SPARSE REPRESENTATION FOR OUTLIERS SUPPRESSION IN SEMI-SUPERVISED IMAGE ANNOTATION

Toru Nakashika, Takeshi Okumura, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
{nakashika,okumura}@me.cs.scitec.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

ABSTRACT

Recently, generic object recognition (automatic image annotation) that achieves human-like vision using a computer has being looked to for use in robot vision, automatic categorization of images, and retrieval of images. For the annotation, semi-supervised learning, which incorporates a large amount of unsupervised training data (unlabeled data) along with a small amount of supervised data (labeled data), is expected to be an effective tool as it reduces the burden of manual annotation. However, some unlabeled data in semi-supervised models contains outliers that negatively affect the parameter estimation on the training stage. Such outliers often cause the over-fitting problem especially when a small amount of training data is used. In this paper, we propose a practical method to prevent the over-fitting in semi-supervised learning, suppressing existing outliers by sparse representation. In our experiments we got 4 points improvement comparing conventional semi-supervised methods, SemiNB and TSVM.

Index Terms— Object recognition, automatic anotation, sparse representation, semi-supervised learning

1. INTRODUCTION

Automatic image annotation, in which the system automatically assigns labels to an image, is one of the most significant tasks in computer vision. Most of the conventional methods are based on a supervised labeling approach in order to achieve an exact classification. However, it has been pointed out that with this approach the training cost is extremely high because an enormous amount of training data must be labeled manually. To reduce the amount of such a troublesome work, a semi-supervised approach has recently attracted considerable attention in machine learning [1][2][3][4][5]. The semi-supervised approach inputs a large amount of non-labeled data (unsupervised data) for the training as well as not so much labeled data. Hence, it helps to improve the training accuracy without using a lot of labeled data.

Descriptions of popular methods using semi-supervised learning in text classification can be found in [2], which introduces TSVM (transductive support vector machine) as a classification model, and in [3], which introduces SemiNB (semi-supervised naive Bayes classifier) as a generative model. TSVM extends the well-known SVM so that it can be trained not only with a few labeled data but also with a large volume of unlabeled data. During the training, labeled data first determine the margin, which classifies unlabeled data. The former SemiNB is a semi-supervised version of Naive Bayes (NB).

Both methods, especially in SemiNB, are adversely affected by the influence of outliers in large amounts of unsupervised data, because they both take whole the unlabeled data as well as labeled data in the training process. The TSVM limits the influence of the outliers only to data around the margin. Therefore, the TSVM is not influenced as much by outliers as SemiNB is, though it is inevitable that the outliers negatively affect the margin estimation. Furthermore, the TSVM is a computationally expensive algorithm. Given a large number of training data, it needs to take an approximate approach that causes weak estimation of the margin.

In consideration of the drawbacks in a semi-supervised approach, we propose an automatic image annotation method where an effective semi-supervised tool, semi-supervised canonical correlation analysis (semi-CCA) [4], and sparse representation [6] collaboratively suppress the influence of outliers. Fig. 1 shows the flow of the proposed method. First, subspaces that maximize the correlation between image features and label features are generated by semi-CCA, using a small amount of labeled data and much unlabeled data. Semi-CCA extends canonical correlation analysis (CCA), so as to avoid over-fitting when it has a few (labeled) training data. Given a large amount of unlabeled data as well as the labeled data, it grabs a global distribution. Since the trained distribution is affected by outliers somewhat, we adopt Regularized Orthogonal Matching Pursuit (ROMP) [7], one of the handy sparsing algorithms. Using sparse representation, it is possible to achieve the automatic annotation that utilizes an abundance of unlabeled data for the semi-supervised learning that is robust to the influence of outliers.

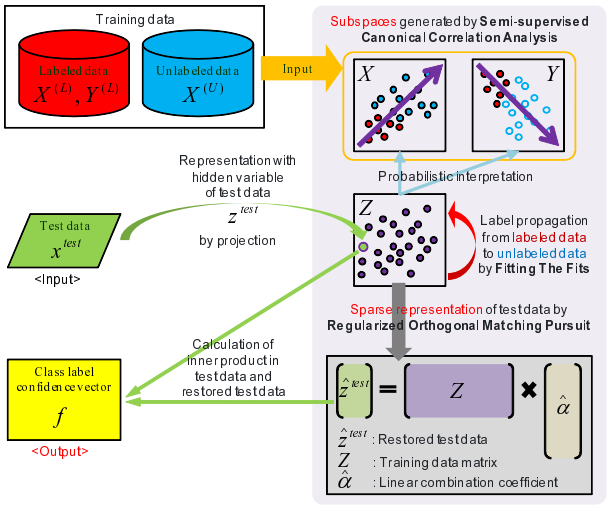

Fig. 1. System flowchart of proposed method.

2. SUBSPACE GENERATION

Due to the high cost of preparing correct labels as training data in automatic image annotation, it is desirable to employ a semi-supervised approach which uses unlabeled data instead of some of labeled data. In this section, we discuss a subspace generation method using the semi-supervised approach called Semi-supervised Canonical Correlation Analysis (semi-CCA) [4]. The semi-CCA is an extended version of Canonical Correlation Analysis (CCA) so that it substitutes unlabeled data for some labeled data. Both methods find the subspace that maximizes a correlation between two different types of features. In this paper, the relationship (correlation) between an image and the accompanying labels are obtained.

2.1. Semi-CCA

Let $\{\mathbf{X}^{(L)}, \mathbf{Y}^{(L)}, \mathbf{X}^{(U)}\}$ be a training data set, where $\mathbf{X}^{(L)} = \{\mathbf{x}_n\}_{n=1}^{N}$ and $\mathbf{Y}^{(L)} = \{\mathbf{y}_n\}_{n=1}^{N}$ are N labeled data, and $\mathbf{X}^{(U)} = \{\mathbf{x}_m\}_{m=1}^{M}$ is M unlabeled data. $\mathbf{x}$ and $\mathbf{y}$ indicate image feature and label feature, respectively (See 2.2). The aim of Semi-CCA or CCA is to find the optimum subspace that maximizes a correlation between projected $\mathbf{x}$ and $\mathbf{y}$:

$$r(\mathbf{w}_x, \mathbf{w}_y) = \frac{\mathbf{w}_x^T \mathbf{S}_{xy}^{(L)} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{S}_{xx}^{(L)} \mathbf{w}_x} \sqrt{\mathbf{w}_y^T \mathbf{S}_{yy}^{(L)} \mathbf{w}_y}} \quad (1)$$

where $\mathbf{w}_x$ and $\mathbf{w}_y$ are projection vectors to the subspace from the original feature space $\mathbf{x}$ and $\mathbf{y}$, respectively. $\mathbf{S}_{**}^{(L)}$ indicates each variance-covariance matrix within the labeled data. For example, $\mathbf{S}_{xy}^{(L)} = N^{-1} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{y}_n^T$.

If unlabeled data $\mathbf{X}^{(U)}$ is not given as the training (in other words, in the case of normal CCA), all that can be done is just to formulate the maximum problem in which the optimum $\mathbf{w}_x$ and $\mathbf{w}_y$ are found to maximize Eq. (1) using a Lagrange multiplier. In that case, the formulation boils down to an eigenvalue problem.

When the amount of labeled data is not adequate, the obtained subspace is inefficiently overfitted to the training data. Hence, unlabeled data are added for correcting a global structure of data distribution in the subspace. In order to do that, PCA is employed using the concept of semi-CCA [4]. In a similar way to CCA, a projection matrix of the PCA can be calculated by solving an eigenvalue problem, in which a variance-covariance matrix of the data is maximized under a normalized orthogonal constraint.

As mentioned above, semi-CCA can be expressed as the combination of two factors: CCA with labeled data and PCA with all data including unlabeled data. Therefore, the semi-CCA formulation is also obtained by combination of the two eigenvalue problems, as in Eq. (2). A projection matrix can ultimately be obtained from the upper eigenvalues using semi-CCA.

$$\mathbf{B} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix} = \lambda \mathbf{C} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix} \quad (2)$$

where,

$$\mathbf{B} = \beta \begin{bmatrix} \mathbf{0} & \mathbf{S}_{xy}^{(L)} \\ \mathbf{S}_{yx}^{(L)} & \mathbf{0} \end{bmatrix} + (1-\beta) \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy}^{(L)} \end{bmatrix} \quad (3)$$

$$\mathbf{C} = \beta \begin{bmatrix} \mathbf{S}_{xx}^{(L)} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy}^{(L)} \end{bmatrix} + (1-\beta) \begin{bmatrix} \mathbf{I}_{D_x} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{D_y} \end{bmatrix} \quad (4)$$

and $\mathbf{S}_{xx} = (N+M)^{-1} \sum_{n=1}^{N+M} \mathbf{x}\mathbf{x}^T$ is a variance-covariance matrix of all image feature vectors including unlabeled images. $\mathbf{I}_{D_x}$ and $\mathbf{I}_{D_y}$ are identity matrices with the size $D_x \times D_x$ and $D_y \times D_y$, respectively. Note that the first term and the second term in Eq. (3) and (4) indicate the terms related to eigenvalue problems of CCA and PCA, respectively. β is a trade-off parameter which determines the effects of CCA and PCA.

The image feature and the label feature are connected via latent variables $\mathbf{z}$ in the subspace. These variables can be calculated by applying the conditional Gaussian model (For more details, see [4]). After this, we can rewrite the training and test data with the latent variables $\mathbf{z}$ for the sake of consideration in the subspace.

2.2. Image Feature and Label Feature

Each image is first divided into small subregions using Normalized Cuts [8]. In each subregion, the following features are extracted:

- Color : Statistics of RGB, HSV, Lab, and YCbCr
- Gabor : Gabor filter and Laplacian-of-Gaussian
- Position : Center position of a region
- Geometric : Area of a region

An image feature vector x is defined as a supervector, where all these features are included. A label feature vector y is a binary vector, where each label is assigned in the subregion or not.

3. ANNOTATION

Recently, classification methods based on sparse representation, in which test data is represented as a linear combination of sparse bases, have been drawing attention in image processing [9][10]. It was reported in these papers that classification results showed favorable robustness of sparse representation against outliers. In this paper, we set out to suppress the effect of outliers that unintentionally appearwhen there is a large amount of unlabeled data, by employing sparse representation.

If a sufficient amount of training data is prepared, an input image $\mathbf{z}^{\text{test}}$ in the subspace can be represented as a linear combination of the training data. Our aim is to find *sparse* coefficients associated with each training data. Those entries are mostly zero, except for a few elements. This can be formulated as a minimizing problem with respect to a coefficient vector $\boldsymbol{\alpha}$ in Eq. (5).

$$\min_{\boldsymbol{\alpha}} ||\boldsymbol{\alpha}||_\epsilon \quad s.t. \quad \mathbf{z}^{\text{test}} = \sum_{n=1}^{N+M} \alpha_n \mathbf{z}_n = \mathbf{Z}\boldsymbol{\alpha} \tag{5}$$

where $\mathbf{Z} \in \mathbb{R}^{D_z \times (N+M)}$ is a training data matrix (D_z is a dimension of subspace feature). $||\boldsymbol{\alpha}||_\epsilon$ indicates l_ϵ norm, which is the number of almost-zero elements in $\boldsymbol{\alpha}$, given by $||\boldsymbol{\alpha}||_\epsilon = (N+M)^{-1} \# \{n | \alpha_n \leq \epsilon\}$ with an experimentally-determined small value ϵ. However, it is computationally difficult to find the optimum vector in Eq. (5) because $||\boldsymbol{\alpha}||_\epsilon$ is indifferentiable. In this paper, we consequently adopt one of the popular greedy algorithms, Regularized Orthogonal Matching Pursuit (ROMP) [7] to solve this optimum problem.

At the end, the test data can be restored by multiplying a training data and the obtained vector $\hat{\boldsymbol{\alpha}}$ as $\hat{\mathbf{z}}^{test} = \mathbf{Z}\hat{\boldsymbol{\alpha}}$. By taking an inner production between the test data and the restored data, a restoration ratio f_c of the label class c can be calculated as

$$f_c = \frac{\mathbf{z}^{test\,T} \hat{\mathbf{z}}_c^{test}}{||\mathbf{z}^{test}||_2 ||\hat{\mathbf{z}}^{test}||_2} = \frac{\mathbf{z}^{test\,T} \mathbf{Z}_c \hat{\boldsymbol{\alpha}}_c}{||\mathbf{z}^{test}||_2 ||\hat{\mathbf{z}}^{test}||_2} \tag{6}$$

where $\mathbf{Z}_c$ is a training data matrix that only contains the data given the label c, and $\hat{\boldsymbol{\alpha}}_c$ is a coefficient vector associated with the training data in $\mathbf{Z}_c$. The restoration ratio f_c implies a confidence of the class c. Therefore, multi-label classification can be realized by calculating all the confidences $f_c (c = 1, \cdots, C)$ (C is the number of label classes).

4. EXPERIMENTS

4.1. Experimental conditions

For image annotation experiments, we used a STAIR image data set [11], which contains 534 images along with pixel-wise 5 labels ("Sky", "Tree", "Road", "Grass" and "Building.") In our experiments, the training and test are conducted using features in each subregion, which is divided by Normalized Cuts. Labeling accuracies for each class and their average are calculated by accumulating the subregion results. The accuracy was evaluated with 3-fold cross validation. Images for the training and the test were randomly selected (400 images for the training and 134 images for the test) three times for each validation.

We conducted two experiments to evaluate our proposed method. In the first experiment, we compared with conventional semi-supervised methods: "SemiNB" and "TSVM". Secondly, we examined these methods in a supervised manner; a supervised variation of our method was compared with "NB" and "SVM," just to see the effectiveness of semi-supervised approach. Here we employed CCA instead of semi-CCA, given a full set of labeled data.

4.2. Results and Discussion

The results of semi-supervised and supervised approaches are shown in Table 1 and Table 2, respectively. Fig. 2 summarizes the results. As shown these tables and figure, the labeling accuracy of our method is higher than not only the other semi-supervised approaches but also the supervised approaches, such as SVM. The other methods, SVM and NB, suffer decreased accuracy in the semi-supervised case. This is, in general, because conventional approaches make extensive use of unsupervised data, and their classifiers were consequently affected by unsupervised factors, especially outliers. On the other hand, our approach increases the accuracy in the semi-supervised case. This is considered to be due to the benefit of semi-supervised learning, which helps the classifier to catch the global structure of data distribution in the case where there is a very small amount of labeled data for the training (due to the effective suppression of outliers using sparse representation).

Comparing the results of "Proposed(supervised)" and "ROMP" in Fig. 1, we can see that our method's accuracy is much higher than the other. The only difference in their methods is that the former approach projects image and label features using CCA, while the latter does not. In sparse representaion such as ROMP, it is known that these methods better perform when the training data is Gaussian or Bernoulli distributed. As previously described, the projected image or label feature are modeled by Gaussian distribution in semi-CCA. This means that ROMP is compatible with CCA (or semi-CCA), and thus, we believe, our proposed method best

Table 1. Labeling accuracies for comparison of proposed method and conventional methods (%).

Label	Sky	Tree	Road	Grass	Build.	**Ave.**
SemiNB	25.3	30.5	70.6	47.5	31.3	41.0
TSVM	82.8	62.0	83.3	74.4	68.6	74.2
Proposed	87.2	66.0	84.4	80.1	75.9	**78.7**

Table 2. Labeling results within supervised approaches (%).

Label	Sky	Tree	Road	Grass	Build.	**Ave.**
NB	43.2	26.7	73.8	54.3	33.1	46.2
SVM	87.8	57.1	85.5	76.9	65.2	74.5
ROMP	54.5	39.2	51.5	44.5	39.9	45.9
Prop.(supervised)	85.0	63.7	89.6	75.4	73.2	77.4

performed due to their synergistic effect, in addition to mere outlier-suppression.

5. CONCLUSION

In this paper, we proposed an effective image annotation method, which suitably combines a semi-supervised approach, Semi-supervised Canonical Correlation Analysis (semi-CCA), and sparse representation, Regularized Orthogonal Matching Pursuit (ROMP). Semi-supervised learning has the advantage of being able to capture a global structure of the true data distribution even when given only a small amount of labeled training data. However, outliers included in unsupervised data often give the negative effect to the classifier construction. Our approach suppresses such outliers in terms of sparse representation in the subspace which is created using semi-CCA.

The experimental results showed the effectiveness of our proposed method satisfactorily. While conventional semi-supervised approaches decreased the labeling accuracy compared with their supervised versions, our sparse-representation-based approach, on the contrary, increased the accuracy, taking advantage of semi-supervised learning fully.

Acknowledgment

This research was supported in part by MIC SCOPE.

6. REFERENCES

[1] X. Zhu, "Semi-supervised learning literature survey," *Proc. IEEE International Conference on Machine Learning (ICML), tutorial*, 2007.

[2] T. Joachims, "Transductive inference for text classification using support vector machines," *Proc. IEEE International Conference on Machine Learning (ICML)*, pp. 200–209, 1999.

[3] K. Nigam, A. Mccallum, and T. Mitchell, "Semi-supervised text classification using EM," *In Semi-supervised Learning*, pp. 33–56, 2006.

[4] A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro, "Semicca: Efficient semi-supervised learning of canonical correlations," *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, pp. 2933–2936, 2010.

[5] M. Culp and G. Michailidis, "An iterative algorithm for extending learners to a semi-supervised setting," *Journal of Computational and Graphical Statistics*, pp. 545–571, 2008.

[6] M. Elad, M. A. T. Figueiredo, and M. Yi, "On the role of sparse and redundant representations in image processing," *Proc. IEEE Special Issue on Applications of Sparse Representation and Compressive Sensing*, pp. 972–982, 2010.

[7] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of Computational Mathematic*, pp. 317–334, 2009.

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Proc. IEEE Computer Vision and Pattern Recognition*, pp. 731–737, 1997.

[9] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Exact recovery of corrupted low-rank matrices by convex optimization," *Proc. IEEE*, 2009.

[10] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *Journal of Ambient Intelligence and Smart Environments*, pp. 103–115, 2009.

[11] "Stanford artificial intelligence robot (stair) image dataset," *http://cs.stanford.edu/group/stair/*.

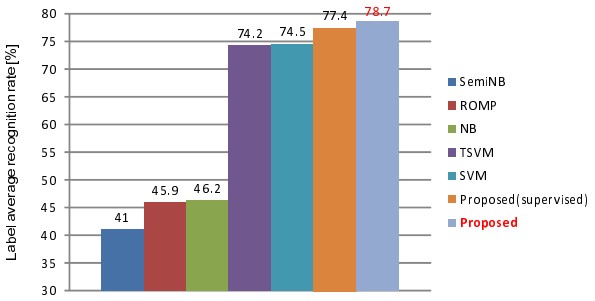

Fig. 2. Labeling accuracies of proposed and conventional methods.