Toward an Information Theoretic Approach to Managing Multiple Decision Makers

Edwin Simpson, Steven Reece
Department of Engineering Science, University of Oxford, UK.
{edwin, reece}@robots.ox.ac.uk

Sarvapali Ramchurn
School of Electronics and Computer Science, University of Southampton, UK.
sdr1@ecs.soton.ac.uk

Stephen J. Roberts
Department of Engineering Science, University of Oxford, UK.
sjrob@robots.ox.ac.uk

Abstract

Citizen science and human computation involves working with multiple, untrusted decision makers. We demonstrate how Bayesian Classifier Combination outperforms a naive Bayes method when classifying documents using unreliable crowdsourced labels. We also present methods for screening workers and selecting informative documents to label. Finally, we explain how the Bayesian Classifier Combination model could be used to unify these steps into a single information theoretic approach.

1 Introduction

For pattern recognition and knowledge-based tasks, human workers often produce good results with only a small set of instructions. In citizen science and other human computation problems, we often need to process large datasets. These typically require large numbers of workers of varying ability, and may involve training artificial decision makers using human-generated examples. In section 2, we compare two methods that combine a small proportion of crowdsourced labels with machine-generated features to classify a large dataset. The first method takes a traditional classification approach with separate training and testing stages. The second method, Independent Bayesian Classifier Combination (IBCC) runs in a single semi-supervised stage and is able to learn the reliability of individual human workers, which it then uses to improve performance of the classifier.

We apply the methods to a text classification problem, the TREC crowdsourcing track\textsuperscript{1}, where documents must be marked as relevant or irrelevant to a particular search query. In this application, the number of labels we can collect is limited by budget. In Section 3 we therefore introduce a trust mechanism to ensure that only informative workers are employed, that each task is priced according to its complexity, and that we informative documents are crowdsourced. The final Section 4 explains how these different steps of managing human workers and combining their output can be integrated by taking an approach that maximises information gain over variables in the IBCC model.

\textsuperscript{1}Part of the Text Retrieval Conference (TREC) run by the US National Institute of Standards and Technology https://sites.google.com/site/treccrowd/
2 Comparison of IBCC with a Traditional Classification Approach

Members of the crowd can be seen as imperfect base classifiers, where each individual has their own probability of producing a particular label. Features occur with different probabilities given each topic, so can also be seen as base classifiers ranging from highly indicative of a topic to completely uninformative. IBCC learns the probabilities of features and crowd responses, simultaneously combining them to infer topic probabilities. This differs from traditional classification, which has separate training and prediction phases. The one-stage classifier is a semi-supervised approach that allows latent structure in the unlabeled documents to influence the results, which is particularly effective when there is limited training data.

2.1 IBCC Model

Figure 1: Graphical Model for IBCC. The shaded node represents observed values, circular nodes are variables with a distribution and square nodes are variables instantiated with point values.

Figure 1 shows the graphical model for IBCC [1]. For a set of documents indexed from 1 to \( N \), the \( \text{i} \)th document has topic \( t_i = 1...J \) that we wish to infer, where \( J \) is the number of topics. We assume \( t_i \) is generated from a multinomial distribution with class probabilities \( \kappa : p(t_i = j|\kappa) = \kappa_j \). For the challenge problem we have assumed that the topics are mutually exclusive based on our understanding of the topics, and that there also exists a “none of the above” topic for documents of unknown topic. The total number of base classifiers (features and turkers) is \( K \) and the set of values assigned by base classifiers to documents is \( c \). A turker \( k \) may produce a label \( c_i^{(k)} \) with values \( l = 1...L \), where \( L \) is the number of topics. A feature \( k \) can have values \( c_i^{(k)} \) of either 0 or 1. The value \( c_i^{(k)} \) is assumed to be generated from a multinomial distribution dependent on the topic of document \( i \), with parameters \( \pi_j^{(k)} : p(c_i^{(k)} = l|t_i = j, \pi_j^{(k)}) = \pi_j^{(k)} \). IBCC assumes conditional independence between base classifiers. Parameters \( \pi_j^{(k)} \) and \( \kappa \) have Dirichlet prior distributions with hyper-parameters \( \alpha_{0,k} = [\alpha_{0,j1},...\alpha_{0,JK}] \) and \( \nu = [\nu_{0,1},...\nu_{0,J}] \) respectively.

Defining \( \Pi = \{ \pi_j^{(k)}|j = 1...J, k = 1...K \} \) and \( A_0 = \{ \alpha_{0,j}|j = 1...J, k = 1...K \} \), the joint distribution is

\[
p(\kappa, \Pi, t, c|A_0, \nu) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_i^{(k)} p(\kappa|\nu) p(\Pi|A_0).
\]  (1)

A key feature of IBCC is that \( \pi_j^{(k)} \) represents a confusion matrix that quantifies the reliability of each crowd member and feature. The prior distribution over each \( \pi_j^{(k)} \) is specified through the hyper-parameters \( A_0 \), which can be regarded as pseudo-counts of prior observations of base classifier outputs. Through our choice of these hyper-parameters, we assume that turkers are likely to be informative but not perfect, allowing us to perform inference without observing true topic labels directly.

We perform inference over the model to predict the unknown variables \( t, \Pi \) and \( \kappa \) using variational Bayes (VB). VB is a principled approximate Bayesian method that follows an iterative procedure similar to Expectation Maximisation. The equations used to estimate posterior distributions over the unknown variables are given in Appendix A.
2.2 Naive Bayes Classifier: Two-Stage Classifier

We compared IBCC to a simpler, more traditional approach using two-stage classification. In the first stage, labeled training data is used to build models for the probability of each class and the probability of features conditioned on each class. The second stage uses this model to infer class probabilities of unlabeled documents. In this Naive Bayes classifier we assumed each Turker was correct in its classification of a document. We also tested a modified classifier in which Turkers are assumed to be unreliable using the trust mechanism defined in Section ??.

We assume that the true label \( t \) is generated from a multi-nomial distribution with probability \( \kappa \): 
\[
p(t = j \mid \kappa) = \kappa_j.
\]
We also assume that the observed features \( c \) are binary so that \( c_i = 1 \) indicates the presence of feature \( i \) in a document and \( c_i = 0 \) indicates its absence. Thus, the observed features are generated from binomial distributions dependent on the class of the true label with parameters \( \beta \): 
\[
p(c_i = b \mid t = j, \beta) = \beta_{jib}.
\]
The parameters \( \beta \) have Beta prior distributions with hyperparameters \( \mu \) and \( \kappa \) have Dirichlet prior distributions \( \nu \). Values for \( \mu \) and \( \nu \) are inferred from the training data as we will demonstrate. Given some test (or unlabeled) document the distribution over the true label \( t^* \) of this document given the features \( c^* \) for that document is,
\[
p(t^* = j \mid c^*, \mu, \nu) = \int_{\pi} \int_{\kappa} p(c^* \mid t^* = j, \beta)p(t^* = j \mid \kappa)Be(\beta \mid \mu)Dir(\kappa \mid \nu) d\beta d\kappa
\]
\[
= \left[ \int_{\beta} p(c^* \mid t^* = j, \beta)Be(\beta \mid \mu) d\beta \right] \int_{\kappa} p(t^* = j \mid \kappa)Dir(\kappa \mid \nu) d\kappa
\]
\[
= \prod_t \left[ \int_{\beta_{jib}} \beta_{jib}Be(\beta \mid \mu) d\beta \right] \int_{\kappa} \kappa_j Dir(\kappa \mid \nu) d\kappa
\]
In the training stage, the labeled data is used to infer the expected likelihoods \( \beta \) and class probabilities \( \kappa \). For any variable \( \theta \) and multi-variate \( \phi \), \( \int \theta Be(\theta \mid \cdot) d\theta \) and \( \int \phi Dir(\phi \mid \cdot) d\phi \) are the means of a Beta and Dirichlet distribution, respectively. Assuming uninformative priors for \( \beta \) and \( \kappa \) and training data \( \{t, c\} \), the posterior means are:
\[
\int \beta_{jib} Be(\beta \mid \mu(t, c)) d\beta = \frac{N^{(i)}_{jb} + 1}{\sum_{d=0}^{1} N^{(i)}_{jd} + 1}, \quad \int \kappa_j Dir(\kappa \mid \nu(t, c)) d\kappa = \frac{N_j + 1}{\sum_{k} (N_k + 1)}
\]
where \( N^{(i)}_{jb} \) is the number of documents labeled \( j \) that contain feature \( c_i = b \) in the training data, \( \{t, c\} \), and \( N_j \) is the number of documents labeled \( j \).

Stage two is the prediction phase and for test features \( c^* \) from a document the probability that the document is class \( j \) is,
\[
p(t^* = j \mid c^*, \beta, \nu) = \prod_t \left[ \frac{N^{(i)}_{jc} + 1}{\sum_{d=0}^{1} N^{(i)}_{jd} + 1} \right] \left[ \frac{N_j + 1}{\sum_{k} (N_k + 1)} \right]
\]
In summary, we can determine the posterior document class probability, \( p(t^* = j \mid c^*, \beta, \nu) \) exactly and efficiently using only the sufficient statistics of Beta and Dirichlet distributions.

2.3 Results

3 Worker Trust Mechanism and Crowdsourcing Process

The classification task required a reasonable level of skill and attention, as workers needed to read every line of text to determine the topic. As a result, we found that many workers did not produce useful classifications so implemented a screening process to prevent paying workers for uninformative responses. The screening process involved crowdsourcing 10 ‘gold’ tasks, i.e. tasks that were previously labeled by trusted individuals. The labeling task required the turker to read the document which was divided into three sections, paste a key sentence from each section into the page, and choose a topic for the document as well as assign a confidence level to the label she gave.

Given this, the trust in each turker was determined using the following formula:
\[
T(w) = (cc + G * cn + (1 - G) * in + 1)/(N + 2)
\]
where $T : \mathcal{N} \rightarrow \mathbb{R}$, $cc$, is the number of tasks where the turker was correct and confident, $cn$ is the number of tasks where the turker was correct but not confident, $in$ is the number of tasks where the turker was incorrect and not confident. $N$ is the total number of tasks they actually performed and $G$ balances the importance of the confidence in correct and incorrect answers. We also disqualified those that did not fill in the required sections.

Given the computed trust levels, we then only allowed those with trust greater than 0.65 to complete the unlabeled tasks, which resulted in approximately 40% of the turkers being available to do tasks. This level was chosen because it ensures that the workers are better than random ($\gg 0.5$) and but still provides a large enough group to process a sufficient number of tasks in the time available.

Given a set of screened workers, we then wished to select the most informative documents to label. Labeling documents with features for which we have few previous examples is likely to produce a high information gain. Therefore, we first clustered documents to produce maximally different groupings, then selected the document from each cluster with highest entropy.

These documents were then chunked into parts not more than 20KB in order to make sure they were readable within 5 mins. Then, depending on their size, a payment level was determined for the labeling tasks required. We chose the following payment levels depending on size: $0.03$ for tasks less than 5KB, $0.05$ for tasks between 5KB and 15KB, and $0.08$ for tasks beyond 15KB. These numbers were chosen so that workers would quickly choose and complete these tasks. In initial test with payments less than $0.03$, tasks were not accepted by workers even after several hours at different points in a given week.

![Figure 2: Trust levels of the 42 turkers that performed our gold tasks.](image)

### 4 Toward an Information Theoretic Approach Using IBCC

**Acknowledgements**

We would like to thank UK research council EPSRC for project ORCHID, grant EP/I011587/1.

**References**


A Inference over IBCC using Variational Bayes

The unknown variables $t$, $\Pi$, and $\kappa$ can be estimated using maximum a posteriori (MAP) estimation [2] or inferred in a Bayesian manner [1] using Gibbs Sampling [3]. Here we apply a principled approximate Bayesian method, variational Bayes (VB) [4], as this converges rapidly to a good approximate solution [5]. VB can be seen as a Bayesian generalisation of the Expectation-Maximisation (EM) algorithm [6].

For the variational treatment of IBCC, VB-IBCC, we assume an approximate form of the posterior distribution over the unknown variables:

$$ q(\kappa, t, \Pi) = q(t)q(\kappa, \Pi) $$

The VB algorithm runs by first initialising the variables $E[\ln \pi_{j;i}]$ and $E[\ln \kappa]$ from their prior distributions. We then iterate over a two-stage procedure, updating each of the factors $q(t)$ and $q(\kappa, \Pi)$ to their optimal values in turn, using the current expectations over parameters not in that factor.

The optimal factor $q^*(t)$ for the topics is defined by taking the joint distribution in Equation (1) and absorbing terms not dependent on $t$ into the normalisation constant:

$$ \ln q^*(t) = E_{\kappa, \Pi}[\ln p(\kappa, t, \Pi, e)] + \text{const.} $$

This factorises into independent data points:

$$ \ln \rho_{ij} = E_{\kappa_j, \pi_j}[\ln p(\kappa_j, t_i, \pi_j, e)] = E_{\kappa}[\ln \kappa_j] + \sum_{k=1}^{K} \sum_{l=1}^{L} p(c_i^{(k)} = l) E_{\pi_j}[\ln \pi_{j,c_i^{(k)}}]. $$

For a base classifier that is a turker, the value of $c_i^{(k)}$ is known and $p(c_i^{(k)} = l)$ is either 0 or 1. However, for features we receive probabilities rather than discrete outputs, hence the term $p(c_i^{(k)} = l)$ in the equation above. We then obtain the approximate probability of a topic, which also gives its expected value:

$$ q^*(t_i = j) = E_{t_i = j} = \frac{\rho_{ij}}{\sum_{c=1}^{N} \rho_{ic}}. $$

For the parameters of the model we have the following optimal factor. Terms involving $\kappa$ and terms involving each confusion matrix in $\Pi$ are independent, so we can factorise $q^*(\kappa, \pi)$ further into

$$ q^*(\kappa, \Pi) = q^*(\kappa) \prod_{k=1}^{K} \prod_{j=1}^{J} q^*(\pi_{j}^{(k)}). $$

In IBCC we assume a Dirichlet prior for $\kappa$, which gives us a Dirichlet posterior for the optimal factor

$$ q^*(\kappa) \propto \text{Dir}(\kappa | \nu_1, ..., \nu_J) \quad \nu_j = \nu_{0,j} + N_j $$

where $\nu$ is updated by adding the data counts to the prior counts $\nu_0$ and $N_j = \sum_{i=1}^{N} E[t_i = j]$ is the expected number of documents of each topic. The expectation of $\ln \kappa$ required to update Equation (4) is therefore:

$$ E[\ln \kappa_j] = \Psi(\nu_j) - \Psi\left(\sum_{i=1}^{J} \nu_i\right) $$
where $\Psi$ is the standard digamma function [7]. For the confusion matrices $\pi_{j}^{(k)}$, the priors are also Dirichlet distributions, giving us a posterior Dirichlet distribution of the form
\begin{equation}
q^*\left(\pi_{j}^{(k)}\right) = \text{Dir}\left(\pi_{j}^{(k)} | \alpha_{j1}^{(k)}, ..., \alpha_{jL}^{(k)} \right), \quad \alpha_{jl}^{(k)} = \alpha_{0,jl}^{(k)} + N_{jl}^{(k)}.
\end{equation}

where $\alpha_{j}^{(k)}$ is updated by adding data counts to prior counts $\alpha_{0,j}^{(k)}$, and $N_{jl}^{(k)}$ is defined as
\begin{equation}
N_{jl}^{(k)} = \sum_{i=1}^{N} p(c_{i}^{(k)} = l) \mathbb{E}[t_{i} = j]
\end{equation}
i.e. the number of times that classifier $k$ has assigned value $l$ to a document of class $j$. The expectation required for Equation (4) is given by
\begin{equation}
\mathbb{E}\left[\ln \pi_{jl}^{(k)}\right] = \Psi\left(\alpha_{jl}^{(k)}\right) - \Psi\left(\sum_{m=1}^{L} \alpha_{jm}^{(k)}\right),
\end{equation}