Cross Language Text Classification via Multi-view Subspace Learning

Yuhong Guo and Min Xiao
Department of Computer and Information Sciences,
Temple University, Philadelphia, PA 19122 USA

Abstract

Cross language classification is an important task in multilingual learning, aiming for reducing the labeling cost of training a different classification model for each individual language. In this paper we develop a novel subspace co-regularized multi-view learning method for cross language text classification. The empirical study on a set of cross language text classification tasks shows the proposed method consistently outperforms a number of inductive methods, domain adaptation methods, and multi-view learning methods.

1 Introduction

With the rapid growth of multilingual data in all aspects of human society, it is very common that documents in different languages share the same set of categories. In such multilingual learning scenarios, applying standard monolingual classification methods directly requires costly and time-consuming document annotation in each language. Thus developing effective cross language text classification methods, which transfer the categorization knowledge from a *source language* to assist classifications in a *target language*, is becoming increasingly important.

Previous work on cross language text classification mainly focuses on the use of automatic machine translation technology. Most of these methods translate documents from the source language to the target language or vice versa, and then apply standard monolingual classification methods [3, 7]. However, due to the difference in language and culture, there exists a word drift problem. That is, while a word frequently appears in one language, its translated version may rarely appear in the other language. This creates a data distribution discrepancy between the translated training documents from the source language and the original testing documents in the target language, which poses a standard domain adaptation problem. Although many domain adaptation methods can be used in cross language text classification on the top of machine translation, e.g., the work in [8, 11, 6, 9], they nevertheless suffer from the information loss and translation error introduced in machine translation process without direct access to the original documents. Multi-view learning methods on the other hand treat each language as one independent view of the data and use both the translated documents and the original documents in each language for text classification [10, 2, 1].

In this paper, we propose a novel subspace co-regularized multi-view learning method to address cross language text classification based on machine translation. Our assumption is that a document and its translated version describe the same data object in two different views. The underlying discriminative subspace representations of the same data object in the two views thus should be very similar regarding the same classification task. We then simultaneously train two different classifiers, one for each language, by formulating a semi-supervised optimization problem that minimizes the training losses on the labeled data in both views and penalizes the distance between the two projected subspace representations of all data objects. We develop a gradient descent optimization algorithm with curvilinear search to solve the proposed optimization problem for a local optimal solution. Our extensive empirical study on a large number of cross language text classification tasks

suggests the proposed approach consistently outperforms a number of comparison inductive methods, domain adaptation methods, and multi-view learning methods. This paper is a reduced version of our previous ICML paper [5].

2 Cross Language Text Classification

Using machine translation, we can translate each document in the source language into a parallel document in the target language, and vice versa. Combing the original and translated data together in each language, we obtain two parallel matrices, $X_1 \in I\!R^{n\times d_1}$ in the source view and $X_2 \in I\!R^{n\times d_2}$ in the target view. The first l rows of X_1 and X_2 form the labeled submatrices, X_1^ℓ and X_2^ℓ, respectively. Their corresponding labels are given as a column vector $\mathbf{y} \in \{-1,+1\}^l$.

2.1 Multi-View Training with Subspace Co-regularization

We assume there is a low-dimensional subspace representation of the data in each view. The predictive function in the ith view is a linear function built over the subspace representation. Since the same classification task is shared between the two views, the underlying predictive subspace representations of the parallel documents in the two views should be very similar. We thus formulate the cross language text classification as a semi-supervised multi-view optimization problem that minimizes the training losses on the labeled data in each view while penalizing the distance between the two view subspace representations of both labeled and unlabeled data. Specifically, we conduct training by minimizing the following regularized loss over the model parameters $\{\Theta_i, \mathbf{w}_i, b_i\}_{i=1}^2$,

$$\begin{aligned} \min_{\{\Theta_i,\mathbf{w}_i,b_i\}} \quad & \sum_{i=1}^{2} \|X_i^\ell\Theta_i\mathbf{w}_i + b_i - \mathbf{y}\|^2 + \alpha_i\|\mathbf{w}_i\|^2 + \gamma\|X_1\Theta_1 - X_2\Theta_2\|_F^2 \qquad (1)\\ \text{s. t.} \quad & \Theta_1^\top\Theta_1 = I, \quad \Theta_2^\top\Theta_2 = I. \end{aligned}$$

where $\Theta_i \in I\!R^{d_i\times m}$ is the linear transformation matrix that projects the input data into the low-dimensional subspace. Below we show that the optimal $\{\mathbf{w}_i, b_i\}$ can be solved in terms of Θ_1 and Θ_2 from the optimization problem.

Lemma 1 *The optimal $\{\mathbf{w}_i^*, b_i^*\}_{i=1}^2$ that solve the optimization problem in Eq. (1) is given by*

$$\mathbf{w}_i^* = (\Theta_i^\top X_i^{\ell\top} H X_i^\ell \Theta_i + \alpha_i I)^{-1}\Theta_i^\top X_i^{\ell\top} H\mathbf{y} \qquad (2)$$

$$b_i^* = \frac{1}{l}\mathbf{1}^\top(\mathbf{y} - X_i^\ell\Theta_i\mathbf{w}_i^*) \qquad (3)$$

for $i = 1, 2$, where $H = I - \frac{1}{l}\mathbf{1}\mathbf{1}^\top$ and $\mathbf{1}$ denotes a column vector of length l with all 1 entries.

Following Lemma 1, the objective function in Eq. (1) can be rewritten as below by replacing $\{\mathbf{w}_i, b_i\}$

$$L(\Theta_1,\Theta_2) = \gamma\big\|X_1\Theta_1 - X_2\Theta_2\|_F^2 + 2\mathbf{y}^\top H\mathbf{y} - \sum_{i=1}^{2}\mathbf{z}_i^\top\Theta_i(\Theta_i^\top M_i\Theta_i + \alpha_i I)^{-1}\Theta_i^\top\mathbf{z}_i$$

where M_i and $\mathbf{z}_i$ are defined as

$$M_i = X_i^{\ell\top} H X_i^\ell \quad \text{and} \quad \mathbf{z}_i = X_i^{\ell\top} H\mathbf{y}.$$

Hence the optimization problem in Eq. (1) can be equivalently re-expressed as

$$\min_{\Theta_1,\Theta_2} \; L(\Theta_1,\Theta_2) \quad \text{s. t. } \Theta_1^\top\Theta_1 = I,\; \Theta_2^\top\Theta_2 = I. \qquad (4)$$

The problem above is a non-convex optimization problem. Nevertheless, the gradient of the objective function with respect to $\{\Theta_1, \Theta_2\}$ can be easily computed, and its part corresponding to each Θ_i is given as

$$\begin{aligned} \nabla_{\Theta_i} L(\Theta_1,\Theta_2) = \; & 2\gamma X_i^\top(X_i\Theta_i - X_{\bar{i}}\Theta_{\bar{i}}) - 2\mathbf{z}_i\mathbf{z}_i^\top\Theta_i(\Theta_i^\top M_i\Theta_i + \alpha_i I)^{-1} \\ & + 2M_i\Theta_i(\Theta_i^\top M_i\Theta_i + \alpha_i I)^{-1}\Theta_i^\top\mathbf{z}_i\mathbf{z}_i^\top\Theta_i\;(\Theta_i^\top M_i\Theta_i + \alpha_i I)^{-1} \end{aligned}$$

for $\{i = 1, \bar{i} = 2\}$ or $\{i = 2, \bar{i} = 1\}$.

2.2 Optimization Algorithm

The non-convex optimization problem (4) is generally difficult to optimize due to the orthogonal constraints. In this work, we use a gradient descent optimization procedure with curvilinear search [12] to solve it for a local optimal solution.

In each iteration of the gradient descent procedure, given the current feasible point (Θ_1, Θ_2), the gradients can be computed using (5), such that

$$G_1 = \nabla_{\Theta_1} L(\Theta_1, \Theta_2), \quad G_2 = \nabla_{\Theta_2} L(\Theta_1, \Theta_2). \tag{5}$$

We then compute two skew-symmetric matrices

$$F_1 = G_1\Theta_1^\top - \Theta_1 G_1^\top, \quad F_2 = G_2\Theta_2^\top - \Theta_2 G_2^\top. \tag{6}$$

It is easy to see $F_1^\top = -F_1$ and $F_2^\top = -F_2$. The next new point can be searched as a curvilinear function of a step size variable τ, such that

$$Q_1(\tau) = \Big(I + \frac{\tau}{2}F_1\Big)^{-1}\Big(I - \frac{\tau}{2}F_1\Big)\Theta_1 \tag{7}$$

$$Q_2(\tau) = \Big(I + \frac{\tau}{2}F_2\Big)^{-1}\Big(I - \frac{\tau}{2}F_2\Big)\Theta_2 \tag{8}$$

It is easy to verify that $Q_1(\tau)^\top Q_1(\tau) = I$ and $Q_2(\tau)^\top Q_2(\tau) = I$ for all $\tau \in I\!R$. Thus we can stay in the feasible region along the curve defined by τ. Moreover, $\frac{d}{d\tau}Q_1(0)$ and $\frac{d}{d\tau}Q_2(0)$ are equal to the projections of $(-G_1)$ and $(-G_2)$ onto the tangent space $\mathcal{Q} = \{(\Theta_1, \Theta_2) : \Theta_1^\top\Theta_1 = I, \Theta_2^\top\Theta_2 = I\}$ at the current point (Θ_1, Θ_2). Hence $\{Q_1(\tau), Q_2(\tau)\}_{\tau\geq 0}$ is a descent path in the close neighborhood of the current point. We thus apply a similar strategy as the standard backtracking line search to find a proper step size τ using curvilinear search, while guaranteeing the iterations to converge to a stationary point.

2.3 Multi-View Testing

After the semi-supervised multi-view training, we obtain two prediction models with model parameters $\{\Theta_i, \mathbf{w}_i, b_i\}_{i=1}^2$. We then conduct multi-view testing on new documents. Specifically, given a test document, $\mathbf{x} \in I\!R^{d_2}$, in the target language, we first translate it into the source language to obtain $\widehat{\mathbf{x}} \in I\!R^{d_1}$. Then we compute the prediction values using the two prediction models

$$f_1(\widehat{\mathbf{x}}) = \widehat{\mathbf{x}}^\top\Theta_1\mathbf{w}_1 + b_1, \tag{9}$$

$$f_2(\mathbf{x}) = \mathbf{x}^\top\Theta_2\mathbf{w}_2 + b_2. \tag{10}$$

The prediction confidence of each predictor can be calculated as $|f_1(\widehat{\mathbf{x}})|$ and $|f_2(\mathbf{x})|$ respectively. We finally set the prediction label for $\mathbf{x}$ as the one predicted from the most confident predictor, i.e.,

$$y = \begin{cases} \text{sign}(f_1(\widehat{\mathbf{x}})) & \text{if } |f_1(\widehat{\mathbf{x}})| > |f_2(\mathbf{x})| \\ \text{sign}(f_2(\mathbf{x})) & \text{otherwise} \end{cases} \tag{11}$$

3 Experiments

We conducted experiments on cross language text classification (CLTC) tasks constructed from a comparable multilingual corpus used in [2], which contains newswire articles written in 5 languages (English(*E*), French(*F*), German(*G*), Italian(*I*), Spanish(*S*)). We constructed a set of 20 binary cross language classification tasks over all possible source-target pairs of 5 languages, using two classes, CCAT and ECAT, as shown in Table 1. For example, **E2F** denotes the task that uses *English* as the source language and uses *French* as the target language. In each task, we used 4000 original documents and 4000 translated documents in each language.

In the experiments, we compared the proposed Subspace Co-regularized Multi-View learning method (**SCMV**) method with five other methods: (1) **TB**, a baseline method that trains a classifier using only the labeled original documents in the target language; (2) **TSB**, a baseline method that trains a classifier on both the labeled original documents in the target language and the labeled documents translated from the source language; (3) **EA++**, the co-regularization based semi-supervised

Table 1: Average classification accuracy results over 10 runs for 20 CLTC tasks.

TASKS	TB	TSB	EA++	MVMV	MVCC	SCMV
E2F	78.60±0.80	79.24±0.51	79.52±0.47	81.13±0.46	83.20±0.38	**86.10**±0.42
E2G	75.65±0.67	75.01±0.51	75.25±0.46	80.37±0.76	81.62±0.54	**83.51**±0.74
E2I	79.80±0.69	76.39±0.98	76.48±1.02	80.01±0.69	83.75±0.64	**84.87**±0.51
E2S	84.54±1.52	85.24±1.01	85.43±1.03	86.30±0.69	89.98±0.42	**92.26**±0.34
F2E	77.04±0.92	80.32±0.47	80.60±0.48	81.15±0.44	82.51±0.36	**83.86**±0.35
F2G	76.41±0.92	76.32±0.62	76.68±0.49	79.66±0.91	81.84±0.76	**83.16**±0.70
F2I	78.32±0.82	77.02±0.78	78.87±0.75	79.53±0.63	82.98±0.47	**83.25**±0.43
F2S	84.77±1.05	86.24±0.71	86.90±0.69	87.53±0.68	90.96±0.44	**92.81**±0.25
G2E	77.04±0.88	78.57±0.37	78.42±0.36	78.68±0.68	80.52±0.50	**82.52**±0.47
G2F	75.93±0.70	77.08±0.51	77.22±0.42	77.99±0.61	80.57±0.48	**83.55**±0.36
G2I	79.88±0.77	78.54±1.05	78.61±0.99	78.07±0.78	81.85±0.54	**84.20**±0.51
G2S	85.82±0.91	86.22±0.55	86.61±0.57	84.73±0.62	89.24±0.37	**90.67**±0.61
I2E	76.98±0.74	76.76±0.42	77.80±0.40	78.86±0.61	80.45±0.47	**81.34**±0.48
I2F	76.88±0.94	78.10±0.35	78.61±0.47	78.11±0.65	80.58±0.60	**81.73**±0.42
I2G	76.79±0.57	76.56±0.55	77.66±0.48	79.69±0.61	80.50±0.53	**84.76**±0.35
I2S	85.36±1.42	87.68±0.50	88.63±0.51	89.42±0.56	90.66±0.33	**94.15**±0.44
S2E	74.35±0.94	74.73±0.63	74.83±0.69	77.89±0.54	79.45±0.58	**80.50**±0.44
S2F	75.89±1.10	77.48±0.58	77.62±0.57	77.93±0.62	82.82±0.22	**84.86**±0.33
S2G	75.88±0.44	74.28±0.40	74.31±0.34	77.91±0.56	80.90±0.44	**81.12**±0.53
S2I	79.36±0.84	79.72±0.69	80.54±0.75	82.46±0.65	87.18±0.46	**88.59**±0.47

domain adaptation method developed in [4], which uses a synthetic source domain formed by translating all documents in the source language into the target language; (4) **MVMV**, the multi-view majority voting method developed in [2]; and (5) **MVCC**, the semi-supervised version of the multi-view co-classification method [1], which penalizes the disagreement of the two view predictions on unlabeled data. Among these methods, only the MVCC uses a logistic regression predictor as base classifier, and all other methods use least squares predictors as base classifiers.

For each CLTC task, we randomly chose 900 labeled and 2100 unlabeled original documents from the source language domain, and chose 100 labeled and 2900 unlabeled original documents from the target language domain for classification model training. Thus in total we had 1000 labeled documents and 5000 unlabeled documents in each language for training. We used the remaining 1000 original documents in the target language for testing. Based on this random data partition procedure, we repeated the E2F experiment 3 times to conduct model parameter selection for *MVCC* and the proposed *SCMV*. We used 10 as the subspace dimension for *SCMV*. The average classification results over 10 repeated runs on the test data in term of accuracy are reported in Table 1. We can see that between the two baseline methods, by exploiting the translated labeled documents from the source language, *TSB* has slight advantages over *TB* on many tasks. The domain adaptation method, *EA++*, however, produced similar performance as the baseline *TSB*. By exploiting both original data and translated data in the two languages, the multi-view methods, *MVMV* and *SCMV*, produced much better results. The proposed *SCMV* on the other hand consistently outperforms the other five methods on all tasks.

4 Conclusion

In this paper, we proposed a novel subspace co-regularized multi-view learning method to address cross language text classification. By training two subspace based prediction models in two language views together while penalizing the distance between the two projected subspace representations of both labeled and unlabeled instances, the underlying discriminative subspace representations can be identified to produce prediction models with better generalization performance. We developed a gradient descent algorithm with curvilinear search to solve the proposed joint optimization problem for a local optimal solution. Our extensive empirical results on a large number of cross language text classification tasks demonstrated the superior performance of the proposed method comparing to a few inductive methods, domain adaptation methods, and multi-view learning methods.

References

[1] M. Amini and C. Goutte. A co-classification approach to learning from multilingual corpora. *Machine Learning*, 79:105–121, 2010.

[2] M. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Process. Systems (NIPS)*, 2009.

[3] N. Bel, C. Koster, and M. Villegas. Cross-lingual text categorization. In *Proc. of the European Conference on Digital Libraries (ECDL)*, 2003.

[4] H. Daumé III, A. Kumar, and A. Saha. Co-regularization based semi-supervised domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[5] Y. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. In *Proc. of the International Conference on Machine Learning (ICML)*, 2012.

[6] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

[7] J. G. Shanahan, G. Grefenstette, Y. Qu, and D. A. Evans. Mining multilingual opinions through classification and translation. In *Proc. of AAAI'04 Spring Symp. on Explor. Attitude and Affect in Text*, 2004.

[8] L. Shi, R. Mihalcea, and M. Tian. Cross language text classification by model translation and semi-supervised learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

[9] C. Wan, R. Pan, and J. Li. Bi-weighting domain adaptation for cross-language text classification. In *Proc. of the Interational Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

[10] X. Wan. Co-training for cross-lingual sentiment classification. In *Proc. of the Annual Meeting of the Association for Comput. Linguistics (ACL)*, 2009.

[11] B. Wei and C. Pal. Cross lingual adaptation: an experiment on sentiment classifications. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

[12] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. Technical report, Rice University, 2010.