1S-1

Learning an Accurate Entity Resolution Model from Crowdsourced Labels

Jingjing Wang[†] **Satoshi Oyama**[†] **Masahito Kurihara**[†] **Hisashi Kashima**[‡]
Hokkaido University[†] The University of Tokyo[‡]

Abstract

We propose a supervised machine learning method for entity resolution using labels collected by crowd workers. Although obtaining a large amount of labeled data via crowdsourcing services can reduce time and monetary cost, it also brings challenges, coping with the variable quality of crowd-generated data and the lack of absolute gold standard. We adopt a dimension reduction approach to project data objects to a low-dimensional latent space, and the data objects is identified by the basis of the distance between data objects. To get an accurate entity resolution model from the inaccurate labels gathered by crowd workers, we introduce regularization to simultaneously learn the true projection and workers projection matrices; that is, to make the true model close to each worker model.

Entity resolution (sometimes referred to as duplicate detection, entity reconciliation, or record linkage) is a task of finding whether different data objects refer to the same real world entity. Entity resolution plays an important role in data integration and data cleaning, especailly under the condition that same name refers to different entities. Entity resolution can also be considered as a link prediction problem by regarding the entity identity as a link. Since in a link prediction problem, data objects and the relationship among them are considered as nodes and edges in a graph, the identify of whether two data objects refer to the same real world entity can be seen as finding whether there is a link between two nodes.

In recent years, many efforts have been done to make entity resolution automatically from labeled training data using machine learning techniques (Bilenko and Mooney 2003; Sarawagi and Bhamidipaty 2002). They all estimates classifier directly from ground true labels. The entity resolution quality is improved, but still far from perfect. The goal of supervised machine learning is not only to get improved performance from training data, but also to obtain predictive models for future data. Especially for some supervised learning tasks it may be infeasible (or very expensive) to obtain objective and reliable labels given by domain experts. Thus, it is important to consider effective ways to gather a large amount of labeles for training data.

To solve this problem, increasing attention has been paid to crowdsourcing. Crowdsourcing services, such as the Amazon Mechanical Turk (AMT), can collect a large amount of labeled data in short period and at low cost. Among the crowd workers, some are highly skilled, some are not. For the highly skilled workers, they can always provide valid labels, but for the unkilled workers, they generally provide labels randomly. Therefore, we have to face the quality control problem of crowd workers.

In this paper, we propose a supervised machine learning method for entity resolution using crowd-generated data. In the learning process, we adopt a dimension reduction approach to project high-dimensional data objects to a low-dimensional latent feature space, and the data objects is identified on the basis of the distance between data objects. That is, the closer two data objects are to each other in the latent space, the greater the likelihood of identified between them. To get an accurate entity resolution model from the inaccurate labels gathered by crowd workers, we introduce regularization to simultaneously learn the true projection matrix and workers projection matrices from the identification labels given by crowd workers; which is, to make the true model close to each worker model. We also add weight to the regularization process. Depending on the work quality of individual worker, it assigns greater weight to more accurate workers.

Supervised Machine Learning from Crowd-generated Labels

Since crowd-generated data is without golden standard, our goal is to learn an accurate true model from the inaccurate crowd-generated labels. We use a different feature projection $\mathbf{W}^{(t)}$ for each crowd worker t, and the corresponding adjacency matrix donated by the worker is $\mathbf{A}^{(t)}$.

Assume that two data objects, $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(u)}$ are known to have a link, the distance between the two data objects in the latent space, should be as small as possible.

$$||\mathbf{W}^{(t)}\mathbf{x}^{(t)} - \mathbf{W}^{(t)}\mathbf{x}^{(u)}||_2^2$$

In the link-prediction process, link prediction is made on the basis of the distances in the latent space.

We use $\mathbf{W}^{(0)}$ for true feature projection. After concatenation of the true model and the workers' models, the parameter matrix to be learned is as $\widetilde{\mathbf{W}} \equiv \left[\mathbf{W}^{(0)}, \mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(T)}\right]$, where T is the number of crowd

workers. The enlarged design matrix obtained by diagonal arrangement of $T+1$ single design matrices is an $N \times (T+1)D$ matrix:

$$\widetilde{\mathbf{\Phi}} \equiv \begin{bmatrix} \mathbf{\Phi} & & & \\ & \mathbf{\Phi} & & \\ & & \ddots & \\ & & & \mathbf{\Phi} \end{bmatrix}$$

where $\mathbf{\Phi}$ is the design matrix defined by $\mathbf{\Phi} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\mathsf{T}$, N is the number of trainning data objects, D is the original dimension number of feature vectors.

The adjacency matrices gathered by crowd workers is a $(T+1)N \times (T+1)N$ matrix:

$$\widetilde{\mathbf{A}} \equiv \begin{bmatrix} \mathbf{0} & & & \\ & \mathbf{A}^{(1)} & & \\ & & \ddots & \\ & & & \mathbf{A}^{(T)} \end{bmatrix}$$

The true projection matrix and workers' projection matrices are independant to each other. To balance the variance of individual crowd workers, we impose an additional requirement for the projection matrix: the true projection matrix should close to every worker's projection matrix. The regularization is defined as:

$$\sum_{t=1}^{T} r_t \|\mathbf{W}^{(t)} - \mathbf{W}^{(0)}\|_2^2$$

where r_t is the weight of each worker. We introduce matrix Λ with size $(T+1)D \times (T+1)D$ for regularization:

$$\Lambda \equiv \begin{bmatrix} (\sum_{t=1}^{T} r_t)\mathbf{I} & -r_1\mathbf{I} & -r_2\mathbf{I} & \cdots & -r_T\mathbf{I} \\ -r_1\mathbf{I} & r_1\mathbf{I} & & & \\ -r_2\mathbf{I} & & r_2\mathbf{I} & & \\ \vdots & & & \ddots & \\ -r_T\mathbf{I} & & & & r_T\mathbf{I} \end{bmatrix}$$

where $\mathbf{I}$ is the $D \times D$ identity matrix. That the follow equation holds is easily shown.

$$\widetilde{\mathbf{W}}\widetilde{\mathbf{\Phi}}^\mathsf{T}\Lambda\widetilde{\mathbf{\Phi}}\widetilde{\mathbf{W}}^\mathsf{T} = \sum_{t=1}^{T} r_t \|\mathbf{W}^{(t)} - \mathbf{W}^{(0)}\|_F^2$$

Using this relationship, we formulate the optimization problem as:

$$\widetilde{\mathbf{W}}^* = \arg\min_{\widetilde{\mathbf{W}}} \mathbf{tr}\left(\widetilde{\mathbf{W}}\left(\widetilde{\mathbf{\Phi}}^\mathsf{T}\widetilde{\mathbf{L}}\widetilde{\mathbf{\Phi}} + \sigma\Lambda\right)\widetilde{\mathbf{W}}^\mathsf{T}\right)$$
$$\text{s. t. } \widetilde{\mathbf{W}}\widetilde{\mathbf{\Phi}}^\mathsf{T}\widetilde{\mathbf{D}}\widetilde{\mathbf{\Phi}}\widetilde{\mathbf{W}}^\mathsf{T} = \mathbf{I}_d$$

where σ is a constant specifying the strength of regularization, where $\widetilde{\mathbf{D}}$ is the diagonal degree matrix in which each element $\widetilde{\mathbf{D}}_{ii} = \sum_j \widetilde{\mathbf{A}}_{ij}$ represents the number of links data object i has, and $\widetilde{\mathbf{L}}$ is the Laplacian matrix defined by $\widetilde{\mathbf{L}} = \widetilde{\mathbf{D}} - \widetilde{\mathbf{A}}$. The optimization problem can be reduced as a generalized eigenvalue problem:

$$\left(\widetilde{\mathbf{\Phi}}\widetilde{\mathbf{L}}\widetilde{\mathbf{\Phi}}^\mathsf{T} + \sigma\Lambda\right)\mathbf{w} = \lambda\widetilde{\mathbf{\Phi}}\widetilde{\mathbf{D}}\widetilde{\mathbf{\Phi}}^\mathsf{T}\mathbf{w}.$$

By finding the smallest eigenvectors of the problem above, we can simultaneously obtain the true model and all worker's models.

AUC Value

	All_woker	3_woker	2_woker
MW-LPP(d=10, σ =50)	0.6022	0.5204	0.5439
MW-LPP(d=10, σ =10)	0.5070	0.5370	0.6132
LPP	0.5512	0.5664	0.5936

Figure 1: AUC comparison on the real crowdsourcing data. For *All_worker*, each data object is voted 5 times, for 3_*worker* each data object is voted 3 times, and for 2_*worker* each data object is voted 2 times.

Experiment

We evaluate our proposed method against the Locality Preserving Projections (LPP) method (He and Niyogi 2004) on a real crowdsourced data set. We use simple MV method to compute consensus labels by equally weighting each worker's vote for LPP method. We call this *Multiple Workers' Locality Preserving Projections (MW-LPP)*. As we can see in Figure 1, our method can achieve comparable or better accuracy when we set appropriate reglarization strength, especially when the number of votes is 5. The regularization constant can be adapted by cross-validation.

Conlusion

In this paper, we propose a new approach to deal with Entity Resolution by directly learning from inaccurate labels donated by crowd workers. In our future woker, we will study better methods to assign weight to workers.

References

Bilenko, M., and Mooney, R. J. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 39–48.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *Advances in Neural Information Processing Systems 16 (NIPS)*, 153–160.

Sarawagi, S., and Bhamidipaty, A. 2002. Interactive deduplication using active learning. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data mining (KDD)*, 269–278.