2M5-OS-07b-3

Accurate Entity Resolution Using Crowdsourcing

Jingjing Wang*1 Satoshi Oyama*1 Masahito Kurihara*1 Hisashi Kashima*2

*1Hokkaido University *2The University of Tokyo

We propose a supervised learning method to derive an accurate model for entity resolution using crowd-generated data. Beginning with discussion on the accuracy of crowd-generated labels, we compared the prediction accuracy of our proposed method using crowd-generated labels with LPP method using consensus labels obtained by majority voting. Experiment results show that our proposed method outperformed LPP method for crowdsourced data with a low standard deviation.

1. Introduction

Recently increasing attention has been paid to crowdsourcing services, such as the Amazon Mechanical Turk (AMT), since they can provide a large amount of labeled data in a short period and at a low cost. However, it also bring challenges, coping with the variable quality of crowd-generated data. Among the crowd workers, some are highly skilled, and some are not. The highly skilled ones generally provide valid labels, while the lesser skilled ones generally provide variable-quality labels.

There are two main ways to process the label uncertainty problem: one is to estimate the ground truth labels [Dawid 79], and the other is to derive a classifier directly from crowd-generated data [Dekel 09, Raykar 10, Kajino 12]. One common approach to estimating consensus labels from individual worker labels is to use simple Majority Voting (MV), which can often achieve relatively good results depending on the accuracy of the workers involved. In MV, the label receiving the most votes is selected as the final aggregated label.

To make use of the information on the distribution of each worker's judgment, we introduce a combination of multiple Laplacians method to entity resolution using crowd-generated labels. In an experiment, we evaluated the label quality of actual crowdsourced data by comparing the labels donated by individual worker and the consensus labels with the ground truth labels. Then, we evaluated the prediction accuracy of the our proposed method using crowdsourced data.

2. Crowdsourced Labeling for Web Entity Resolution

We used data of four person names *("David Lodge,""Michael Howard,""Paul Clough,""Thomas Baker,")* obtained from the *Searching Information about Entities in the Web (WEPS)* dataset*1. We used words in Web pages as features and used binary features; that is, the value of the corresponding feature was set to one if a term appeared in the entry and to zero otherwise. Then, the web pages were assigned as *human intelligence tasks (HITs)* to crowd workers in *Lancers*2. Since one web page has been assigned to multiple HITs, a worker would label the same web page several times. In this case, a pair of web pages were assmumed linked once they were labeled as the same person.

Contact: Jingjing Wang, Hokkaido University, Kita 14 Nishi 9 Kita-ku Sapporo, 011-706-6815, 011-706-7831, wangjingjingi@yahoo.co.jp

∗1 http://nlp.uned.es/weps/weps-1/weps1-data

∗2 http://www.lancers.jp

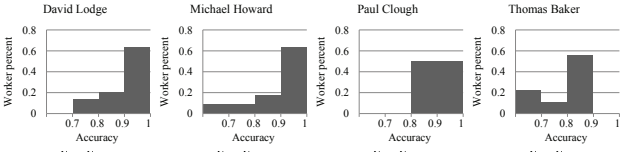

Figure 1: Accuracy of labels donated by individual worker.

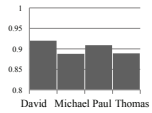

Figure 2: Average accuracy.

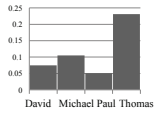

Figure 3: Standard deviation.

From Figure 1, 2, and 3, we can find out that the accuracy of labels obtained from individual workers was satisfactory while there was some dispersion across workers (the average accuracy was up to 0.9 and standard deviation was less 0.1 for most times).

3. Combination of Multiple Laplacians

The linear projection $\mathbf{W}$ from the original D-dimensional feature space to a d-dimensional latent feature space is learned from training data consisting of data objects known to have or not to have links between them. Assuming data objects $\mathbf{x}$ and $\mathbf{y}$ are known to have a link. The distance between them ($\|\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{y}\|_2^2$) in the latent space, should be as small as possible.

The link between two data objects with unknown link status is predicted on the basis of the distance between them after they are mapped to the latent space by using $\mathbf{W}$.

Assume a set of N training data objects $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in $\mathbb{R}^D$. The process for finding the optimal projection matrix $\mathbf{W}^*$ can be seen as:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{t}^{T} \sum_{i,j} r_t A_{ij}^{(t)} \|\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_j\|_2^2,$$

where $\|\cdot\|$ is the Euclidean norm, r_t is the weighting parameter. The adjacency matrix $\mathbf{A}^{(t)}$ is defined as $\mathbf{A}^{(t)} = \{A_{ij}^{(t)}\}$, which is donated by worker t. $A_{ij}^{(t)}$ represents the link status between the data objects in the training data set.

$$A_{ij}^{(t)} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ have a link,} \\ 0 & \text{otherwise.} \end{cases}$$

The problem can be formulated as an optimization problem:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathbf{tr}\left(\mathbf{W}\mathbf{\Phi}^\mathsf{T}\sum_t^T r_t \mathbf{L}^{(t)}\mathbf{\Phi}\mathbf{W}^\mathsf{T}\right)$$

where $\mathbf{\Phi}$ is the design matrix defined by $\mathbf{\Phi} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\mathsf{T}$, $\mathbf{L}^{(t)}$ is defined as the Laplacian matrix $\mathbf{L}^{(t)} = \mathbf{D}^{(t)} - \mathbf{A}^{(t)}$, $\mathbf{D}^{(t)}$ is a diagonal degree matrix in which the entries are column sums of $\mathbf{A}^{(t)}$, $D_{ii}^{(t)} = \sum_j A_{ij}^{(t)}$. Considering the variability of the crowd workers, higher weights are assigned to labels of more accurate workers.

And then the problem can be reduced as a generalized eigenvalue problem:

$$\mathbf{\Phi}\sum_t^T r_t \mathbf{L}^{(t)}\mathbf{\Phi}^\mathsf{T}\mathbf{w} = \lambda\mathbf{\Phi}\mathbf{\Phi}^\mathsf{T}\mathbf{w}.$$

The optimal projection matrix $\mathbf{W}^*$ is obtained by finding d eigenvectors with the smallest positive eigenvalues. After data objects mapped to the latent space by projection matrix, the links are identified based on the distance between them and the prediction accuracy can be calculated by comparing our predicted result with the truth labels.

4. Weight Measurements

We evaluated the prediction accuracy of our proposed method with or without weighting. Without weighting, the labels donated by the crowd worker were treated equally; that is, $r_t = 1/T$. We also used two weighting measurements, one was assigning higher weights to workers who finished more tasks which is $r_t = \frac{Finished\,task\,number}{Total\,task\,number}$, another one was depending on the similarity between the collected labels.

If a worker's labels were more similar to the labels donated by other crowd worker, we assumed that that crowd worker was more qualified and assigned a higher weight to that worker. The similarity between labels $\mathbf{A}_i$ donated by worker i and labels $\mathbf{A}_j$ donated by worker j is defined as

$$Sim_{ij} = \frac{(\mathbf{A}_i \cdot \mathbf{A}_j)}{||\mathbf{A}_i||}.$$

Thus, the weight of worker i is defined as the sum of similarities $r_i = \sum_{j\neq i, j\in\mathcal{T}} Sim_{ij}$, where $\mathcal{T}$ is the set of crowd workers, $\mathcal{T} = \{1, \ldots, T\}$. In practice, we add normalization to make the sum weight of all workers to 1.

5. Experiment

We experimentally evaluated the prediction accuracy of our proposed method without weighting (Uni), with weighting based on finished tasks (Task), with weighting based on similarity (Sim), and the conventional LPP method [He 04] for the crowdsourced labels. Majority voting was used to generate consensus labels. Because the number of votes affects the majority voting result to some extent, while the original crowd-generated data (All_worker) was obtained by 5 times voting, we randomly chose labels to generate sampled crowd-generated data $(4\_worker, 3\_worker, 2\_worker, 2\_worker)$.

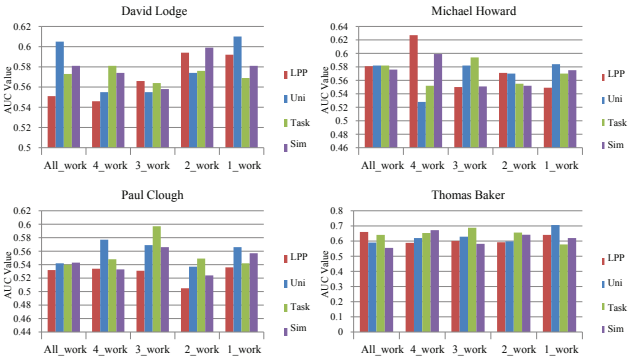

Figure 4: Prediction accuracy.

By comparing accuracy distribution in Figure 3 with prediction accuracy of our proposed method and LPP method in Figure 4, we can see except for *"Thomas Baker"*, who has the highest standard deviation, our proposed method with or without weighting has the same or better performance than the LPP method on original crowdsourced data. Furthermore, when the result of LPP methed outperformed ours, the gap can be reduced by using measure of similarity, for example, sampled crowdsourced data (2_worker) of *"Thomas Baker"* and sampled crowdsourced data (4_worker) of *"Michael Howard"*. In addition, our proposed method without weighting outperformed the LPP method on sampled crowdsourced data (1_worker).

6. Conclusion

In this paper, we evaluated the label quality of an actual crowdsourced data for entity resolution and compared the prediction accuracy of our combination of multiple Laplacians method to the conventional locality preserving projections method. Our analysis on label quality showed that the accuracy of labels donated by individual workers was satisfactory while there was some dispersion across workers. The evaluation on the prediction accuracy indicates that the distribution of the accuracy of workers affects the accuracy of learned prediction models: the high standard deviation of worker label reduces the prediction accuracy. Thus, our future work will focus on improving the measure used for assigning weights to workers in the learning process.

References

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, *Journal of the Royal Statistical Society. Series C* (1979)

[Dekel 09] Dekel, O. and Shamir, O.: Vox Populi: Collecting High-Quality Labels from a Crowd, in *COLT* (2009)

[He 04] He, X. and Niyogi, P.: Locality Preserving Projections, in *NIPS 16*, pp. 153–160 (2004)

[Kajino 12] Kajino, H., Tsuboi, Y., and Kashima, H.: A Convex Formulation for Learning from Crowds, in *AAAI* (2012)

[Raykar 10] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L.: Learning From Crowds, *JMLR* (2010)