generation process* selects a mixture component j with probability α_j. Mixture component j is the value of a random variable Z. Once j is fixed, the generation process draws the s independent vectors $x^{(v)}$ according to the likelihoods $P(x^{(v)}|j)$. The likelihoods $P(x^{(v)}|j)$ are assumed to follow a parametric model $P(x^{(v)}|j, \Theta)$ (distinct views may of course be governed by distinct distributional models).

The *learning task* involved is to estimate the parameters $\Theta = (\Theta^{(1)}, \ldots, \Theta^{(s)})$ from data. The *sample* consists of n observations that either contain only the visible attributes $x_i^{(v)}$ in all views v of the instances x_i (unsupervised multi-view learning) or additionally contains some examples that are labeled with the mixture component that they originate from (semi-supervised learning). The attributes $x_i^{(v)}$ in all views v of the instances x_i. The vector Θ contains priors $\alpha_j^{(v)}$ and parameters of the likelihood $P(x_i^{(v)}|j, \Theta^{(v)})$, where $1 \leq j \leq m$ and m is the number of mixture components assumed by the model (clusters). Given Θ, we will be able to calculate a posterior $P(j|x^{(1)}, \ldots, x^{(s)}, \Theta)$. This posterior will allow us to assign a cluster membership to any instance $x = (x^{(1)}, \ldots, x^{(s)})$. The *evaluation metric* is the impurity of the resulting clusters as measured by the entropy; the elements of each identified cluster should originate from the same true mixture component.

Dasgupta et al. [7] have studied the relation between the consensus among multiple independent hypotheses and their error rate. Let us review a very simple result that motivates our approach, it can be derived easily from their general treatment of the topic. Let $h^{(v)}(x) = \text{argmax}_j \, P(j|x^{(v)}, \Theta^{(v)})$ be two independent clustering hypotheses in views $v = 1, 2$. For clarity of the presentation, let there be two mixture components. Let x be a randomly drawn instance that belongs to

mixture component 1, and let both hypotheses $h^{(1)}$ and $h^{(2)}$ have a probability of at least 50% of assigning x to the correct cluster 1. In this case, we observe that

$$P(h^{(1)}(x) \neq h^{(2)}(x)) \geq \max_v P(h^{(v)}(x) \neq 1).$$

That is, the probability of a disagreement $h^{(1)}(x) \neq h^{(2)}(x)$ is an upper bound on the risk of an error $P(h^{(v)}(x) \neq 1)$ of either hypothesis $h^{(v)}$.

We give a brief *proof* of this observation. In Equation 1 we distinguish between the two possible cases of disagreement; we utilize the independence assumption and order the summands such that the greater one comes first. In Equation 2, we exploit that the error rate be at most 50%: both hypotheses are less likely to be wrong than just one of them. Exploiting the independence again takes us to Equation 3.

$$\begin{aligned}
&P(h^{(1)}(x) \neq h^{(2)}(x)) \\
&= P(h^{(v)}(x) = 1, h^{(\bar{v})}(x) = 2) + P(h^{(v)}(x) = 2, h^{(\bar{v})}(x) = 1) && (1)\\
&\quad \text{where } v = \text{argmax}_u P(h^{(u)}(x) = 1, h^{(\bar{u})}(x) = 2) \\
&\geq P(h^{(v)}(x) = 2, h^{(\bar{v})}(x) = 2) + P(h^{(v)}(x) = 2, h^{(\bar{v})}(x) = 1) && (2)\\
&= \max_v P(h^{(v)}(x) \neq 1) && (3)
\end{aligned}$$

In unsupervised learning, the risk of assigning instances to wrong mixture components cannot be minimized directly, but with the above argument we can minimize an upper bound on this risk.

Even though the goal is to maximize $P(X|\Theta)$, EM iteratively maximizes an auxiliary (single-view) criterion $Q^{SV}(\Theta, \Theta_t)$. The criterion refers to the visible variables X, the invisibles Z (the mixture component), the optimization parameter Θ and the parameter estimates Θ_t of the last iteration. Equation 4 defines $Q^{SV}(\Theta, \Theta_t)$ to be the expected log-likelihood of $P(X, Z|\Theta)$, given X and given that the hidden mixture component Z be distributed according to $P(j|x, \Theta_t)$.

$$\begin{aligned}
Q^{SV}(\Theta, \Theta_t) &= E[\log P(X, Z|\Theta)|X, \Theta_t] && (4)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} P(j|x_i, \Theta_t) \log(\alpha_j P(x_i|j, \Theta)) P(j|x_i, \Theta_t) && (5)
\end{aligned}$$

We want to maximize the likelihood in the individual views and the consensus of the models because we know that the disagreement bounds the risk of assigning an instance to an incorrect mixture component. Equation 6 defines our *multi-view Q function* as the sum over s single-view Q functions minus a penalty term $\Delta(\cdot)$ that quantifies the disagreement of the models $\Theta^{(v)}$ and is regularized by η.

$$\begin{aligned}
&Q^{MV}(\Theta^{(1)}, \ldots, \Theta^{(s)}, \Theta_t^{(1)}, \ldots, \Theta_t^{(s)}) \\
&= \sum_{v=1}^{s} Q^{SV}(\Theta^{(v)}, \Theta_t^{(v)}) - \eta\Delta(\Theta^{(1)}, \ldots, \Theta^{(s)}, \Theta_t^{(1)}, \ldots, \Theta_t^{(s)}) && (6)
\end{aligned}$$

When the regularization parameter η is zero, then $Q^{MV} = \sum_v Q^{SV}$. In each step, multi-view EM then maximizes the s terms Q^{SV} independently. It follows immediately from Dempster et al. [6] that each $P(X^{(v)}|\Theta^{(v)})$ increases in each step and therefore, if the views are independent, $P(X|\Theta) = \prod_v P(X^{(v)}|\Theta^{(v)})$ is maximized.

The disagreement term Δ should satisfy a number of desiderata. Firstly, since we want to minimize Δ, it should be convex. Secondly, for the same reason, it should be differentiable. Given Θ_t, we would like to find the maximum of $Q^{MV}(\Theta, \Theta_t)$ in one single step. We would, thirdly, appreciate if Δ was zero when the views totally agree.

We construct Δ to fulfill these desiderata in Equation 7. It contains the pairwise cross entropy $H(P(j|x_i^{(v)}, \Theta_t^{(v)}), P(j|x_i^{(u)}, \Theta^{(u)}))$ of the posteriors of any pair of views u and v. The second cross entropy term $H(P(j|x_i^{(v)}, \Theta_t^{(v)}), P(j|x_i^{(v)}, \Theta^{(v)}))$ scales Δ down to zero when the views totally agree. Equation 8 expands all cross-entropy terms. At an abstract level, Δ can be thought of as all pairwise Kullback Leibler divergences of the posteriors $P(j|x_i^{(v)}, \Theta^{(v)})$ in all views. Since the cross entropy is convex, Δ is convex, too.

$$\begin{aligned}
&\Delta(\Theta^{(1)}, \ldots, \Theta^{(s)}, \Theta_t^{(1)}, \ldots, \Theta_t^{(s)}) \\
&= \frac{1}{s-1} \sum_{v \neq u} \sum_{i=1}^{n} \Big(H(P(j|x_i^{(v)}, \Theta_t^{(v)}), P(j|x_i^{(u)}, \Theta^{(u)})) \\
&\qquad - H(P(j|x_i^{(v)}, \Theta_t^{(v)}), P(j|x_i^{(v)}, \Theta^{(v)})) \Big) \qquad (7) \\
&= \frac{1}{s-1} \sum_{v \neq u} \sum_{i=1}^{n} \sum_{j=1}^{m} P(j|x_i^{(v)}, \Theta_t^{(v)}) \log \frac{P(j|x_i^{(v)}, \Theta^{(v)})}{P(j|x_i^{(u)}, \Theta^{(u)})} \qquad (8)
\end{aligned}$$

In order to implement the M step, we have to maximize $Q^{MV}(\Theta, \Theta_t)$ given Θ_t. We have to set the derivative to zero. Parameter Θ occurs in the logarithmized posteriors, so we have to differentiate a sum of likelihoods within a logarithm. Theorem 1 solves this problem and rewrites Q^{MV} analogously to Equation 5.

Equation 10 paves the way to a multi-view EM algorithm. The parameters Θ occur only in the log-likelihood terms $\log P(x_i^{(v)}|j, \Theta^{(v)})$ and $\log \alpha_j^{(v)}$ terms, and Q^{MV} can be rewritten as a sum over local functions Q_v^{MV} for the views v. It now becomes clear that the M step can be executed by finding parameter estimates of $P(x_i^{(v)}|j, \Theta^{(v)})$ and $\alpha_j^{(v)}$ independently in each view v. The E step can be carried out by calculating and averaging the posteriors $P^{(v)}(j|x_i, \Theta_t, \eta)$ according to Equation 11; this equation specifies how the views interact.

Theorem 1. *The multi-view criterion Q can be expressed as a sum of local functions Q_v^{MV} (Equation 9) that can be maximized independently in each view v. The criterion can be calculated as in Equation 10, where $P^{(v)}(j|x_i, \Theta_t, \eta)$ is the averaged posterior as detailed in Equation 11 and $P(j|x_i^{(v)}, \Theta_t^{(v)})$ is the local*

posterior of view v, detailed in Equation 12.

$$Q^{MV}(\Theta^{(1)},\ldots,\Theta^{(s)},\Theta_t^{(1)},\ldots,\Theta_t^{(s)})$$
$$= \sum_{v=1}^{s} Q_v^{MV}(\Theta^{(v)},\Theta_t^{(1)},\ldots,\Theta_t^{(s)}) \tag{9}$$
$$= \sum_{v=1}^{s}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} P^{(v)}(j|x_i,\Theta_t,\eta)\log\alpha_j^{(v)}\right. \tag{10}$$
$$\left. + \sum_{i=1}^{n}\sum_{j=1}^{m} P^{(v)}(j|x_i,\Theta_t,\eta)\log P(x_i^{(v)}|j,\Theta^{(v)})\right)$$
$$P^{(v)}(j|x_i,\Theta_t^{(1)},\ldots,\Theta_t^{(s)},\eta) \tag{11}$$
$$= (1-\eta)P(j|x_i^{(v)},\Theta_t^{(v)}) + \frac{\eta}{s-1}\sum_{\bar{v}\neq v} P(j|x_i^{(\bar{v})},\Theta_t^{(\bar{v})})$$
$$P(j|x_i^{(v)},\Theta_t^{(v)}) = \frac{\alpha_j^{(v)}P(x_i^{(v)}|j,\Theta^{(v)})}{\sum_k \alpha_k^{(v)}P(x_i^{(v)}|k,\Theta^{(v)})} \tag{12}$$

The proof of Theorem 1 as well as the detailed derivation of the resulting EM algorithm can be found in [9].

3 Link Farm Discovery

Search engines shape our perception of the web. Striving to maximize the visibility of their businesses, many commercial web sites employ search engine optimization tools that inflate the page rank of a target web page. This is achieved by creating a dense web of pages that point at the target ("link farms"). In order to direct the crawlers of search engines to such a generated web, links are posted to open discussion forums on the web. It is estimated that possibly 75 out of 150 web servers that exist today are operated by search engine optimizers in order to manipulate Google's search results.

In order to maintain the quality of their search results, search engines have to utilize relevance measures that cannot easily be manipulated by the owners of the pages. It is therefore a crucial classification problem to identify link spam. Many features of the URL, the page itself, and properties of the surrounding pages have been identified as being discriminative [10,11]. We have collected a training corpus that consists of 1000 labeled and several thousand unlabeled pages, roughly 50% of the data are link spam. Each example contains a page together with all pages that are connected via inbound and outboud links. We calculate many intrinsic and contextual features.

The classification problem is not only difficult but also adversarial [12]: as soon as search engines employ a filtering technique, search engine optimizers will modify their generating tools to bypass the filter and dodge identification. Copies of the link spam data set can be obtained from the author.

Acknowledgment

This abstract refers to joint work with Steffen Bickel and Isabel Drost. The author is supported by grant SCHE540/10-1 of the German Science Foundation DFG.

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Conference on Computational Learning Theory. (1998) 92–100
2. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proc. of the 33rd Annual Meeting of the Association for Comp. Linguistics. (1995)
3. Bickel, S., Scheffer, T.: Multi-view clustering. In: Proceedings of the IEEE International Conference on Data Mining. (2004)
4. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In: Proceedings of Information and Knowledge Management. (2000)
5. McCallum, A., Nigam, K.: Employing em in pool-based active learning for text classification. In: Proceedings of the International Conference on Machine Learning. (1998)
6. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B **39** (1977)
7. Dasgupta, S., Littman, M., McAllester, D.: PAC generalization bounds for co-training. In: Proceedings of Neural Information Processing Systems (NIPS). (2001)
8. Henzinger, M., Motwani, R., Silberstein, C.: Challenges in web search engines. SIGIR Forum **36** (2004)
9. Bickel, S., Scheffer, T.: Multi-view EM for mixture models. Unpublished manuscript (2005)
10. Davison, B.: Recognizing nepotistic links on the web. In: AAAI Workshop on Artificial Intelligence for the Web. (2000)
11. Fetterly, D., Manasse, M., Najork, M.: Spam, damn spam, and statistics: using statistics to locate spam web pages. In: Proceedings of the International Workshop on the Web and Databases. (2004)
12. Dalvi, N., Domingos, P., Mausam, Sanghai, S., Verma, D.: Adversarial classification. In: Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining. (2004)