Hierarchical Mixtures of GLMs for Combining Multiple Ground Truths

Joseph Reisinger
Department of Computer Sciences
The University of Texas at Austin
joeraii@cs.utexas.edu

Sugato Basu
Google
basu@google.com

Roberto Bayardo
Google
rbayardo@google.com

Abstract

In real-world machine learning problems it is often the case that the gold-standard for a particular learning problem is not accurately reflected by any one particular data set. For example, when modeling the landing-page quality associated with a search result, labels from human evaluators are often biased towards "brand-name" sites, whereas labels derived from conversions can potentially confound search abandonment and successful conversion. In this paper we propose a class of models for characterizing and isolating the *relative bias* of a prediction problem across multiple data sets. These models can be used either as tools for data analysis, with the goal of calculating the divergence to the hypothetical gold-standard, or as smoothing procedures aimed at capturing as much shared structure between the domains as possible.

1 Introduction

We consider a class of learning problems where labeled data for the exact target concept is unavailable, but multiple sources of *related* data are available, and we wish to combine these biased data sources to construct an estimate of the true target concept. Without additional human knowledge, this problem is theoretically intractable, as data from the source domains can be combined arbitrarily. However, by making model assumptions about how the data sets are likely to agree, we can combine them in a way that systematically cancels out individual idiosyncracies. Specifically, using mixtures of simple linear classifiers, we can remove orthogonal bias components from multiple data sources, retaining their shared signal.

Although this problem class shares many features in common with Domain Adaptation and Multi-task learning, it is separate from each of them as it potentially treats *every* source domain as unreliable. Instead, we base our approach on multivariate Related-Studies models [cf. 6] and mixture-of-experts models [cf. 9].

Our particular approach, constrained mixture of generalized linear models (CM-GLM), extends the basic mixture of GLMs framework with additional domain constraints: Each domain inherits mixture components from two sets: (1) shared components common across all domains and (2) domain-specific components designed to capture additional idiosyncracies. We will develop this model in subsequent sections and introduce a simple inference procedure based on conditional EM.

To motivate this work, we will address the specific problem of predicting conditional webpage quality given a search query. Conditional measures of webpage quality can be acquired either automatically, e.g. bounce-rate, conversion-rate, or at some incurred cost, e.g. human subject evaluation. Although human evalution data tends to be more pure in general, it typically demonstrates significant biases, due to the artificial nature of the problem. For example, human evaluators often prefer "big name" websites over smaller ones, despite the fact that they often do not contain as relevant information. Bias is introduced because the raters are asked to "think like a typical web surfer" but cannot actually come to do so.

2 Background

Our approach to modeling multiple ground-truths borrows from previous work in domain adaptation and multi-task learning. Chelba and Acero propose a simple model of domain adaptation where a target model is centered on a source model, e.g. via regularization $\Omega_{\text{CA}} = ||\boldsymbol{\theta}_{\text{target}} - \boldsymbol{\theta}_{\text{src}}||_2^2$ [3]. Note that the fundamental asymmetry of this model makes it less suitable in cases where both domains may be biased.

Daumé proposes a "frustratingly easy" approach to domain adaptation where source $\boldsymbol{\theta}_{\text{source}}$ and target $\boldsymbol{\theta}_{\text{target}}$ corrections to a set of shared weights $\boldsymbol{\theta}_{\text{shared}}$ are fit simulatenously [4]. Our model extends this approach by inferring data-dependent mixture structure.

Hannah et. al propose DP-GLM, a Dirichlet Process mixture of Generalized Linear Models where the data are grouped and each group is fit with a separate GLM [5]:

$$\begin{aligned} F_0 &\sim \text{DP}(\boldsymbol{\alpha}\mathbb{G}_0), \\ \boldsymbol{\theta}_i = (\boldsymbol{\theta}_i^x, \boldsymbol{\theta}_i^y)|F_0 &\sim F_0, \\ X_i|\boldsymbol{\theta}_i^x &\sim f_x(\cdot|\boldsymbol{\theta}_i^x), \\ Y_i|x_i, \boldsymbol{\theta}_i^y &\sim \text{GLM}(\cdot|X_i, \boldsymbol{\theta}_i^y). \end{aligned}$$

where $\boldsymbol{\theta}$ is a parameter vector drawn from set of cluster centers F_0 and f links $\boldsymbol{\theta}$ to the covariate distribution (e.g. $f = \mathcal{N}$ for mixture of Gaussians). This model is capable of capturing heteroscedasticity as the variance of the response function can vary across mixture components. We further extend this model to combine multiple ground truths.

3 Constrained Mixture of GLMs

We add additional constraints to the basic DP-GLM accounting for multiple ground truths. We first describe this *Constrained Mixture* of GLMs (CM-GLM) in terms of a finite model; then derive a Bayesian nonparametric extension.

Data X_i^d from $d \in D$ domains are fit using $D + 1$ sets of mixture components M_d: one component set specific to each domain M_d and one set of components shared across all domains, M_{shared}. This class of models generalizes several previous approaches:

- When $M_d = 0$ for all d, we recover standard DP-GLM with M_{shared} clusters, combining all domains into a single regression problem.
- When $M_{\text{shared}} = 0$ and $M_d \geqslant 1$ for all d we have d independent regression problems.

Other interesting cases include: $M_{\text{shared}}{=}1$, $M_j{>}1$, where the per-domain distributions are more expressive than the shared distribution, and the converse case $M_{\text{shared}}{>}1$, $M_j{=}1$, where the shared distribution is more expressive.

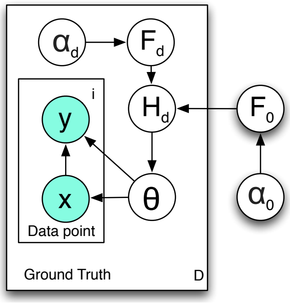
To extend this finite model to the nonparametric case, we borrow mixture combination machinery from [7] and the DP-GLM from [5]. Each ground truth inherits clusters from a shared mixture model F_{shared} and a specific mixture model F_d for capturing any noise or idiosyncratic structure that the ground truth exhibits on top of the shared structure. In particular we replace the cluster distribution $F_0 \sim \text{DP}(\boldsymbol{\alpha}\mathbb{G}_0)$ step from the DP-GLM with

$$\begin{aligned} F_{\text{shared}} &\sim \text{DP}(\boldsymbol{\alpha}_{\text{shared}}\mathbb{G}_{\text{shared}}) \\ F_d &\sim \text{DP}(\boldsymbol{\alpha}_d\mathbb{G}_d) \\ H_d &= \epsilon F_d + (1 - \epsilon)F_{\text{shared}}, \end{aligned}$$

where H_j is a new measure that is the weighted average of F_{shared} and F_d. Since the F are discrete with probability 1, this amounts to reweighting all their components.

Under this model, the clusters from the DPMM draw F_{shared} are shared across all the domains, while F_j accounts for the idiosyncrasies of domain j. The base measures $\{\mathbb{G}_j\}$ and concentration parameters $\{\boldsymbol{\alpha}_j\}$ control the expressiveness of the model. Furthermore, these parameters can be tuned to model any known structure in the ground truth data sets.

This model naturally captures domain-dependent response variance and heteroscedasticity: each mixture component is responsible for every covariate location, and hence parameter dispersion between the different components leads to higher posterior response variance.



4 Case Study: Landing Page Quality

We evaluate CM-GLM on “post-click quality” i.e. determining the relevance of a particular landing page L to a Web search query Q, or Web query / advertisement pair (Q, A).

Our three ground-truth datasets are:

(**Human-Eval; HE**) A small collection of landing page relevance scores made by trained human evaluators. Coverage is low due to the high cost of label acquisition, and although variance is low, these labels exhibit significant bias, especially because raters do not have the context of the queries (e.g., based on the session) of actual web users when they rate the landing page quality w.r.t. just the query.

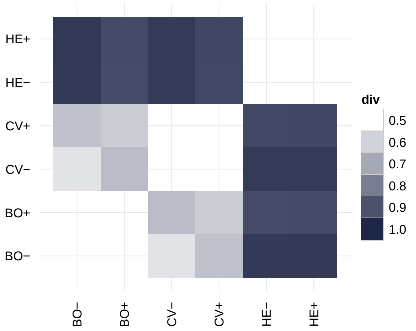
(**Conversion; CV**) The *conversion* after a click measures whether the user performed an action on the landing page that was specified by advertiser as a conversion event, e.g., user makes an item purchase. Since conversions are advertiser-reported, some of the conversion events can be noisy. Furthermore, many landing pages rated highly by human evaluators may not actually track conversions [1, 10].

(**Bounce; BO**) A bounce event on a webpage indicates that the user “bounces off” from the page, i.e., user clicks through, does not find the desirable information on the ad landing page after click-through and then returns relatively quickly to the page from where the original click was issued [8]. Bounces are typically negative indicators of conditional landing page quality. Although easy to collect, Bounce has high variance and is difficult to use as a quality signal directly.

The target concepts for each of the these data sets differ. Despite best efforts in training evaluators, HE data may still contains unwanted biases that keep it from accurately reflecting landing page quality. We apply CM-GLM as a smoothing method to generate a reliable signal indicating landing page quality, focusing on combining Bounce (BO) data with Human-eval (HE) data, in order to address potential biases. Our working assumptions are:

1. HE data is much sparser than CV, so much so that including HE as a feature in an BO-only model would get washed out. But,
2. HE is typically more precise than BO (modulo some bias that we’ll try to model). Finally,
3. BO targets can be a more reliable indicator of ”real-world” landing page quality in some cases, but have significantly higher variance and noise than HE.
4. (Optional) HE and BO have the same data distribution.

Thus our approach is to leverage both data sets in a way that can reduce the bias of HE while *simultaneously* reducing the variance of BO. This is accomplished using the generative CM-GLM, specifically identifying clusters of data points where there is strong agreement between HE and BO and areas where there is little agreement.



Model	Misclassification Rate		
	HE	CV	
GLM HE	0.2859	-	0.4080
GLM CV	0.4838	0.2890	-
GLM BO	0.4330	-	0.2941
CM-GLM HE/CV	0.2944	0.2853	-
CM-GLM HE/BO	0.2942	-	0.3062

Figure 1: (**left**) Empirical divergences between domain hypothesis class pairs (see Note 1). (**right**) Classification performance for CM-GLM vs. single-domain baselines.

Figure 1 shows the empirical divergences computed between each of the domains.[1] In general, the BO/CV domains are more related to each other than to HE, and it is easier to distinguish negative CV/BO instances from HE instances.

Due to the availability of large amounts of data in each domain, the strong bias induced by CM-GLM does not lead to improvement on the test accuracy. Since we are not interested in performance on any of the three domains *per se*, we instead focus on model interpretation; in particular comparing feature coefficient z-scores between the CM-GLM model and standalone regression models.

Furthermore, the shared BO/HE component places strong emphasis on content vertical features in general than BO alone, indicating that expert human raters are more sensitive to content mismatch. Based on these observations, an actionable outcome from our analysis would be to retrain human raters to be more judicious when evaluating apparent content mismatches.

5 Discussion and Future Work

CM-GLM can be applied broadly in situations where *data integration* or synthesis between multiple sources is important. For example, in computer vision it is often necessary to combine different ground truth models of image annotation, e.g. LabelMe and Peekaboom. Since these data sources are collected under different sets of annotator constraints, they show may exhibit significant relative bias.

(**Expressivity**) The model can only assign entire feature vectors to clusters, and hence cannot capture differences in the reliability of feature subspaces. For example, it might be the case that coefficients for certain subsets of features should be shared across multiple sources, but others should not. It would be straightforward to extend the model from a mixture to an admixture to address this problem.

(**Mean-shift bias**) Although the model is capable of identifying idiosyncracies in each model, it is not capable of distinguishing between additive biases and pure noise. For example, HE and BO might both admit similar feature covariance structures, but require additive offsets to the shared structure for some features. This kind of sharing can be expressed using a “frustratingly-easy” style model and could be included in our model as the base model for the shared structure.

References

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2006. ACM.

[1] A classifier is fit to discriminate between each pair of domain hypothesis classes (e.g. HE+/BO- corresponds to learning a model capable of disriminating positive human-eval instances and negative bounce instances); divergence is measured as $D = 1 - m$, where m is the misclassification rate [2].

[2] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems 21*, Cambridge, MA, 2008. MIT Press.

[3] C. Chelba and A. Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, 2006.

[4] H. Daumé III. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007.

[5] L. A. Hannah, D. M. Blei, and W. B. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, pages 1923–1953, July 2011.

[6] P. Muller, G. L. Rosner, M. D. Iorio, and S. Maceachern. A nonparametric bayesian model for inference in related longitudinal studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):611–626, 2005.

[7] P. Mller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric bayesian models. *Journal Of The Royal Statistical Society Series B*, 66(3):735–749, 2004.

[8] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2009. ACM.

[9] M. K. Titsias and A. Likas. Mixture of experts classification using a hierarchical mixture model. *Neural Computation*, 14, 2002.

[10] S. Xu, H. Jiang, and F. C. M. Lau. Mining user dwell time for personalized web search reranking. Barcelona, Spain, 2011. July 16-22.