Unsupervised Aggregation for Classification Problems with Large Number of Categories

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

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

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:
1. Labeled data is sparse and expensive to annotate
2. 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.

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:
1. Conditional independence assumption is replaced by a weaker exchangeability assumption
2. The notion of category types is incorporated to account for variability of the judge expertise depending on the category
3. We evaluate our method on synthetic data and on a practical task of aggregating syntactic dependency trees.

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

Conditional Independently Experts

Assume that we observe only categories predicted by K experts (y1, ..., yK) for N examples (yj ∈ D).

Generative story (for every example):
1. Draw the true category y∗ ~ P0
2. For every expert k ∈ {1, ..., K}:
   a. Decide if the expert is correct: r = 1 if r = 0 otherwise yk = y∗
   b. Draw measure ϕk from a Dirichlet distribution

In practice, it is often the case that though |D| is large for each example x, all the experts predict a set of categories from a small confusion set |Y(x)| ≤ |D|.

For the dependency parsing experiments (below) 23 experts predicted only 3.6 different categories per example out of around 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.

Exchangeable Experts

Generative story (for every example):
1. Draw measure G from a Dirichlet process DP(α, P0)
2. Draw the true category y∗ ~ P0
3. For every expert k ∈ {1, ..., K}:
   a. Draw measure ϕk from a Dirichlet distribution
   b. Draw the true category yk ~ P0
4. If r = 1 then yk := y∗
5. Else yk ~ G

The exchangeability assumption, though not 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 of each expert is assumed independent of the category.

Incorporating Category Types

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:
1. Recall for each type t: θk = (θk,1, ..., θk,T)
2. Distribution of false positives over types: ϕk = (ϕk,1, ..., ϕk,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 variant, virtually ignoring the vote distribution (see the paper for details).

Estimation

The EM algorithm is used to estimate parameters of both versions of the model:
1. E-step: the posterior probabilities of y∗ and r are estimated for every example
2. 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 Naive Bayes model
- Each expert is a Naive Bayes model estimated on a dataset with a randomly selected proportion of category types
- No exchangeability of experts’ votes is enforced

Number of experts: K ∈ [3, 20]
Number of types: T = 3
Number of categories: |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 Parsers

- 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

- Experiment CoNLL-07 shared task participants producing parses for 10 languages
- N of parsers K ∈ [3, 20]
- Number of types: T = 1 (more types: work in progress)