“© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Domain Adaptation for Gaussian Process Classification

Kai Yang[1,2,3], Wanggen Wan[1,2]
[1]School of Communication and Information Engineering,
Shanghai University, 200444, Shanghai, China
[2]Institute of Smart City,
Shanghai University, Shanghai, China
{yangkaijyjy@163.com}

Jie Lu[3]
[3]Decision System & e-Service Intelligence Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology,
University of Technology Sydney, P.O. Box 123, Broadway NSW, Australia

***Abstract*—Traditional machining learning method aims at using the labeled data or unlabeled data to train a mathematic model then it can be used to predict the unlabeled data for Data mining problem, but it requires that the data which be trained should have same distribution with the predicting data. For the real world datasets, it is hard to get enough training datasets which has the same distribution. Thus, how to train a good mathematic model by using different distribution data is crucial problem, and the researchers using the probability view to solve transfer classification problem is relative less. In this paper, we propose a transfer classification algorithm based on the Gaussian Process model, which can be used to solve the homogeneous transfer classification problem. We use the probability theory to propose a novel classification transfer learning model based on the Gaussian Process (GP) model. We experiment on the synthetic and real-world datasets and compare to other method, the result has verified the effectiveness of our approach.**

Keywords—Gaussian Process; Domain Adaptation; Transfer Learning; Homogeneous

I. INTRODUCTION

The Machine Learning techniques already have made great contribution for the Artificial Intelligence (AI). However, many machine learning methods have good performance should be satisfied a same condition, it is the feature space and the distribution of the training data and the test data should be same. When the distribution of the data are different, the traditional machine learning method can't train a good model to predict the data. In order to solve this problem, recently, transfer learning has been widely studied [1-3].

Until now, many researchers already have got some good performance for classification transfer, Literature [4] is using the boosting theory to solve the classification transfer problem, during the training model, it add the penalty weight item to reweight the target data by using the source data. Literature [5] is using the SVM theory to consider the transfer classify problem, in order to help train a good model by using the auxiliary data, the author add the regularization item to the objective loss function, it can constrain the source data to train the model more similarity with the auxiliary data. Literature [6] it proposed the transfer learning kernel to deal with a domain-invariant kernel by using the kernel theory, it use the spectral kernels for the target eigensystem on source domain. Literature [7] it use the Gaussian Process model to solve the transfer learning regression problem by probability theory. The researchers using the probability theory to solve transfer classification problem is relative less. In this article, we proposed the classification transfer algorithm based on the GP model by using the probability theory. Compare to other algorithm, the proposed algorithm has good performance.

The rest of the paper is organized as follows. In Section 3, we propose our algorithm. In Section 4, we do the experiment on synthetic datasets and real world datasets. Finally, we give some concludes some future works.

II. METHOD

In this article, we consider the homogenous transfer learning classification problem based on the Gaussian Process (GP) model, the GP model is different from the SVM model which the classical classification method, the GP is based on the Bayes theory and use the probability distribution to replace the point distribution.

A. Gaussian Process

The linear function $f = w^T x$, we assume that $y = f + \varepsilon$, where ε is noise term, the ridge regression estimate objective function of f is

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \| w \|^2 \qquad (1)$$

where $\|w\| = w^T w$ is the regularization item of weights w.

From the Bayes theory view, the regularization of weights is come from the prior distribution of the weights, thus, the regularization is assumed to be satisfied the Gaussian distribution. For the posteriori distribution of w is

$$p(w \mid D) = \frac{1}{Z} p(y \mid w, X) p(w) \propto \exp\{-\frac{1}{2} J(w)\}, \qquad (2)$$

where $Z = p(y|x)$, and

$$J(w)=\frac{1}{\sigma^2}\sum_i (w^T x_i - y_i)^2 + \| w \|^2 \qquad (3).$$

If $\sigma^2=\lambda$, the estimate function is the same with the ridge regression, thus, the Gaussian Process (GP) is defined as $f \sim GP(m,k)$, where m is mean function, k is convenience function, GP is the joint Gaussian distributions [8].

B. Laplace Approximation Binary Classifier

For the classification problem,

$$p(y \mid x)=\frac{p(y)p(x \mid y)}{\sum_{c=1}^{C} p(C_c)p(x \mid C_c)}, \qquad (4)$$

where C is the classifier, the method is same as the SVM method, the linear regression model $p(C_1 \mid x)=\lambda(x^T w)$, where $\lambda(z)=\frac{1}{1+\exp(-z)}$.

We consider the homogenous transfer learning problem, we define the source domain S and target domain T, for the $p(y \mid X)$, where $X=(X_S, X_T)$, $y=(y_S, y_T)$, the objective function is $p(y_T \mid y_S, X_S, X_T)$, we use the new transfer kernel K is

$$K_{nm}=\begin{cases} k(x_n, x_m), & x_n, x_m \in X_S \text{ or } X_T \\ \lambda k(x_n, x_m), & otherwise \end{cases} \qquad (5).$$

The transfer kernel should be a positive semi-definite (PSD) matrix, where the $|\lambda| \le 1$, the proof as shown follow:

For any number λ let

$$K(\lambda)=\begin{bmatrix} K_{11} & \lambda K_{12} \\ \lambda K_{21} & K_{22} \end{bmatrix} \qquad (6)$$

$K(\lambda)$ to be a positive semidefinite (PSD) matrix merely means that for all vectors x and y of suitable dimensions, it satisfied that

$$\begin{aligned} 0 &\le (x' \;\; y')K(\lambda)\begin{pmatrix} x \\ y \end{pmatrix} \\ &= x'K_{11}x + 2\lambda y'K_{21}x + y'K_{22}y \end{aligned} \qquad (7)$$

This is what we have to prove when $|\lambda| \le 1$.

We are told that $K(1)$ is PSD matrix,

Where $K(1)=\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$. (8)

We claim that $K(-1)$ also is PSD. This follows by negating y in formula (7):

$$\begin{aligned} 0 &\le (x' \;\; -y')K(1)\begin{pmatrix} x \\ -y \end{pmatrix} \\ &= x'K_{11}x + 2(-y')K_{21}x + (-y')K_{22}(-y) \\ &= x'K_{11}x + 2(-1)y'K_{21}x + y'K_{22}y \\ &= (x' \;\; y')K(-1)\begin{pmatrix} x \\ y \end{pmatrix} \end{aligned} \qquad (9)$$

Notice that $K(\lambda)$ can be expressed as a linear function of the extremes $K(1)$ and $K(-1)$:

$$K(\lambda)=\frac{1-\lambda}{2}K(-1)+\frac{1+\lambda}{2}K(1) \qquad (10)$$

When $|\lambda| \le 1$ both coefficients $\frac{1-\lambda}{2}$ and $\frac{1+\lambda}{2}$ are non-negative. Thus,

$$\begin{aligned} &(x' \;\; y')K(\lambda)\begin{pmatrix} x \\ y \end{pmatrix} \\ &=\left(\frac{1-\lambda}{2}\right)(x' \;\; y')K(-1)\begin{pmatrix} x \\ y \end{pmatrix}+\left(\frac{1+\lambda}{2}\right)(x' \;\; y')K(1)\begin{pmatrix} x \\ y \end{pmatrix} \\ &\ge 0+0=0 \end{aligned} \qquad (11)$$

Because x and y are arbitrary vector, so we have proven the $K(\lambda)$ is a positive semidefinite (PSD) matrix, where the $|\lambda| \le 1$.

According to the Gaussian Process (GP) model and the new transfer kernel, we combine them to propose a transfer classification algorithm by using the Laplace Approximation method [9].

The training model algorithm as follow:

Algorithm: The training model algorithm

Input K transfer kernel, $y=\pm 1$ task label, $p(y_T \mid f_T)$ likelihood function

$f_S := 0$

Repeat

$L_S := cholesky(K_{SS})$

$W_S := -\nabla\nabla \log p(y_S \mid f_S)$

$L_S := cholesky(I + W_S^{\frac{1}{2}} K_{SS} W_S^{\frac{1}{2}})$

$$b_S := W_S f_S + \nabla \log p(y_S \mid f_S)$$

$$f_1 := K_{SS} a_S$$

$$f_2 := \lambda K_{TS} L_S^T \setminus L_S f_1$$

$$W_T := -\nabla\nabla \log p(y_T \mid f_T)$$

$$L_T := cholesky(I + W_T^{\frac{1}{2}} K_{TT} W_T^{\frac{1}{2}})$$

$$b_T := W_T f_T + \nabla \log p(y_T \mid f_T)$$

$$a_T := b_T - W_T^{\frac{1}{2}} L_T{}^T \setminus (L_T \setminus (W_T^{\frac{1}{2}} K_{TT} b_T))$$

Until convergence

$$\log q(y \mid X, \theta) := -\frac{1}{2} a_T{}^T f_2 + \log p(y_T \mid f_2) - \sum_i \log L_{Tii}$$

Return $\hat{f} := f_2$, $\log q(y_T \mid X_T, \theta)$

The prediction function as follow:

Algorithm: The prediction function

Input: $\hat{f}$ trained mode, X training data, $y = \pm 1$ task label, K transfer kernel, $p(y \mid f)$ likelihood function, x_* test data.

$$W := -\nabla\nabla \log p(y \mid f)$$

$$L := cholesky(I + W^{\frac{1}{2}} K W^{\frac{1}{2}})$$

$$\bar{f}_* := \lambda K_{TS}{}^T \nabla \log p(y \mid \hat{f})$$

$$v := L \setminus (W^{\frac{1}{2}} \lambda K_{ST})$$

$$V[f_*] := K_{TT} - v^T v$$

$$\bar{\pi}_* := \int \sigma(z) N(z \mid \bar{f}_*, V[f_*]) dz$$

Return $\bar{\pi}_*$

III. EXPERIMENT

In this experiment, we first use the synthetic dataset to test the proposed algorithm, we assume the $X_S = X_T$, $P(X_S) \neq P(X_T)$, $y_S = y_T$, $P(y_S \mid X_S) \approx P(y_T \mid X_T)$, the Figure 1 has shown the source domain datasets and its classify result, the Figure 2 has shown the target domain datasets and its classify result, the Figure 3 has shown the combined datasets with source and target, we can see the distribution of the source domain and target domain is different, but the classify result is depending on the target domain or not on the source domain, this result has shown the proposed algorithm has transfer characteristic.

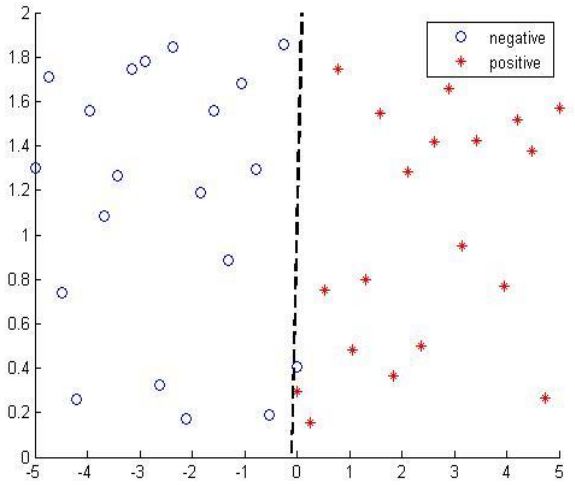

Figure 1. Source domain datasets and classify result.

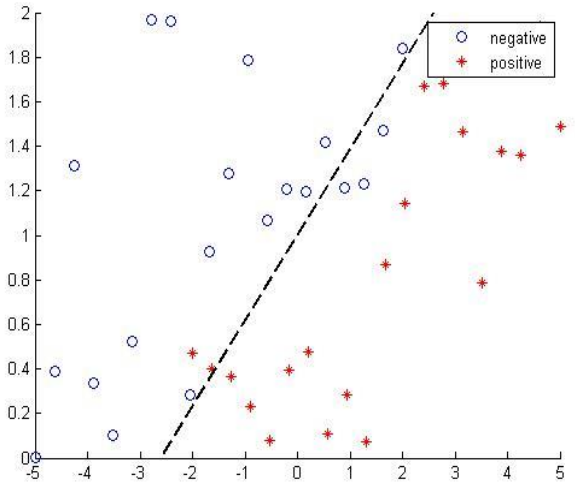

Figure 2. Target domain datasets and classify result.

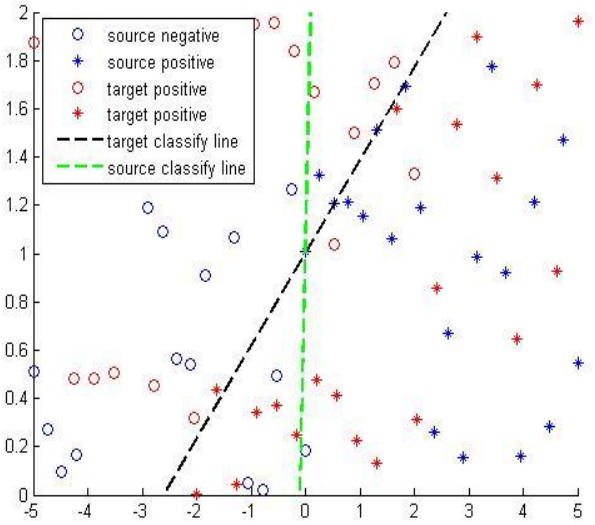

Figure 3. Result of the proposed algorithm

For the real-world datasets 20-newsgroups datasets, we select 3 groups datasets, Talk, Sci and Rec, the datasets as shown in Table 1, and the result as shown in Table 2. Compare to the boosting algorithm [4], the proposed algorithm has better performance.

TABLE 1. THE DESCRIPTION OF 20-NEWSGROUPS DATASETS.

Data Set	Feature	Source domain		Target domain	
		Postive	Negtive	Postive	Negtive
Rec vs Talk	30165	1000	1000	1000	1000
Rec vs Sci	29644	1000	1000	1000	1000
Sci vs Talk	33151	1000	1000	1000	1000

TABLE 2. CLASSIFICATION ACCURACIES OF THE EXPERIMENT RESULT

Data Set	TrAdaBoost	Proposed algorithm
Rec vs Talk	80.43%	81.52%
Rec vs Sci	78.71%	79.58%
Sci vs Talk	74.40%	74.71%

IV. CONCLUSION

In this article, we use the probability theory and the new transfer kernel to propose a transfer classification algorithm based on the Gaussian Process (GP) model, the proposed algorithm can solve the homogenous transfer classification problem, through the synthetic datasets and real-world 20 newsgroups datasets, the result has shown the effective. In the future work, we will consider the multiclass problem and the heterogonous transfer classification problem.

ACKNOWLEDGMENT

This work is supported by the Key Project of Shanghai Scientific Committee (No. 17511106802).

REFERENCES

[1] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[2] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-taught learning: Transfer learning from unlabeled data,” in Proceedings of the 24th International Conference on Machine Learning, Corvalis, Oregon, USA, pp. 759–766, June 2007.

[3] N. D. Lawrence and J. C. Platt, “Learning to learn with the informative vector machine,” in Proceedings of the 21st International Conference on Machine Learning. Banff, Alberta, Canada: ACM, July 2004.

[4] Dai, W., Yang, Q., Xue, G.-R. & Yu, Y., 'Boosting for transfer learning', Proceedings of the 24th international conference on Machine learning, ACM, pp. 193-200, 2007.

[5] Wu, P. & Dietterich, T.G., 'Improving SVM accuracy by training on auxiliary data sources', Proceedings of the twenty-first international conference on Machine learning, ACM, pp. 110-118, 2004.

[6] Long, M., Wang, J., Sun, J. & Philip, S.Y., 'Domain invariant transfer kernel learning', IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1519-32, 2015.

[7] Cao, B., Pan, S.J., Zhang, Y., Yeung, D.-Y. & Yang, Q., 'Adaptive Transfer Learning', AAAI, vol. 2, pp. 407-412, 2010.

[8] Rasmussen, C.E., 'Gaussian processes in machine learning', Advanced lectures on machine learning, Springer, pp. 63-71, 2004.

[9] Williams, C.K. & Rasmussen, C.E., 'Gaussian processes for machine learning', the MIT Press, vol. 2, no. 3, pp. 33-47, 2006.