Annotation models for crowdsourced ordinal data

Vikas C. Raykar
Siemens Healthcare, Malvern, PA, USA
vikas.raykar@siemens.com

Shipeng Yu
Siemens Healthcare, Malvern, PA, USA
shipeng.yu@siemens.com

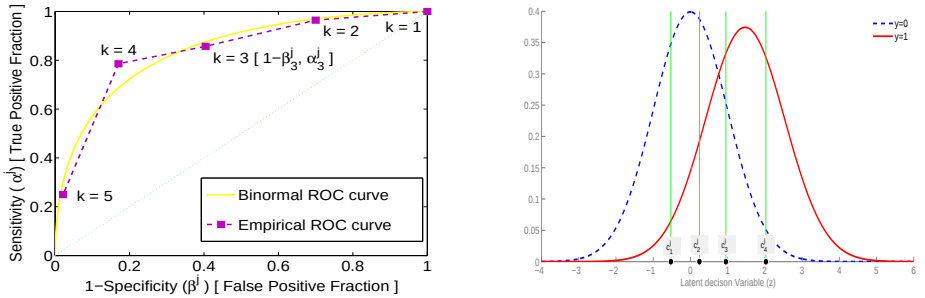
Abstract

In supervised learning when acquiring good quality labels is hard, practitioners resort to getting the data labeled by multiple noisy annotators. Various methods have been proposed to estimate the consensus labels for binary and categorical labels. A commonly used paradigm to annotate instances when the labels are inherently subjective is to use ordinal scales. In this paper we propose annotator models based on Receiver Operating Characteristic (ROC) curve analysis to consolidate the ordinal annotations from multiple annotators. The models lead to simple Expectation-Maximization (EM) algorithms that estimate both the consensus labels and annotator performance jointly. Experiments indicate that the proposed algorithm is superior to the commonly used majority voting rule.

1 Introduction

Most supervised learning algorithms expect a well defined supervision for the training instances in the form of labels. With the advent of crowdsourcing services (Amazon's Mechanical Turk AMT being a prime example) it has become relatively easy and inexpensive to acquire labels from a large number of annotators in a relatively short amount of time. However one drawback of crowdsourcing is that we do not have control over the quality of the annotators. In order to get good quality labels requestors typically get each instance labeled by multiple annotators and these multiple annotations are then consolidated either using a simple majority voting rule or more sophisticated methods that model and correct for the annotator biases.

A commonly used paradigm to annotate instances when the labels are inherently subjective is to use *ordinal scales* where the annotator is asked to rate an instance on a certain discrete ordinal scale say $\{1, \ldots, K\}$. For example, rating a restaurant on a scale of 1 to 5, or assessing the malignancy of a breast lesion on a BIRADS scale [BIR, 1995] of 1 to 5. An ordinal scale expresses rank and there is an implicit ordering in the labels $1 < \ldots < K$. Recently a lot of methods have been proposed for consolidating the *binary* and *categorical* labels from many different annotators that model and correct for the annotator biases [Dawid and Skene, 1979, Smyth et al., 1995, Raykar et al., 2009, 2010] and/or task complexity [Carpenter, 2008, Whitehill et al., 2009, Welinder et al., 2010]. The main contribution of this paper is to extend these ideas to ordinal scales and propose a method to *consolidate the discrete ordinal annotations from multiple annotators*. Our annotator models (§ 2) are based on Receiver Operating Characteristic (ROC) analysis. We propose two models, one based on empirical ROC model and the other a parametric binormal ROC model. Both these models lead to simple EM algorithms that estimate both the consensus labels and the annotator parameters jointly. Experiments on data from different domains indicate that the proposed algorithm is superior to the commonly used majority voting rule. The ROC based models have an added advantage that the annotators can be ranked using the area under the estimated ROC curve. This is especially useful for crowdsourcing services to rank the annotators (for monetary incentives) and to weed out low quality annotators.



(a) ROC models for an annotator (b) Binormal ROC model

Figure 1: (a) The ROC model for an annotator with $K = 5$ ordinal scales. Each annotator is modeled by the sensitivity and specificity for each threshold ($k = 1, \ldots, 5$). The dotted line shows the empirical ROC curve and the smooth line the parametric binormal ROC curve. (b) The latent decision variable binormal ROC model.

We will first discuss two commonly used methods to consolidate ordinal annotations. Let $y_i^1, \ldots, y_i^M \in \{1 < \ldots < K\}$ be the ordinal scales assigned to the i^{th} instance by the M annotators, and let y_i be the actual (unobserved) label.

Direct averaging One method we have sometimes seen used (but incorrect) to consolidate ordinal scales from multiple annotators is to average them, that is, $\widehat{y}_i = (1/M) \sum_{j=1}^{M} y_i^j$. The ordinal values only express an order and hence we cannot directly average the ratings. When using an ordinal scale, the central tendency can be described by the median but the mean cannot be defined [Stevens, 1946]. For example, a BIRADS score of 5 means the lesion is more malignant than a score of say 3, which is more than a score of 1. But we cannot compute the difference between two values, because the difference between 5 and 3 may not be comparable to the difference between 3 and 1.

Categorical labels Sometimes for simplicity the ordinal labels as considered categorical and methods proposed for categorical labels are used to consolidate the annotations [Raykar et al., 2010, Ipeirotis et al., 2010]. Treating ordinal labels as categorical implicitly ignores the order information.

Majority voting rule A commonly used strategy is to use the labels on which the majority of the annotators agree as an estimate of the actual label, $\widehat{y}_i = \arg\max_k \Pr[y_i = k]$ where $\Pr[y_i = k] = (1/M) \sum_{j=1}^{M} \delta(y_i^j, k)$, where $\delta(s,t) = 1$ if $s = t$ and 0 otherwise. Majority voting implicitly assumes that all annotators are equally good. A drawback of crowdsourcing is that we have no control over the quality of the annotators. For example, if there is only one true expert and the majority are novices, then majority voting would favor the novices since they are in a majority.

2 Annotator models for ordinal scales based on ROC analysis

It is conceptually easier to think of the true label to be binary, that is, $y_i \in \{0, 1\}$. For example in mammography a lesion is either malignant(1) or benign(0) and the BIRADS ordinal scale is a means for the radiologist to quantify his uncertainty based on the digital mammogram. The radiologist assigns a higher value of the label if he or she thinks the true label is closer to one.

Empirical annotator ROC model Similar to the annotation models used for binary labels [Dawid and Skene, 1979, Smyth et al., 1995, Raykar et al., 2010] we model each annotator by the *sensitivity* (true positive fraction) and *specificity* (1-false positive fraction), but the main difference is that we now define the sensitivity and specificity for each ordinal label (or threshold) $k \in \{1, \ldots, K\}$. Let α_k^j and β_k^j be the sensitivity and specificity respectively of the j^{th} annotator corresponding to the threshold k, that is, $\alpha_k^j := \Pr[y_i^j \geq k \mid y_i = 1]$ and $\beta_k^j := \Pr[y_i^j < k \mid y_i = 0]$ for $k = 2, \ldots, K$. Note that $\alpha_1^j = 1$, $\beta_1^j = 0$ and $\alpha_{K+1}^j = 0$, $\beta_{K+1}^j = 1$. Hence each annotator is parameterized by a set of $2(K-1)$ parameters $[\alpha_2^j, \beta_2^j, \ldots, \alpha_K^j, \beta_K^j]$. This corresponds to an empirical ROC curve for the annotator as shown in Fig. 1(a). The area under the empirical ROC curve (area under the dotted line in Fig. 1(a)) can be computed as $\text{AUC}^j = (1/2) \sum_{k=1}^{K} (\alpha_{k+1}^j + \alpha_k^j)(\beta_{k+1}^j - \beta_k^j)$, and can be used as a summary metric to rank the annotators, good annotators have an AUC close to one while random annotators (spammers) have an AUC close to 0.5.

Parametric annotator ROC model The *latent decision variable(LDV) framework* [Dorfman and Alf, 1969, Pepe, 2004] is a popular conceptual framework to develop parametric ROC models for ordinal scale data. For each annotator j we assume that there is an unobserved latent continuous variable z_i^j corresponding to the annotator's perception of the i^{th} instance, and there are $K-1$ *decision threshold* values $[c_0^j = -\infty] < c_1^j < \ldots < c_{K-1}^j < [c_K^j = \infty]$ (see Fig. 1(b)). The annotator assigns the label $y_i^j = k$ if $c_{k-1}^j < z_i^j < c_k^j$. Two annotators with the same discriminatory power may perceive the instances similarly but may give different ratings because their internal decision thresholds are different. The *binormal form* [Pepe, 2004] for the ROC assumes that the latent variable z_i^j has a separate normal distribution corresponding to $y_i = 1$ and $y_i = 0$, that is, $z_i^j \mid y_i = 0 \sim \mathcal{N}(z_i^j \mid \mu_0^j, (\sigma_0^j)^2)$ and $z_i^j \mid y_i = 1 \sim \mathcal{N}(z_i^j \mid \mu_1^j, (\sigma_1^j)^2)$, with the assumption that $\mu_1^j > \mu_0^j$. Without loss of generality we can assume that $z_i^j \mid y_i = 0$ is a standard normal, that is, $\mu_0^j = 0$ and $\sigma_0^j = 1$. Hence $z_i^j \mid y_i = 0 \sim \mathcal{N}(z_i^j \mid 0, 1)$ and $z_i^j \mid y_i = 1 \sim \mathcal{N}(z_i^j \mid \mu^j, (\sigma^j)^2)$ with the assumption that $\mu_j > 0$. Under this model for a particular false positive fraction (FPF) t (1-specificity) the true positive fraction (TPF) (sensitivity) can be parameterized as $\text{ROC}^j(t) = \text{TPF}^j(t) = \Phi\left(a^j + b^j \Phi^{-1}(t)\right), \quad t \in [0, 1]$, where we define $a^j = \mu^j/\sigma^j$, $b^j = 1/\sigma^j$, and Φ is the cdf of the standard normal distribution. Hence under the binormal LDV model each annotator is parameterized by $K+1$ parameters, the two ROC parameters (a^j, b^j) and the $K-1$ decision thresholds $(c_1^j, \ldots, c_{K-1}^j)$. This corresponds to a smooth binormal ROC as shown in Figure 1(a). The AUC for the binormal ROC curve is given by [Pepe, 2004] $\text{AUC}^j = \Phi(a^j/\sqrt{1 + (b^j)^2})$ and can be used as a summary metric to rank the annotators.

3 Maximum likelihood parameter estimation via EM algorithm

Let $\mathcal{D} = \{y_i^1, \ldots, y_i^M\}_{i=1}^N$ be the observed N ordinal annotations from M annotators. Let $y_i \in \{0, 1\}$ be the actual (unobserved) binary label for the i^{th} instance and let $p = \Pr[y_i = 1]$ be the (unknown) prevalence of the positive class. Let $\boldsymbol{\alpha}^j = [\alpha_2^j, \ldots, \alpha_K^j]$ and $\boldsymbol{\beta}^j = [\beta_2^j, \ldots, \beta_K^j]$ be the sensitivity and specificity vector for different thresholds for annotator j, with boundary conditions $\alpha_1^j = 1$, $\beta_1^j = 0$ and $\alpha_{K+1}^j = 0$, $\beta_{K+1}^j = 1$. Given all the observed ordinal annotations $\mathcal{D}$ the maximum likelihood estimator of the parameters $\boldsymbol{\theta} = [\boldsymbol{\alpha}^1, \boldsymbol{\beta}^1, \ldots, \boldsymbol{\alpha}^M, \boldsymbol{\beta}^M, p]$ can be effectively computed via the EM algorithm [Dempster et al., 1977] summarized in Algorithm 1 (EM-ROC). Each iteration of the EM algorithm consists of two steps: an Expectation(E)-step and a Maximization(M)-step. The M-step involves maximization of a lower bound on the log-likelihood that is refined in each iteration by the E-step. A similar algorithm can be derived for the binormal LDV model. The only complication is that we do not have a close form solution in the M-step and we have to use a gradient descent procedure.

4 Experimental results

We first illustrate the proposed algorithm on a simulated data containing 500 instances annotated on a scale of 1 to 5 by 3 annotators. The annotators were simulated according to the binormal LDV model. We compare our two proposed algorithms EM-ROC (based on the empirical ROC model) and EM-BLDV (based on the binormal latent decision variable model) algorithms with the majority voting (MV) rule. Figure 2 plots the estimated (solid line) and the actual (dotted line) binormal ROC curve for the three annotators for MV, EM-ROC, and EM-BLDV respectively. It can be seen the MV under estimates the performance of the annotators. In each plot the solid black line shows the ROC curve of the estimated consensus labels. The EM-ROC and the EM-BLDV algorithms are clearly superior to the majority voting rule. We also report experimental results on some publicly available linguistic and image annotation data collected using the Amazon's Mechanical Turk and some medical annotation data. Table 1 summarizes the datasets along with a brief description of the tasks. Table 2 summarizes the results for all the datasets in terms of the AUC of the resulting consolidated ground truth. The results are average over 100 bootstrap replications. The 95% confidence intervals (CI) are also shown. To compute the accuracy we use a threshold of 0.5 on the estimated probabilities. In terms of the AUC the EM-ROC and the EM-BLDV have similar performance and both are superior to the majority voting rule. In practice we prefer EM-ROC since it is much simpler to implement and also numerically more stable than EM-BLDV.

Input: Ordinal annotations $y_i^j \in \{1 < \ldots < K\}$, $j = 1, \ldots, M$(annotators), $i = 1, \ldots, N$(instances)

Outputs:
—soft probabilistic consensus labels $\mu_i = \Pr[y_i = 1 | y_i^1, \ldots, y_i^M]$, $\forall i = 1, \ldots, N$.
—annotator empirical ROC curve parameters $[\alpha_1^j, \beta_1^j, \ldots, \alpha_K^j, \beta_K^j]$ $\forall j = 1, \ldots, M$
—the prevalence of the positive class $p = \Pr[y_i = 1]$
—the empirical AUC^j for each annotator.

Initialize $\mu_i = (1/M) \sum_{j=1}^{M} \delta(y_i^j, k)$ via soft majority voting, where $\delta(s, t) = 1$ if $s = t$ and 0 otherwise.
repeat

M-step Update the model parameters

Update prevalence $p \leftarrow (1/N) \sum_{i=1}^{N} \mu_i$.
Update the annotator ROC curve parameters $\forall j = 1, \ldots, M$, $\forall k = 1, \ldots, K$.

$$\alpha_k^j \leftarrow \sum_{\ell \geq k} \frac{\sum_{i=1}^{N} \mu_i \delta(y_i^j, \ell)}{\sum_{i=1}^{N} \mu_i}, \quad \beta_k^j \leftarrow \sum_{\ell < k} \frac{\sum_{i=1}^{N} (1 - \mu_i) \delta(y_i^j, \ell)}{\sum_{i=1}^{N} (1 - \mu_i)}. \tag{1}$$

E-step Re-estimate the consensus labels

Recompute the probabilistic consensus labels $\forall i = 1, \ldots, N$ as $\mu_i \leftarrow A_i p / (A_i p + B_i (1 - p))$ where

$$A_i = \prod_{j=1}^{M} \prod_{k=1}^{K} [\alpha_k^j - \alpha_{k+1}^j]^{\delta(y_i^j, k)}, \quad B_i = \prod_{j=1}^{M} \prod_{k=1}^{K} [\beta_{k+1}^j - \beta_k^j]^{\delta(y_i^j, k)}. \tag{2}$$

until *convergence*

Compute the area under the ROC curve AUC$^j = \frac{1}{2} \sum_{k=1}^{K} (\alpha_{k+1}^j + \alpha_k^j)(\beta_{k+1}^j - \beta_k^j)$.

A binary label can be obtained by applying a threshold (say 0.5) on μ_i. In the experiments reported in this paper we used a convergence criterion based on the change in the model parameters. The AUC can be used to sort the annotators.

Algorithm 1: EM-ROC: The proposed EM algorithm for the empirical ROC model.

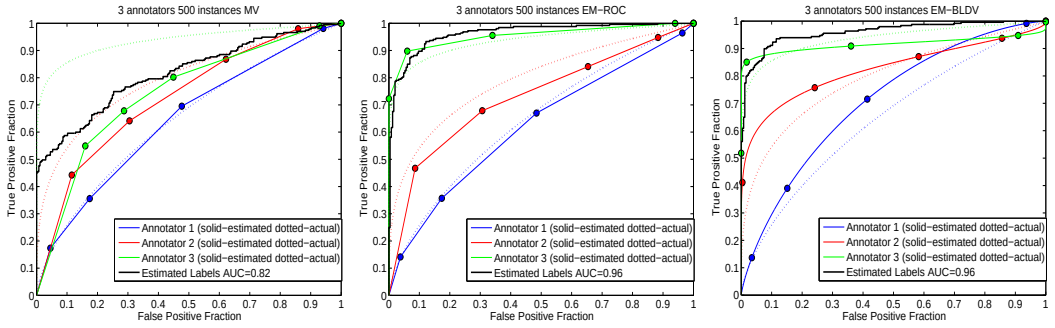


Figure 2: Results for a simulated data containing 500 instances annotated on a scale of 1 to 5 by 3 annotators. Each plot shows the estimated (solid line) and the actual (dotted line) ROC for the three annotators for the majority voting (MV) and the two proposed algorithms, EM-ROC and EM-BLDV. In each plot the solid black line is the ROC curve of the estimated consensus labels.

References

Amazon Mechanical Turk. URL `https://www.mturk.com`.

Breast imaging reporting and data system. *American College of Radiology*, 1995.

B. Carpenter. Multilevel bayesian models of categorical data annotation. Technical Report available at http://lingpipe-blog.com/lingpipe-white-papers/, 2008.

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

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.

Table 1: *Datasets* N is the number of instances and M is the number of annotators. M^* is the mean/median number of annotators per instance. N^* is the mean/median number of instances labeled by each annotator.

Dataset	Type	N	M	M^*	N^*	Brief Description
breast	ordinal[1 5]	75	3	3/3	75/75	breast cancer[Proprietary] Each radiologist reviews the breast MRI data and assesses the malignancy of each lesion on a BIRADS scale.
colon	ordinal[0 10]	420	7	7/7	420/420	colon cancer[Proprietary] Each radiologist reviews a segment of the colon and assesses the malignancy on a scale of 0(no cancer) to 10.
wosi	ordinal[0 10]	30	10	10/10	30/30	word similarity [Snow et al., 2008] Numeric judgements of word similarity.
surprise	ordinal[0 100]	100	38	10/10	26/20	affect recognition [Snow et al., 2008] Each annotator is presented with a short headline and asked to rate it on a scale of [0 100] for the emotion surprise.

Table 2: *Comparison of the various methods.* MV is the soft majority voting algorithm. EM-ROC and EM-BLDV are the proposed EM algorithms. The results are averaged over 100 bootstrap replications. The 95% confidence intervals are also shown.

Data	AUC			Accuracy		
	MV	EM-ROC	EM-BLDV	MV	EM-ROC	EM-BLDV
breast	0.75 0.46 0.89	0.89 0.77 0.98	0.91 0.79 0.98	0.62 0.51 0.72	0.84 0.71 0.93	0.83 0.65 0.92
colon	0.70 0.60 0.77	0.88 0.78 0.93	0.93 0.89 0.98	0.88 0.85 0.91	0.94 0.92 0.96	0.91 0.85 0.96
wosi	0.95 0.85 1.00	0.98 0.89 1.00	0.96 0.86 1.00	0.44 0.23 0.60	0.95 0.87 1.00	0.89 0.70 1.00
surprise	0.61 0.42 0.85	0.77 0.56 0.91	0.76 0.56 0.90	0.88 0.81 0.95	0.79 0.69 0.87	0.86 0.78 0.93

D. D. Dorfman and E. Alf. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. *Journal of Mathematical Psychology*, 6(487-496), 1969.

P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, 2010.

M. S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2004.

V. C. Raykar, S. Yu, L .H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *ICML 2009: Proceedings of the 26th International Conference on Machine Learning*, pages 889–896, 2009.

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, April 2010.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems 7*, pages 1085–1092. 1995.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and Fast—but is it good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, 2008.

S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684), 1946.

P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pages 2424–2432. 2010.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043. 2009.