A Semi-supervised Approach to Transferring the Learned Knowledge for Indoor Location Estimation

Shoko Suzuki, Yuta Tsuboi, Hisashi Kashima, Shohei Hido,
Toshihiro Takahashi, Tsuyoshi Idé, Rikiya Takahashi, and Akira Tajima
IBM Research
Tokyo Research Laboratory
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502, Japan

Abstract

We describe our approach that we used for Task2 "Transferring the Learned Knowledge for Indoor Location Estimation" in ICDM Data Mining Contest 2007. We formulated the task as a transduction problem under a distribution change, and employed a semi-supervised learning approach based on the Laplacian eigenmap followed by the nearest neighbour classifier.

1 A formulation as a transduction problem under a distribution change

We formulate the indoor location estimation problem as a transductive multi-class classification problem.

Let the whole data set consists of $N = 5,503$ instances, whose i-th data is given as $(\mathbf{x}^{(i)}, y^{(i)})$, where $\mathbf{x}^{(i)} \in \mathbb{R}^{101}$ is the vector of the received signal strength (RSS) values from the WiFi Access Points (APs), and $y^{(i)} \in \{1, 2, \ldots, 247\}$ is the location label assigned to the RSS vector. As for the unobserved RSS values, we filled them by -100, since all RSS values are in the range of $[-100, 0]$ and unobserved RSS value implies that it was too weak to detect.

The datasets are divided into two types of domains the *source domain* and the *target domain*, whose data distributions are considered to be similar but not the same. The source domain we are given a relatively large number (621) of labeled data denoted by D_{SL} , and $1,701$ unlabeled data D_{SU}. On the other hand, in the target domain, we have only 53 labeled data denoted by D_{TL}, and $3,128$ unlabeled data denoted by D_{TU}.

The task is to predict the location labels for D_{TU}, the unlabeled data in the target domain. Note that since the inputs of the test data set are given in advance of the test phase, we can regard the problem as a transduction problem where test inputs are explicitly used.

2 A semi-supervised approach using the Laplacian eigenmap

The task can be basically considered as a multi-class classification problem. Since the data distribution in the source domain and that in the target domain are different, an obvious solution is to train a classifier using only D_{TL} (and possibly D_{TU} in addition). However, the number of the labeled data in the target domain is even smaller than the number of location labels. We have to "transfer" information from the source domain.

Therefore, our strategy is to use one of the supervised learning approaches [2] using all the data except D_{TU}, the unlabeled data in the source domain. Our intention behind this is as follows. We use all of the labeled data since the number of the target labeled data is far from enough. As for the unlabeled data, we use only the target unlabeled data, since we would like to work in the intrinsic feature space in the target domain and to avoid suffering from the source data distribution.

Our approach consists of two steps. First we apply a nonlinear dimension reduction technique called the *Laplacian eigenmap* [1] to obtain an appropriate feature space for the target domain, and then use the nearest neighbour classifier to predict the location labels for the target unlabeled data.

The Laplacian Eigenmap is a nonlinear unsupervised dimension reduction technique. It treats each data instance as a node in a weighted graph, whose weighted edge for node i and node j (corresponding to the i-th instance and the j-th instance, respectively) is defined as a heat-kernel like function as follows,

$$w^{(i,j)} = \exp\left(\frac{-\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|_2^2}{(c\sigma)^2}\right), \tag{1}$$

where $\|\cdot\|_2$ is the 2-norm, and $\sigma \approx 4.23$ is a constant defined as the mean of the standard deviations estimated with respect to each of the axes, and c is a parameter to be tuned.

The Laplacian eigenmap considers the Laplacian matrix L represents an intrinsic structure of the data distribution, which is defined as

$$L = D - W, \tag{2}$$

where W is a matrix each of whose element is given by $w^{(i,j)}$, and D is a diagonal matrix each of whose diagonal element $d^{(i,i)}$ is defined as

$$d^{(i,i)} = \sum_{\ell} w^{(i,\ell)}. \tag{3}$$

To find the new coordinates of the data embedded into the new feature space induced by Laplacian matrix, we solve the following generalized eigenvalue problem,

$$L\mathbf{y} = \lambda D\mathbf{y}, \tag{4}$$

Let $\mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_k$ be the $k+1$ smallest eigenvectors and $\lambda_0, \lambda_1, \cdots, \lambda_k$ be the $k+1$ smallest eigenvalues of L, respectively. Since $\lambda_0 = 0$ always holds for Laplacian matrices, k vectors $\mathbf{y}_1, \cdots \mathbf{y}_k$ can be regarded as the k dimensional coordinates, where original data are embedded.

In addition, since the obtained coordinates are normalized by the constraint

$$\mathbf{y}^T D\mathbf{y} = 1, \tag{5}$$

we further rescale them to obtain the new coordinates $\mathbf{z}_1, \cdots, \mathbf{z}_k$ as

$$\mathbf{z}_i = \frac{1}{\lambda_i}\mathbf{y}_i, \tag{6}$$

so that the new coordinates reflect the scale of the original data distribution. This feature is not implemented in the standard Laplacian eigenmap, but we found it effective in our preliminary experiments.

Once we obtained the new coordinates of the data, we can apply standard supervised classification algorithms by using D_{SL} and D_{TL} as training data, and obtain the predictions for the unlabeled target data D_{TU}. In our submission, we employed the nearest neighbour classifier, since we found that the nearest neighbour classifier is quite robust to distribution changes through our preliminary experiments.

3 The algorithm

Based on the discussion in the previous section, each step of the algorithm is summarized as follows. In our submission, the parameter c was set to 25, based on the result of a 10-fold cross validation using the target labeled data. Similarly, k was set to 20.

1. Prepare a data set with the labeled source data D_{SL}, the labeled and unlabeled target data (D_{TL}, D_{TU}), where each instance is a 101-dimensiona RSS vector. All the missing values are filled with the value -100.

2. Calculate the heat kernel for all instance pairs by (1) with the parameter c to be tuned by cross validation, and the diagonal matrix D defined by (3), to obtain the Laplacian matrix L by (2).

3. Find the $k+1$ smallest eigenvectors $\mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_k$ and the $k+1$ smallest eigenvalues $\lambda_0, \lambda_1, \cdots, \lambda_k$ of the generalized eigenvalue problem (4). Remove $\mathbf{y}_0$ with the smallest eigenvalue, and calculate the new coordinates $\mathbf{z}$ by (6), where k is the parameter to be tuned by cross validation.

4. Apply the nearest neighbour classifier to the unlabeled target data to predict their labels. The nearest neighbour to each instance is calculated by the 2-norm in the new feature space $\mathbf{z}_1, \cdots, \mathbf{z}_k$.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 6(15):1373–1396, 2003.

[2] X. Zhu. Semi-supervised learning literature survey. Technical Report Computer Sciences, TR 1530, University of Wisconsin Madison, 2006.