Unsupervised Aggregation for Classification Problems with Large Number of Categories

Ivan Titov, Alexandre Klementiev, Kevin Small, and Dan Roth

{titov@mmci.uni-saarland.de, aklement@jhu.edu, kevin.small@tufts.edu, danr@illinois.edu}

COGNITIVE COMPUTATION GROUP

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CLUSTER OF EXCELLENCE

Aggregation

Multiple experts utilizing different data views, representations, or modeling assumptions are often available for a given task. Aggregating their predictions normally yields more accurate and robust predictions.

Previous work on aggregation makes at least one of the following 3 assumptions:

1. A small number of categories
2. Presence of labeled data
3. Votes of experts are independent conditioned on the true category

For majority of problems with a large number of categories both assumptions (2) and (3) are violated:

- Labeled data is sparse and expensive to annotate
- Even though the set of categories is large, for every example there exists a small subset of categories (confusion set) such that any 'reasonable' expert would predict a category from this set.

Our Approach

We propose a generative model for unsupervised aggregation of experts with a large (possibly infinite) number of categories by relaxing the conditional independence assumption on their votes.

Key aspects:

- Conditional independency assumption is replaced with a weaker exchangeability assumption
- The notion of category types is incorporated to account for variability of the judge expertise depending on the category

We evaluate our method on synthetic data and on a practical task of aggregating syntactic dependency trees.

[1] [Genest & Zidek, Statistical Science (1986)]
[2] [Kahn , Ph.D. thesis, Stanford Univ. (2004)]

Most of this work was done while the authors were at UIUC.

Supported by NSF grant ITR IIS-0428472, DARPA funding under the Bootstrap Learning Program, Excellence Cluster on Multimodal Computing and Interaction (MMCI), Saarbruecken and by MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.

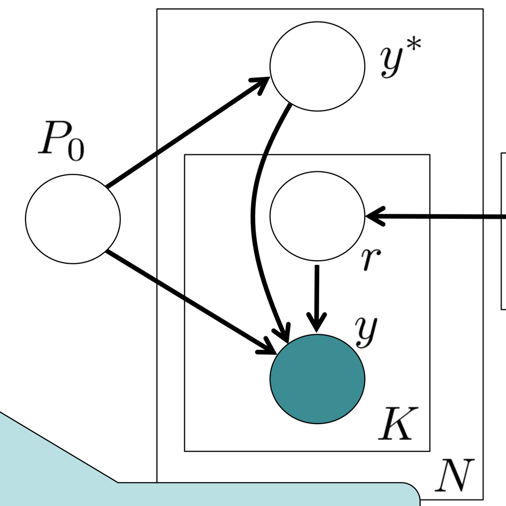
Conditionally Independent Experts [1,2]

- Assume that we observe only categories predicted by K experts $(y_1, \ldots, y_K)$ for N examples ($y_k \in \mathcal{Y}$).

Generative story (for every example) :

Draw the true category $y^* \sim P_0$
For every expert $k \in \{1, \ldots, K\}$
 Decide if the expert is correct $r_k \sim Bernoulli(\theta_k)$
 If $r_k = 1$ then $y_k := y^*$
 else $y_k \sim \frac{1}{(1-P_0(y^*))} P_0, \quad y_k \neq y^*$

Intuitively, the agreement signal is used as surrogate supervision



Accuracy parameter for every expert

In practice, it is often the case that though $|\mathcal{Y}|$ is large for each example x, all the experts predict a set of categories from a small confusion set $\mathcal{Y}_c(x) \subset \mathcal{Y} : |\mathcal{Y}_c(x)| \ll |\mathcal{Y}|$.

For the dependency parsing experiments (below) 23 experts predicted only 3.6 different categories per example out of around $|\mathcal{Y}| \approx 1000$.

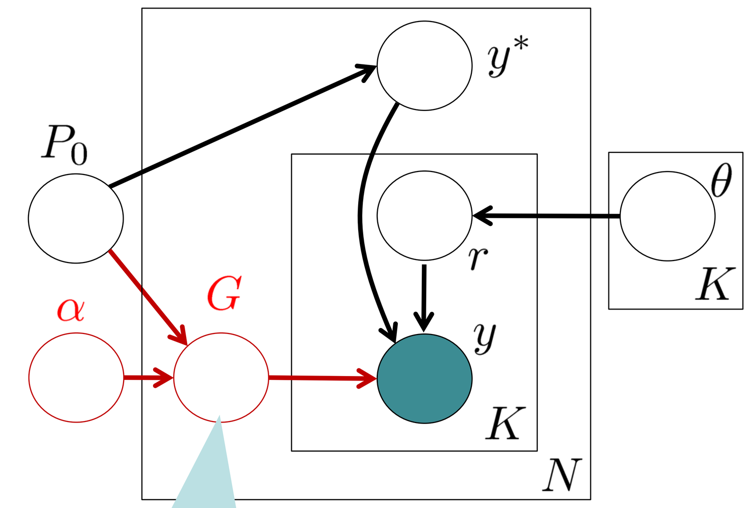
Clearly, this distribution of votes violates the conditional independence assumption. In this case, under fairly general conditions (see details in the paper), predictions of the aggregation model are guaranteed to agree with a majority vote.

The model cannot explain overlaps in predictions unless the corresponding category is the true category

Exchangeable Experts

Generative story (for every example) :

Draw a measure G from a Dirichlet process $DP(\alpha, P_0)$
Draw the true category $y^* \sim P_0$
For every expert $k \in \{1, \ldots, K\}$
 Decide if the expert is correct $r_k \sim Bernoulli(\theta_k)$
 If $r_k = 1$ then $y_k := y^*$
 else $y_k \sim G$



A probability measure modeling the confusion set

The exchangeability assumption, though may not be realistic in all cases, results in a much better approximation of vote distributions, allowing for smaller supports and explaining high agreement between incorrect experts.

One drawback of this approach is that the expertise θ_k of each expert k is assumed independent of the category.

Incorporating Category Types

In this work, we assume that mapping from categories to category types is known. Potentially, it can be inferred.

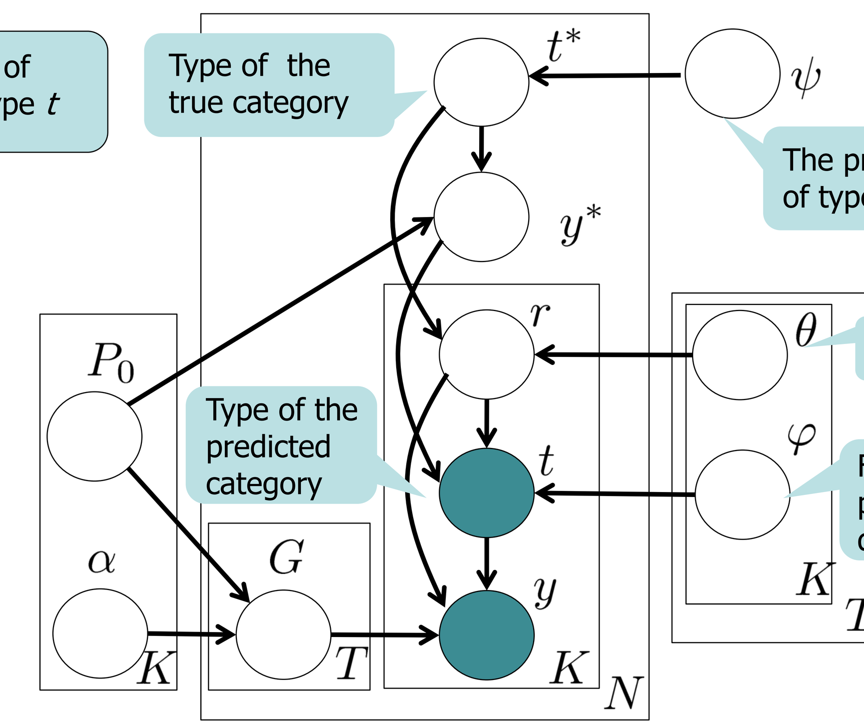
- We assume that there exists a finite and relatively small set of *category types* such that experts' accuracies differ significantly across types but remains constant or similar for categories within each type
- We characterize each expert with two sets parameters:
  - Recall for each type t: $\theta_{\mathbf{k}} = (\theta_{k,1}, \ldots, \theta_{k,T})$
  - Distribution of false positives over types $\varphi_{\mathbf{k}} = (\varphi_{k,1}, \ldots, \varphi_{k,T})$

Note that modeling only the recall parameters is not sufficient:
On a difficult example, where there is little agreement among the experts , the model would tend to predict a category corresponding to the lowest recall parameters, virtually ignoring the vote distribution (see the paper for details).

Generative story (for every example) :

For $t \in \{1, \ldots, T\}$
 Draw measure G_t from a Dirichlet process $DP(\alpha_t, P_{0,t})$
Draw the true type from the distribution of types $t^* \sim \psi$
Draw the true category $y^* \sim P_{0,t}$
For every expert $k \in \{1, \ldots, K\}$
 Decide if the expert is correct $r_k \sim Bernoulli(\theta_{k,t})$
 If $r_k = 1$ then $t_k := t^*$, $y_k := y^*$
 else
  select wrong type $t_k \sim Bernoulli(\varphi_{k,t})$
  select wrong category $y_k \sim G_{t_k}$

The distribution of categories for type t



Type of the true category

The prior distribution of types

Recall

Type of the predicted category

False positives' distribution

Estimation

The EM algorithm is used to estimate parameters of both versions of the model:

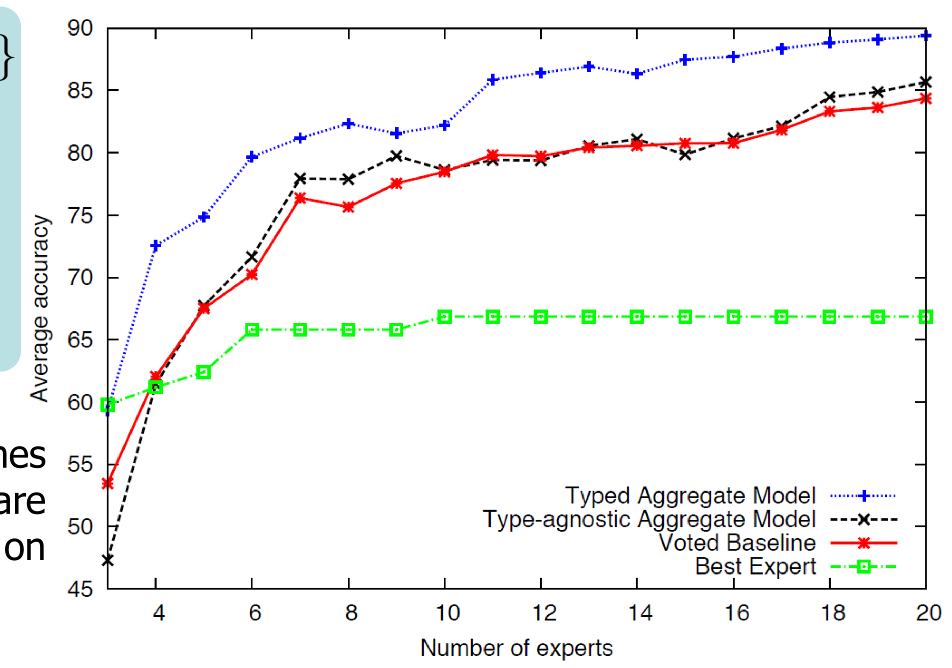
- E-step: the posterior probabilities of y^* (and t^*) are estimated for every example
- M-step: the model parameters are re-estimated to maximize the expected log-likelihood function.

In our experiments, the method appears insensitive to the initialization parameters. It converges in less than 10 iterations (or 1 minute on a standard desktop PC).

Synthetic experiments

- The data is generate from a random Naïve Bayes model
- Each expert is a Naïve Bayes model on a dataset with a randomly selected proportion of category types
- No exchangeability of experts' votes is enforced

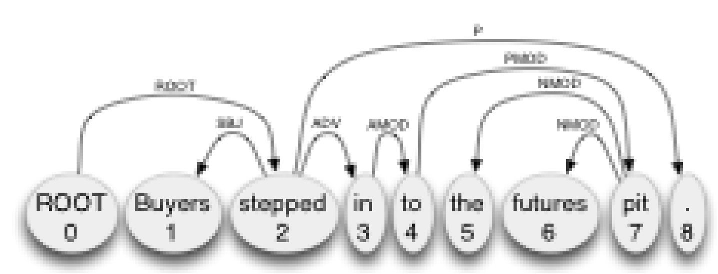
- *Num. of experts:* $K \in \{3, \ldots, 20\}$
- *Number of types:* $T = 3$
- *Number of categories:* $|\mathcal{Y}| = 150$
- *Number of examples:* $N = 1000$
- *Averaged over 5 runs*



Results for the model which assumes conditional independence of experts are significantly below the voted baseline on most experiments.

Aggregating Dependency Parses

- Dependency parse: a graph representing syntactic relations between words in a sentence
- Each word has exactly one syntactic head and a relation label: a single category is a (head, relation) pair
- Incorporating multiple types of dependencies (short/long, root/non-root) is work in progress



- *Experts*: CoNLL-07 shared task participants producing parses for 10 languages
- *N. of parsers*: $K \in \{3, \ldots, 23\}$
- *Number of types*: $T = 1$ (more types: work in progress)