Bayesian Aggregation of Categorical Distributions with Applications in Crowdsourcing

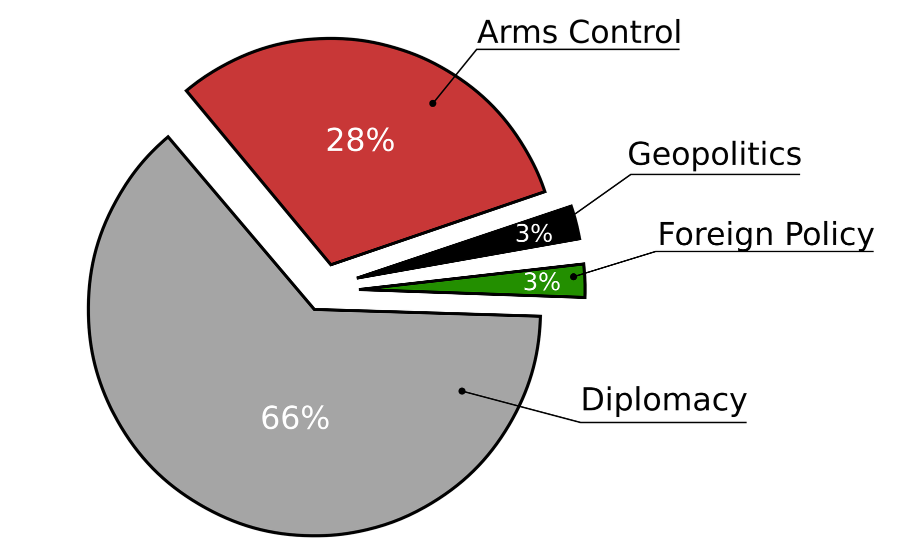
UNIVERSITY OF Southampton

A. Augustin, M. Venanzi, A. Rogers, N. R. Jennings

Motivation

- A whole new range of problems can be solved by aggregating judgments of proportions.
- Applications include information retrieval and recommendation:

Source: BBC News, 2017



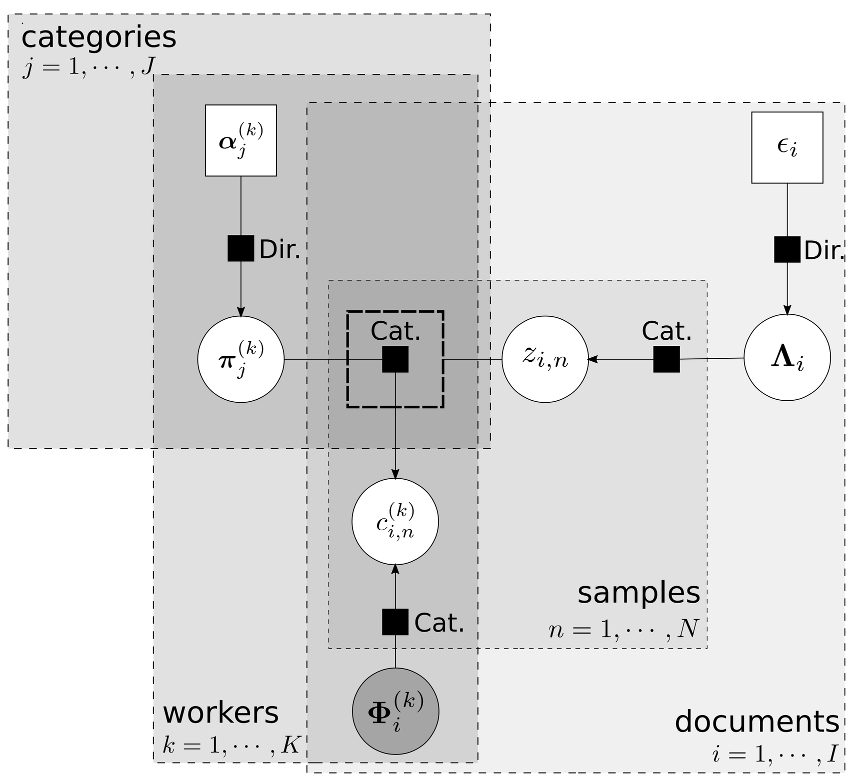
Example judgment of proportions

Challenges

- Malicious participants (i.e. **spammers**) constitute up to 45% of all contributors [1].
- Spammers increase the cost of judgment acquisition and degrade accuracy.

Aggregating Judgments over Multiple Categories (MBCC)

- Key innovations: elicitation and sampling of judgments of proportions in the form of probability distributions [2] to improve on the accuracy of aggregation.
- We extend the Independent Bayesian classifier combination (**IBCC**) [3] to deal with documents having multiple categories.
- Workers' ability and the aggregation are learnt in either an **unsupervised** or **semi-supervised** fashion.



Factor graph of MBCC

Procedure 1 Generative process of MBCC.

```
Input: the confusion matrices Π and the category proportions Λ

for each document i ∈ {1, ..., I} do
    for each sample n ∈ {1, ..., N} do
        Sample z_{i,n} ~ Cat(Λ_i)
        for each worker k ∈ {1, ..., K} do
            Sample c^{(k)}_{i,n} ~ Cat(π^{(k)}_{z_{i,n}})
        end for
    end for

for each worker k ∈ {1, ..., K} do
    for each category j ∈ {1, ..., J} do
        Φ^{(k)}_{i,j} = β^{(k)}_{i,j} Σ_{n=1}^{N} δ(c^{(k)}_{i,n} − j)
        where β^{(k)}_{i,j} is a normalising constant.
    end for
end for

return Φ
```

$\Phi_i^{(k)}$ Judgment of document *i* by worker *k* (categorical dist.)

$z_{i,n} \sim \mathrm{Cat}(\Lambda_i)$ Sampled category *n* from aggregated distribution of document *i*

$c_{i,n}^{(k)} \sim \mathrm{Cat}\left(\pi_{z_{i,n}}^{(k)}\right)$ Sampled category *n* from worker *k*'s judgment of document *i*

$\Lambda_i \sim \mathrm{Dir}(\epsilon_i)$ Aggregated distribution of document *i*

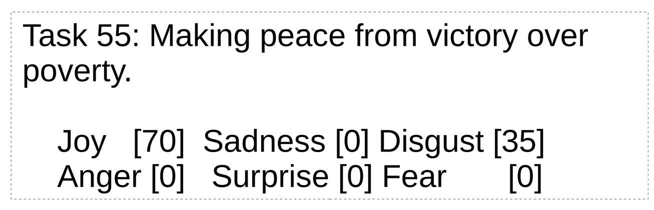
$$\Pi^{(k)} = \begin{pmatrix} \pi_{1,1}^{(k)} & \cdots & \pi_{1,J}^{(k)} \\ \vdots & \ddots & \vdots \\ \pi_{J,1}^{(k)} & \cdots & \pi_{J,J}^{(k)} \end{pmatrix} = \begin{pmatrix} \vdots \\ \pi_j^{(k)} \\ \vdots \end{pmatrix}$$

Confusion matrix of worker *k* with $\pi_j^{(k)} \sim \mathrm{Dir}\left(\alpha_j^{(k)}\right)$

- The joint distribution over all variables is given by
$$p(\mathbf{z}, \mathbf{c}, \Lambda, \Pi) = \prod_{i=1}^{I} \mathrm{Cat}(\Lambda_i)\,\mathrm{Dir}(\epsilon_i) \times \prod_{k=1}^{K} \prod_{n=1}^{N} \mathrm{Cat}\left(\pi_{z_{i,n}}^{(k)}\right) \prod_{j=1}^{J} \mathrm{Dir}\left(\alpha_j^{(k)}\right).$$
- Inference is performed by variational message passing on the posterior distribution $p(\mathbf{z}, \Lambda, \Pi|\mathbf{c})$.

Datasets

1. **SemEval**: 6 sentiments in 100 news headlines (1,000 judgments)
2. **IAPR-TC12**: 6 objects in 16 urban and rural scenes (441 judgments)
3. **Colours**: 10 colours in 20 countries' flag (460 judgments)



SemEval

IAPR-TC12

Colours

- Each dataset is augmented with additional synthetic spammers to explore the loss of accuracy as they increase in number.
- The distribution of each spammer *k* for each document *i* shares a prior Dirichlet distribution such that
$$\Phi_i^{(k)} \sim \mathrm{Dir}(\mathbf{1}).$$

Benchmarks

- **Uniform** distribution
- Arithmetic average distribution (**LinOp**) [4]
- **Median** distribution
- **IBCC** [3]

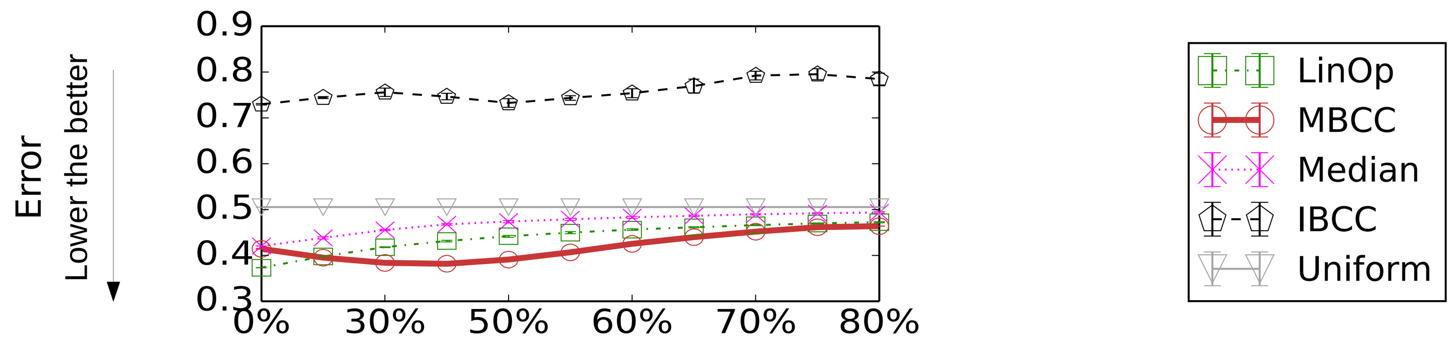
Error Metric

Average **Euclidean distance** between ground truth and the inference.

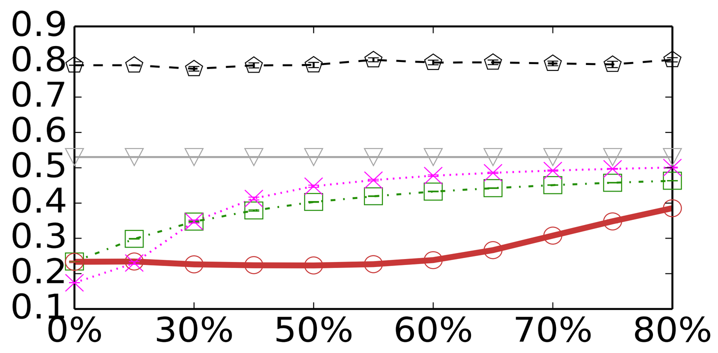
Results

The aggregation error is robust to an increasing ratio of spammers:

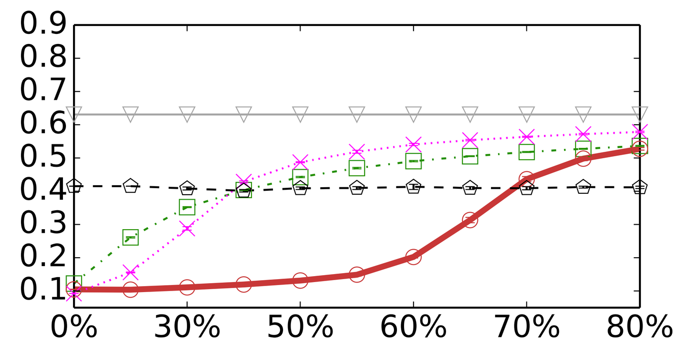
- up to 28% improvement in accuracy,
- comparable level of accuracy when 60% of the workers are spammers, as other approaches do when there are no spammers.



Colours dataset



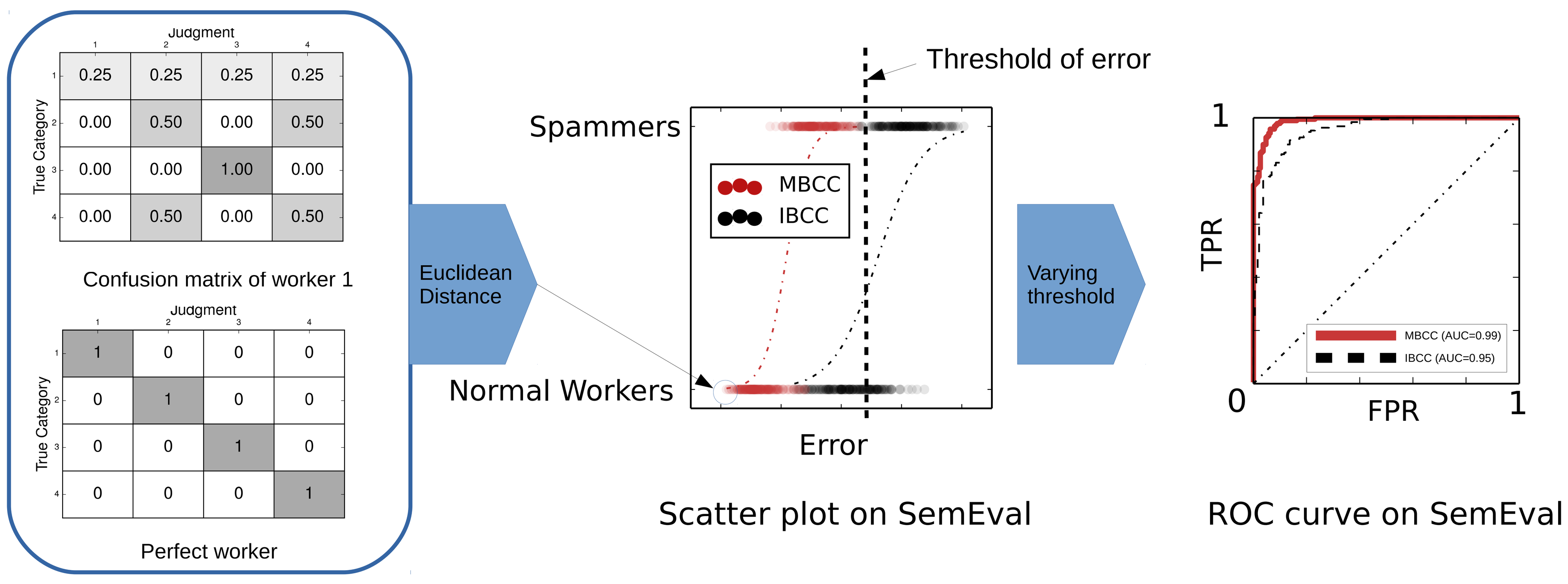
SemEval dataset



IAPR-TC12 dataset

The accuracy of classifying diligent workers from spammers is improved:

- up to fivefold improvement in expected number of misclassified spammers in comparison to existing methods.



Scatter plot on SemEval

ROC curve on SemEval

References

[1] Vuurens et al., "How Much Spam Can You Take? An Analysis of Crowd sourcing Results to Increase Accuracy", CIR, 2011.
[2] X. Geng, “Label distribution learning”. IEEE Transactions on Knowledge and Data Engineering, 2016.
[3] Kim et al., "Bayesian Classifier Combination", AISTATS, 2012.
[4] M. Bacharach, “Normal Bayesian Dialogues”, Journal of the American Statistical Association, 1979.

Author contact details
Email: aa7e14@soton.ac.uk

IJCAI 2017
This work was supported by General Electric Aviation.