A Multilevel Bayesian Model of Categorical Data Annotation

Author(s)
Affiliation
Address
xxx@yyy.zzz

Abstract

This paper demonstrates the utility of multilevel Bayesian models of data annotation for classifiers. The observable data is the set of categorizations of items by annotators from which data may be missing at random or may be replicated. Estimated individual-level parameters include category prevalence, the "true" category of each item, and the accuracy in terms of sensitivity and specificity of each annotator. The multilevel parameters represent average annotator performance and variance. Samples from the posterior category distribution may be used for probabilistic supervision and evaluation of classifiers, as well as in gold-standard adjudication and active learning. We demonstrate the effectiveness of our approach with simulated data and two real data sets (RTE-1 and MUC-6).

1 Introduction

The goal of this paper is to demonstrate the utility of multilevel (or hierarchical) Bayesian approaches to modeling the data annotation process (also known as coding, rating, grading, tagging, and labeling). We focus on binary (dichotomous) categorical data annotation, as commonly used to create training and evaluation data for automatic classifiers. Related tasks such as ranking, rescoring, or sequence tagging can often be reduced to classification.

We are able to reliably infer the true categories of items and the underlying accuracy of annotators in terms of sensitivity (accuracy on category 1 items) and specificity (accuracy on category 0 items) given only the set of annotations. We are also able to infer the prevalence of category 1 items, which combined with sensitivity and specificity provide estimated false positive and false negative rates.

The Bayesian models we discuss have the power to jointly reason about all parameters at all levels. The multilevel priors help to smooth individual estimates by combining annotator-specific data with group-level data for all annotators. The group-level data represents average annotator performance as well as inter-annotator variability. The models allow annotations to be missing or replicated at random (a varying panel design).

We perform inference with Gibbs sampling, a Markov chain Monte Carlo (MCMC) method. Gibbs sampling draws parameter assignments from their posterior distributions. This allows general posterior predictive inferences to be made by averaging results over the samples. Posterior distributions over accuracies and prevalence allow Bayesian estimates of agreement statistics such as κ.

Data annotation for classification involves the construction of a coding standard describing how items should be categorized. A coding standard may be nothing more than shared knowledge among a group of annotators. A coding standard typically includes example instances and their corresponding categories. More formal coding standards include written guidelines attempting to specify which items should be assigned to which categories.

In practice, several problems invariably arise. Annotators make mistakes relative to their own understanding of the coding standard. Two annotators might interpret the coding standard differently, each

consistently annotating to a different understanding of the ambiguous standard. This may even be the case for an annotator or group of annotators whose standards evolve over time. Furthermore, coding standards are almost always too vague to determine how every particular item should be annotated.

We represent the impurity in a gold standard through the uncertainty in our category assignments. Measuring this uncertainty is important for deciding how to further annotate data or refine a coding standard.

Uncertainty in category assignments may be propagated directly to applications such as training and evaluating classifiers through multiple imputation or probabilistic supervision. With Gibbs sampling, we create multiple versions of the data set, each with category assignments determined by a Gibbs sample.

2 A Model of Categorical Data Annotation

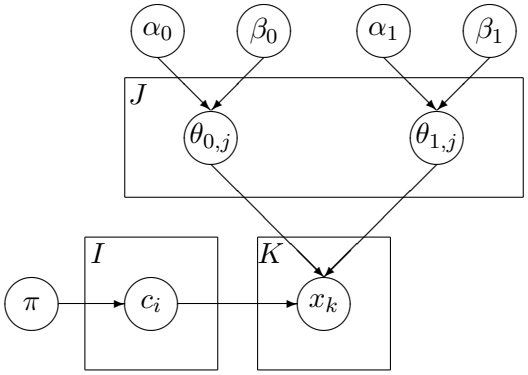
We present our model in Figure 1 in both graphical plate notation and in sampling notation. The rectangular plates represent multiple nodes; I is the number of items being categorized, J is the number of annotators, and K is the total number of annotations.

The variable π represents the prevalence, which is the percentage of categories assigned to category 1. In Bayesian models, all estimated variables require priors. The prevalence π is drawn from a uniform prior, though this is only shown in the sampling notation. We express the uniform distribution as $\mathsf{Beta}(1,1)$, because the beta distribution is the conjugate prior for the binomial. In general, $\mathsf{Beta}(x|\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$. With $\alpha = 1, \beta = 1$, the beta reduces to a uniform distribution.

The variable $c_i \in \{0,1\}$, $1 \leq i \leq I$, represents the true category of item i, and is drawn from a Bernoulli distribution parameterized by the prevalence. In general, $\mathsf{Bernoulli}(c|\theta) = \theta$ if $c = 1$, and $(1-\theta)$ if $c = 0$.

The variables $\theta_{0,j} \in [0,1]$ and $\theta_{1,j} \in [0,1]$, $1 \leq j \leq J$, represent the specificity and sensitivity of annotator j. These are drawn from beta priors with parameters α_0, β_0 and α_1, β_1.

For specifying priors, it is convenient to reparameterize the beta distribution $\mathsf{Beta}(\alpha,\beta)$ in terms of mean $\alpha/(\alpha+\beta)$ and scale $\alpha+\beta$. We put a uniform prior on the mean and a uniform prior on the inverse square scale $1/(\alpha+\beta)^2$. The uniform priors on the means are expressed as $\mathsf{Beta}(1,1)$ priors. The uniform prior on the inverse square scale entails a Pareto prior $\mathsf{Pareto}(1.5)$ on the scale $\alpha+\beta$ (Gelman et al., 2003). In general, $\mathsf{Pareto}(x|\theta) \propto x^{-(\theta+1)}$



$$
\begin{aligned}
c_i &\sim \mathsf{Bernoulli}(\pi) \\
\theta_{0,j} &\sim \mathsf{Beta}(\alpha_0, \beta_0) \\
\theta_{1,j} &\sim \mathsf{Beta}(\alpha_1, \beta_1) \\
x_k &\sim \mathsf{Bernoulli}\ (c_{ii_k}\theta_{1,jj_k} \\
&\qquad + (1 - c_{ii_k})(1 - \theta_{0,jj_k})) \\
\\
\pi &\sim \mathsf{Beta}(1,1) \\
\alpha_0/(\alpha_0+\beta_0) &\sim \mathsf{Beta}(1,1) \\
\alpha_0+\beta_0 &\sim \mathsf{Pareto}(1.5) \\
\alpha_1/(\alpha_1+\beta_1) &\sim \mathsf{Beta}(1,1) \\
\alpha_1+\beta_1 &\sim \mathsf{Pareto}(1.5)
\end{aligned}
$$

Figure 1: *The model in graphical and sampling notations.*

The variables $x_k \in \{0,1\}$, $1 \leq k \leq K$, represent the annotations. There are two index arrays, ii and jj indicating the annotator and item for an annotation, specifically that annotator jj_k labels item ii_k with label x_k. The value x_k is drawn as a Bernoulli with parameter $c_{ii_k}\theta_{1,jj_k} + (1-c_{ii_k})(1-\theta_{0,jj_k})$. This breaks down to $x_k \sim \mathsf{Bernoulli}(\theta_{1,jj_k})$ if the category of the item $c_{ii_k} = 1$, and $x_k \sim \mathsf{Bernoulli}(1-\theta_{0,jj_k})$ if the category $c_{ii_k} = 0$. Specifically, if the category is 1, the annotator has their sensitivity chance of a correct annotation of 1, and if the category is 0, the annotator has their specificity chance of a correct annotation of 0.

In order to create a Gibbs sampler, we must compute the conditional probability of each variable

given all other variables. This model neatly decomposes into three steps: (1) category probability given prevalence, specificity/sensitivity, and data annotation, (2) prevalence given categories, and (3) sensitivity/specificity and beta priors given categories.

For (1), the conditional is straightforward:

$$
\begin{aligned}
&P(c_i = 1|\pi, x, \theta_0, \theta_1) \\
&\quad \propto \pi \textstyle\prod_{ii_k=i} p(x_k|\theta_{1,jj_k}) \\
&\quad = \pi \textstyle\prod_{ii_k=i} x_k\theta_{1,jj_k} + (1 - x_k)(1 - \theta_{1,jj_k})
\end{aligned}
$$

$$
\begin{aligned}
&P(c_i = 0|\pi, x, \theta_0, \theta_1) \\
&\quad \propto \pi \textstyle\prod_{ii_k=i} p(x_k|\theta_{0,jj_k}) \\
&\quad = \pi \textstyle\prod_{ii_k=i} x_k(1 - \theta_{0,jj_k}) + (1 - x_k)\theta_{0,jj_k}
\end{aligned}
$$

Because the beta is conjugate to the binomial, step (2) is straightforward. Sampling (3) decomposes into independent beta-binomial estimates of $\theta_0, \alpha_0, \beta_0$ over category 0 items and $\theta_1, \alpha_1, \beta_1$ based on category 1 items.

Estimates of the multilevel beta parameters (α_0, β_0) and (α_1, β_1) may be used to predict the performance of new annotators. For a new annotator and a batch of items with P positive instances ($c_i = 1$) and N negative instances ($c_i = 0$), the number of true positives is distributed as $\text{Beta-Bin}(\alpha_1, \beta_1, P)$ with true negatives as $\text{Beta-Bin}(\alpha_0, \beta_0, N)$. In general, $\text{Beta-Bin}(k|\alpha, \beta, n) = \int_0^1 \text{Bin}(k|\theta)\text{Beta}(\theta|\alpha, \beta)\, d\theta$. The beta-binomial has a mean of $\alpha/(\alpha+\beta)$ and variance roughly inverse to the scale $\alpha + \beta$.

3 Simulated Data Evaluation

We evaluate our models ability to fit with simulated data.[1] We simulated 20 annotators over 1000 items with a 50% missingness at random rate not balanced by item or annotator. The prevalence was simulated at $\pi = 0.20$, and the multilevel parameters at $(\alpha_0, \beta_0) = (40, 8)$ and $(\alpha_1, \beta_1) = (20, 8)$. The simulated values for θ_0 and θ_1 were drawn from their respective betas.

[1]Simulations, data manipulations and graphical display was performed with R 2.7.1 (R Core Development Team, 2008). Gibbs sampling was performed with WinBUGS 1.4.3 (Spiegelhalter et al., 2008). Communication between BUGS and R and evaluation used R2WinBUGS 2.1.8(Sturtz et al., 2005). Amazon Mechanical Turk services were carried out using their command line tools over the 2008-08-02 API release.

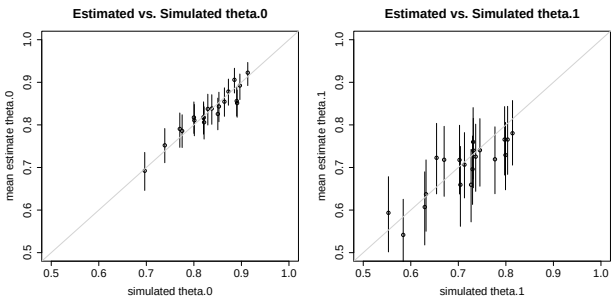


Figure 2: *Posterior mean (circles) and 95% intervals (vertical lines) for θ_0 and θ_1. The 45 degree line represents perfect estimation.*

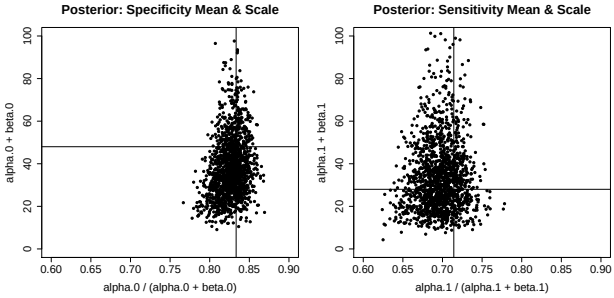


Figure 3: *Posterior specificity and sensitivity beta distributions. Lines are drawn at the simulated values of the parameters.*

For the simulated and real data, we ran the Gibbs samplers multiple times for 1000 iterations from dispersed starting points, discarded the first half of the chains, and computed potential scale reduction ($\hat{R}$) values (Gelman and Rubin, 1992) very close to 1 for all parameters in the remaining 500 samples.

Figure 2 shows how well the posterior estimates for the θ values match the simulated true values. As expected, the posterior intervals for the sensitivity estimates are wider than for specificity, because with a prevalence of 0.2, there are four times more negative items than positive items. Further note that the posterior intrevals are wider than they would be for simple binomials with the same means and number of samples. As annotators accrue more annotations, posterior intervals become tighter. The posterior uncertainty highlights the danger of basing agreement statistics on such uncertain accuracy estimates. Figure 3 shows a scatterplot of posterior samples for the beta parameters reparameterized as scales and means. As with annotator-level effects,

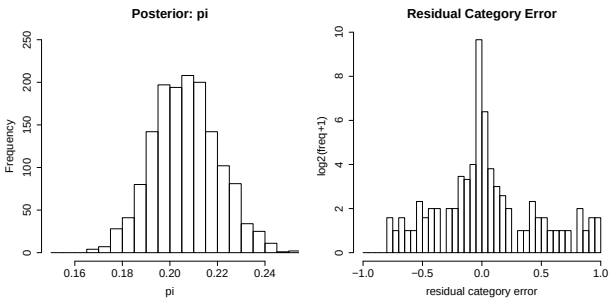


Figure 4: Left: *Posterior histogram of prevalence samples; imulated value* $\pi = 0.2$ *with sample prevalence 0.206.* Right: *Residual error histogram plotting the difference between the true category and estimated probability. Posterior prevalence.*

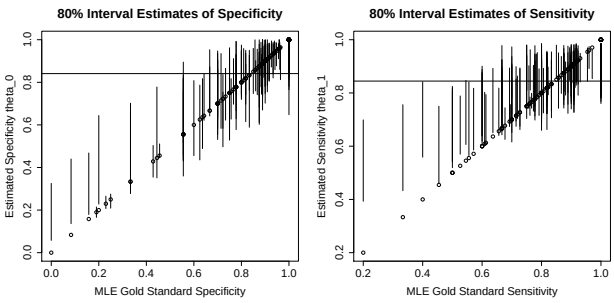


Figure 5: *80% posterior interval estimates of specificity and sensitivity parameters* θ_0 *and* θ_1 *for annotators is plotted versus their sample specificity and sensitivity evaluated against the RTE-1 test data. The circles indicate the specifity and sensitivity against the gold standard, with vertical lines representing 80% posterior interval estimates. The horizontal lines are at the beta prior's estimated means for specificity and sensitivity,* $\alpha_0/(\alpha_0 + \beta_0)$ *and* $\alpha_1/(\alpha_1 + \beta_1)$.

the group-level posteriors are tighter for specificity than sensitivity. Figure 4 (left) shows that the posterior distribution for the prevalence parameter π provides a tight estimate on the simulated value. Figure 4 (right) plots residual error error made by the model, of which 25/1000 items (2.5%) have absolute errors greater than 0.5. These error rates would be smaller if specificities and sensitivities were higher or with more annotators per item. With simulated annotations missing completely at random at a 50% level, some items receive very few annotations and thus remain highly uncertain.

4 RTE-1 Data

(Snow et al., 2008) used the Amazon Mechanical Turk to re-annotate the 800 item test set from the First Recognizing Textual Entailment Challenge (RTE-1) (Dagan et al., 2006). Items consists of a text (e.g. "The city Tenochtitlan grew rapidly and was the center of the Aztec's great empire.") and a hypothesis (e.g. "Tenochtitlan quickly spread over the island, marshes, and swamps."), with the task being to determine if the text implies the hypothesis. (Dagan et al., 2006) were intentionally vague in describing what they meant by entailment in terms of certainty, world knowledge, tense and so forth.

Gold-standard data was balanced ($\pi = 0.5$). Each example was labeled by two annotators with an agreement rate of 80%. The gold standard was made up of the subset of the data in agreement, with 13% more examples later removed.

The top-level instructions provided by Snow et al. to the mechanical Turk annotators were terse, but more specific, restricting world knowledge more than the gold standard implied. They collected ten annotations per item by Amazon mechanical Turk workers. The annotators were not selected for linguistic training. Each annotator performed between 20 and 800 annotations over a span of four days.

In Figure 5, we show posterior 80% intervals for the estimated specificity and sensitivity parameters θ_0 and θ_1. The figure illustrates the strong smoothing effect of the prior, especially on estimates with high uncertainty, which are mostly due to low counts (most annotators labeled 20 items). The 95% posterior interval for prevalence π is $(.45, .52)$; for prior specificity mean $\alpha_0/(\alpha_0 + \beta_0)$ $(.81, .87)$ and scale $\alpha_0 + \beta_0$ $(2.0, 3.9)$; for prior sensitivity mean $\alpha_1/(\alpha_1+\beta_1)$ $(.82, .87)$ and scale $\alpha_1+\beta_1$ $(6.9, 17.6)$.

Figure 6 (left) plots model-based estimates of specificity and sensitivity against performance measured by gold standard, as well as the number of annotations per annotator. Clearly annotators vary with respect to their bias towards 0 or 1 annotations, as well as in overall accuracy.

39% of all annotators performed no better than chance, as indicated by the diagonal green line in Figure 6. The bias of a random annotator is indicated by the position along the line. Our model is able to automatically filter out the effect of the noise. If we remove all annotators with an estimated sensi-

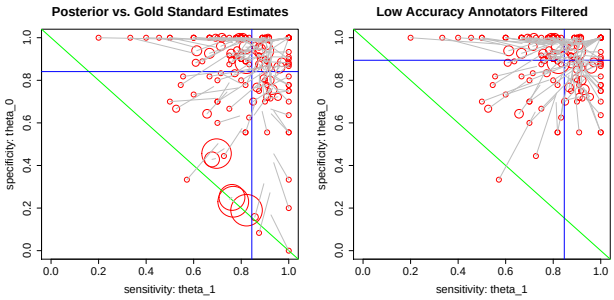


Figure 6: Left: *Each circle represents a different annotator. The circles are centered on the empirical maximum likelihood estimate versus the RTE-1 gold standard and sized proportionally to the number of items annotated. The gray lines run to the model estimates. The horizontal and vertical blue lines are at the beta prior means. The diagonal green line is chance performance.* Right: *Same diagram after removing random annotators and re-estimating.*

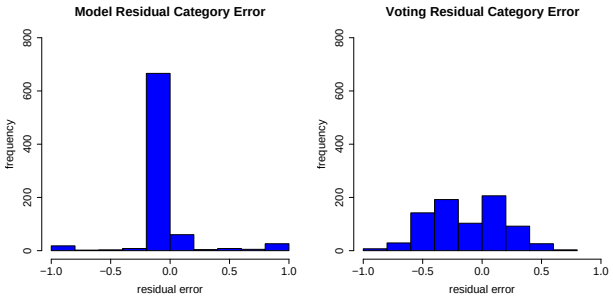


Figure 7: *Residual category errors for model-based category estimates and simple voting-based estimates versus the gold standard.*

tivity or specificity below 50% and refit the model, we have Figure 6 (right). Removing low-accuracy annotators proportionally reduces the number of annotators required to reproduce an "expert" annotation, strengthening the results in (Snow et al., 2008).

The very worst annotator, as measured against the gold standard by being below chance performance, is not filtered, because the prior exerts a strong pull on annotators who only labeled 20 examples. A better pruning strategy would be to remove annotators who did not perform significantly better than chance, for instance, as calculated by a κ statistic versus the gold standard (positive κ is above the green line).

The posterior 95% intervals and mean for prevalence π remain unchanged, but mean prior specificity is now $(.88, .92)$ with scale $(6, 16)$ and sensitivity is now $(.82, .87)$ with scale $(7, 22)$, showing more accuracy and less variance among annotators.

We plot how well our model-based estimates fare against a simple voted estimated versus the gold standard. Both results are plotted in Figure 7. There were 50 cases where voting produced residual error greater than 0.5 and 65 cases which were tied. If ties are broken by coin flipping, that's an expected number of errors of $50 + 65/2 = 82.5$, for an expected accuracy of .897. Taking the most likely categories according to model-based estimates produces 58 errors, for an accuracy of .928. Pruning low-accuracy annotators (without recourse to the gold standard), the model makes 55 errors versus the gold standard, a .931 accuracy. Simple voting after the low-accuracy annotators are removed (based on our initial model fit, but presumably this could be done in some other way) leads to 49 errors and 22 ties, for an expected number of errors of 59.5 and an expected accuracy of .926.

We further estimated a non-hierarchical version of the model with a $\mathsf{Beta}(1, 1)$ prior to simulate maximum likelihood inference; that model had one less error than the hierarchical model. The hierarchical model's smoothing results in a slightly less tight fit to the training data, but hopefully a better estimate of underlying ability and thus more generalizability to future data because of less overfitting to the training data. By pulling accuracy estimates away from 1, it also disperses the category estimates away from 0 or 1.

Annotating a corpus could proceed by continuing to supply annotators until posterior intervals were tight around 0 or 1. Under this approach, some of the categories are still suspect. There are 109 items whose posterior category mean is in (.01,.99), indicating uncertainty in the model. Of the remaining 691 items with mean category posterior below .01 or above .99, 18 have category assignments that do not match the gold standard.

The residual errors were problematic in that at least half of the disagreements seemed to favor the mechanical Turk annotators given our understanding of the coding standard. For instance, the annotators did not conclude that someone reported missing was abducted, or that someone shot was killed, or

that filing for an IPO entailed going public. Perhaps the most problematic case had text "Microsoft was established in Italy in 1985" and hypothesis "Microsoft was established in 1985". The annotators inferred the hypothesis, which world knowledge says is false. But just looking at the text, it's unclear whether "in Italy" is a restrictive modifier, in which case the first sentence is false. Unfortunately, inference is not a relation between texts, but rather requires context and disambiguation.

5 MUC-6 Data

MUC-6 is a corpus that includes named entity mention annotations in sentences of newswire (Grishman and Sundheim, 1995). We had mechanical Turk annotators re-annotate the person-name section of the data, which constituted 190,124 tokens, 4127 of which were part of person names according to the gold standard.

To reduce the problem to a simple token-wise classification problem, we presented an interface with a checkbox below each token, with the instructions "Please check the boxes below all **person names** in the text." Just below that were the instructions "Please **do not** include titles (e.g. "President"), honorifics (e.g. "Mr."), pronouns (e.g. "She"), or names in companies (e.g. "Charles Schwab Corp."), but please **do** include punctuation within names (e.g. "T. S. Eliot")." Data was presented to annotators in complete sentences with at least 400 tokens per form, with all tokens containing a capital letter or punctuation bold-faced to guide the annotators (there were 63,920 bold-faced tokens).

In fitting the model, we restrict attention to the 6753 tokens that were marked as being part of person names by at least one annotator or in the gold standard. This has the effect of over-estimating prevalence with respect to all tokens, which in turn affects the posterior category estimates. All fits were very tight.

A total of 239 annotators participated over a span of four days. Annotator sensitivities and specificites are plotted in the same way as for the RTE-1 data in Figures 8 and 9.

In Figure 10, we show the residual category error by the model contrasted with simple voting. As with the RTE-1 data, the model outperforms simple voting for MUC-6. The total number of absolute residual errors greater than or equal to 0.5 is 233 (96.5% accuracy) according to the model and 242.5 for voting (199 voted errors, 87 tie votes; 96.4% accuracy).

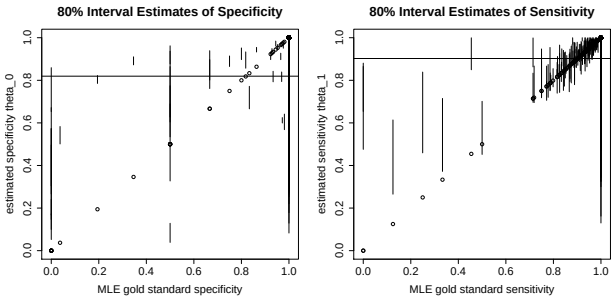


Figure 8: Left: *Named entity specificity residuals.* Right: *Named entity sensitivity residuals. (c.f. Figure 5.)*

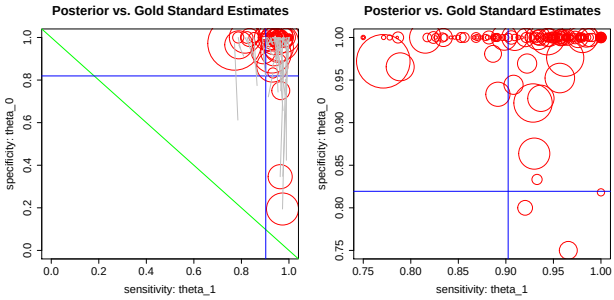


Figure 9: Left: *Named entity sensitivity and specificity residuals and annotation for 50 annotators.* Right: *Upper quadrant, all annotators, without fits. (c.f. Figure 6.)*

Person entity annotation is an easier task than textual entailment, at least as specified. Performance is even better than the 96.5% accuracy versus the gold standard reflects. Error analysis over

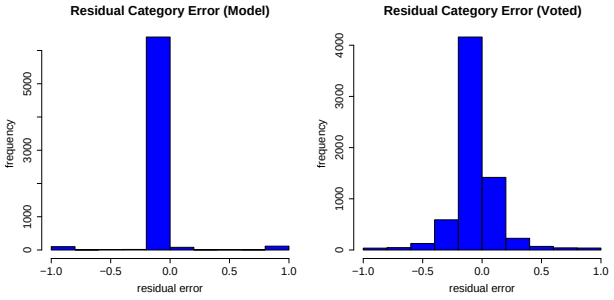


Figure 10: Left: *Category prediction residuals from model.* Right: *Category residusals with simple voting. (c.f. Figure 7.)*

the corpus again shows the mechanical Turk annotators to have correctly assigned tags where the gold standard is in error. The annotators correctly rejected several companies incorrectly annotated as persons in the gold standard, such as "Webster's", "Seagram", "Du Pont", "Buick-Cadillac" and "Moon" used as orgnaizations. They even teased out the plural fictional character "erstwhile Phineas Foggs". They had trouble with some world knowledge, wrongly annotating "Tass" as a person. They also missed name-internal punctuation (e.g. "J. E. "Buster" Brown"), despite the instructions, which we attribute to the checkbox-based user interface.

Elements that were uncertain in the model's estimate were primarily persons used as organizations, such as "the Clinton administration". This measure of uncertainty is a good way to guide the development of annotation standards.

Far more than half of the residual errors versus the gold standard are actually errors in the gold standard itself, indicating an impurity of at least 3% in the gold standard. For the effects of this on evaluating classifiers and hence annotators, see (Lam and Stork, 2003).

6 Bayesian κ

The family of κ "chance corrected agreement" statistics (Cohen, 1960) is widely used, despite a number of well-studied problems due to bias, prevalence, the lack of probabilistic interpretation, and difficulties in extending to multiple annotators or varying panel designs (Reidsma and Carletta, 2008; Artstein and Poesio, In press). The definition $\kappa = (A - E)/(1 - E)$ takes A to be the annotator agreement rate and E the chance agreement rate.

Given annotator sensitivity, specificity and prevalence, we can compute expected agreement between annotators j and j' from the formula:

$$\begin{aligned} A_{j,j'} &= \pi(\theta_{1,j}\theta_{1,j'} + (1-\theta_{1,j})(1-\theta_{1,j'})) \\ &+ (1-\pi)(\theta_{0,j}\theta_{0,j'} + (1-\theta_{0,j})(1-\theta_{0,j'})) \end{aligned}$$

The chance agreement rate can be computed given the overall true prevalence (Siegel and Castellan, 1988):

$$E = \pi\pi + (1-\pi)(1-\pi)$$

or with a slightly more complex formula relative to annotator-specific distributions. With a sequence of posterior samples $\pi^{(n)}, \theta_0^{(n)}, \theta_1^{(n)}$ we can compute a posterior distribution $p(\kappa_{j,j'}|x)$ over the κ statistic between annotators j and j'.

7 Why No Item-Level Predictors?

Although it is possible to extend these models to general item- and annotator-level predictors (Uebersax and Grove, 1993; Qu et al., 1996; Albert et al., 2001) using a logistic or probit generalized linear model, our simulation and real data experiments have shown that even with ten annotators per item, item-level difficulty parameters have very wide posteriors, rendering their Bayesian predictions very similar to models without item-level difficulty parameters. Furthermore, item-level difficulty parameters allow a second explanation of errors, greatly widening the posteriors on annotator accuracy.

Another approach that has been tried is a mixture model of regular and "easy" items (Espeland and Handelman, 1989; Albert et al., 2001; Klebanov et al., 2008). Such a model resembles a zero-inflated Poisson (Gelman et al., 2003) and improves the fit of posterior marginal all-1 and all-0 item annotations. It is otherwise rather unrealistic in the setting where there are many noisy annotators.

8 Multinomials and Ordinals

It is straightforward to extend these models to the multinomial setting. The Bernoulli distributions for prevalence and category are replaced with corresponding discrete (multinomial) distributions, and the beta priors are replaced with Dirichlet priors (Dawid and Skene, 1979). To properly estimate the Dirichlet and multinomial parameters will require more data, but otherwise nothing changes.

Ordinal outcomes, as involved in sentiment and other rating tasks, may be peformed with ordinal logistic regression (Uebersax and Grove, 1993).

9 Previous Work

(Dawid and Skene, 1979) introduced a multinomial model where annotator's responses vary by category, thus generalizing the notion of sensitivity and specificity. They used EM to find maximum likelihood point estimates. (Hui and Walter, 1980)

discusses identifiability of this model. (Mendoza-Blanco et al., 1996) apply Bayesian inference to the uncertainty in sensitivity and specificity estimates in estimating prevalence in a model like Dawid and Skene's.

(Smyth et al., 1995) apply Dawid and Skene's model to estimate gold standards from annotations by four geologists of radar images collected by the Magellan spacecraft of Venus for volcanos. (Smyth, 1995) shows that probabilistic supervision works well for simulated data.

(Bruce and Wiebe, 1999) perform an EM estimate over a model very similar to Dawid and Skene's to estimate annotator accuracies and prevalence from which they could assign a single true category for each item in a corpus.

(Espeland and Handelman, 1989) introduce latent classes for items that are unambiguously positive or negative in the sense that all tests are expected to return the same result. (Klebanov et al., 2008) apply a similar mixture of easy and regular cases to filter unreliable examples from a gold standard.

(Joseph et al., 1995) introduce a binomial model for sensitivity and specificity with independent beta priors. They estimate priors using moment matching from a collection of clinician's estimates. With only two annotators, the priors exert a strong influence.

(Uebersax and Grove, 1993) introduce a variant of the item-response model (Lord, 1980; Rasch, 1980) in which categories are unknown (latent) to the problem of agreement analysis in an ordinal response setting. They model items with a single latent trait generated by binary normal mixtures of 0 and 1 items. Annotators are conveniently modeled by a bias and a variance (error) term, which forces all accuracies to yield non-negative κ statistics. They also discuss missing data, as well as replicated data (multiple annotations of a single item by a coder).

(Qu et al., 1996) introduce another item-response-like model, but separately model annotator sensitivity and specificity. They hand-tune item-difficulty interaction parameters given known correlations among clinical diagnostics (e.g. blood tests). (Dendukuri and Joseph, 2001) apply Bayesian inference to Qu et al.'s model, again fixing priors by interviewing human experts.

(Basu et al., 2000) develop a Bayesian estimate of the kappa statistic (chance adjusted inter-rater agreement) by supplying a beta prior for binomial responses and then reasoning about values of kappa derived in the posterior beta distribution.

(Sheng et al., 2008) study human annotators for classification problems, analyzing agreement with simple uniformity and independence assumptions, with a goal of directing active assignment of annotators to new items or to relabeling existing items.

(Snow et al., 2008) consider how many noisy annotators would be required to recreate a "gold standard", analyzing five existing natural language corpora with analyses generated from Amazon's Mechanical Turk service. They applied Dawid and Skene's model with Laplace-smoothed estimates based on the known gold standard.

10 Conclusions

The model introduced in this paper generalizes previous models of categorical data annotation (Dawid and Skene, 1979; Joseph et al., 1995; Bruce and Wiebe, 1999) to full Bayesian inference, including group-level parameters and missing/replicated data.

We have shown through simulation that the model parameters are identifiable under realistic conditions. Application to multiple partial annotations in the textual entailment and named-entity annotation domains demonstrates the practical utility of the model. Our category estimates refine the active learning approach of (Sheng et al., 2008).

The resulting output of the model is a multiply imputed data set with a set of parameter samples $\theta^{(n)}$. It remains to be seen whether the simulations in (Smyth, 1995) are borne out with real data for training and evaluating classifiers.

Source Code and Data Distribution

In order to support experimental replication, all of the data and Amazon Mechanical Turk, Java, R, and BUGS source code for this paper is available at `xxx.yyy.zzz`.

Acknowledgments

We would like to thank XXX, YYY and ZZZ for discussion and WWW for funding.

References

Paul S. Albert, Lisa M. McShane, Joanna H. Shih, and the U. S. National Cancer Institute Bladder Tumor Marker Network. 2001. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assyas in bladder tumors. *Biometrics*, 57:610–619.

Ron Artstein and Massimo Poesio. In press. Inter-coder agreement for computational linguistics. *Computational Linguistics*.

Sanjib Basu, Mousumi Banerjee, and Ananda Sen. 2000. Bayesian inference for kappa from single and multiple studies. *Biometrics*, 56:577–582.

Rebecca F. Bruce and Janyce M. Wiebe. 1999. Recognizing subjectivity: a case study of manual tagging. *Natural Language Engineering*, 1:1–16.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Ed. and Psych. Meas.*, 20:37–46.

Ido Dagan, Oren Glickman, and Bernardo Magnini, 2006. *The PASCAL Recognising Textual Entailment Challenge.*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28.

Nandini Dendukuri and Lawrence Joseph. 2001. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, 57(1):158–167.

M. A. Espeland and S. L. Handelman. 1989. Using latent class models to characterize and assess relative-error in discrete measurements. *Biometrics*, 45:587–599.

Andrew Gelman and Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. Chapman Hall/CRC, Boca Raton, Florida, 2nd edition.

Ralph Grishman and Beth Sundheim. 1995. Design of the muc-6 evaluation. In *Proceedings of the Sixth Message Understanding Conerence*, Columbia, Maryland.

S. L. Hui and S. D. Walter. 1980. Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171.

Lawrence Joseph, Theresa W. Gyorkos, and Louis Coupal. 1995. Bayesian estimation of disease prevalance and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Proc. of the COLING 2008 Workshop on Human Judgments in Computational Linguistics*, Manchester.

Chuck P. Lam and David G. Stork. 2003. Evaluating classifiers by means of test data with noisy labels. In *IJCAI*.

Frederic M. Lord. 1980. *Applications of Item-Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, New Jersey.

J. R. Mendoza-Blanco, X. M. Tu, and S. Iyengar. 1996. Bayesian inference on prevalence using a missing-data approach with simulation-based techniques: application to hiv screening. *Statistics in Medicine*, 15:2161–2176.

Yinsheng Qu, Ming Tan, and Michael H. Kutner. 1996. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810.

R Core Development Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Georg Rasch. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, Chicago, expanded edition.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of KDD '08*.

S. Siegel and N. J. Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems*. MIT Press.

Padhraic Smyth, 1995. *Computational Learning Theory and Natural Learning Systems Volume III*, chapter Learning with probabilistic supervision. MIT Press, Cambridge, Massachussetts.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast – but is it good? evaluating nonexpert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, Honolulu, Hawaii.

D. J. Spiegelhalter, A. Thomas, and N. G. Best. 2008. WinBUGS version 1.4.3 user manual. Technical report, MRC Biostatistics Unit.

S. Sturtz, U. Ligges, and A. Gelman. 2005. R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16.

J. S. Uebersax and W. M. Grove. 1993. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49(3):823–835.