Dictionary-Based Transfer Learning for One-Class Text Classification

Haoxin Xie[a], Bo Liu[b], Yanshan Xiao[c]

[a]howhyn@gmail.com, [b]csbliu@gmail.com, [c] xiaoyanshan@gmail.com

Department of Computer Science, Guangdong University of Technology Guangzhou, China

Abstract— One-class classification requires only positive samples to build the model. In this paper, we propose a novel method for one-class text learning based on transfer learning and dictionary learning (DTOC). We integrate transfer learning and dictionary learning into one-class text model. In order to solve DTOC model, the interactive framework updates each variable by fixing other variables, and updates one-class classifier and dictionary alternatively to obtain the predictive classifier. Finally, extensive experiments show that our method has higher competitive classification accuracy and less sensitive to label noise than existing state-of-the-art methods.

Keywords—one-class learning, transfer learning, dictionary learning

1. INTRODUCTION

Text classification is a basic task in the field of natural language processing [1]. It attracts growing attention in news analysis, information retrieval, sentiment analysis, spam classification and public opinion monitoring. To date, text classification traditionally divided into on-line and off-line classification [2]. In on-line classification, a time ordered sequence of coordinates is capturing the real-time text data of users when building classification. In off-line classification, only the entire text data has been collected completely before building classification.

In most text classifications, One-class text classification (OC) has attracted growing attention in the problem of class imbalance on text data [3,4]. The approach to OC selects the one class as the target class, and the other classes as the nontarget classes, where the number of the target class is far less than the nontarget classes. Then OC builds the classifier based on the target class. For OC, the authors in [4] build one-class naive Bayes on the target data and examine the clustering quality of the classification. Shravan et al. [5] utilize principal component analysis to reduce dimension and build the one-class classifier to perform classification.

Despite much progress in one-class text classification, most of studies still focus on single task learning. However, there always exist limited labeled examples of a new task, where a new task is regarded as the target task. To overcome the above problem, we need to transfer knowledge from the source task to the target task so as to build the one-class text classifier. For example, in the 20 Newsgroup dataset, when the user’s interest changes from “sport baseball” to “sport hockey”, we hope to transfer knowledge from the labeled samples of “sport baseball”, and then construct one-class text classifier for the limited labeled examples of “sport hockey”. Another important observation is that, dictionary learning methods are proposed to represent the input

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334296

data by using a linear combination of synthesis dictionary, whose low-rank data representations generally reduce redundancy and the impact of noise data. For example, the work in [6] proposes a denoising method based on dictionary learning and wavelet analysis, which can reduce the calculation time and improve the efficiency. The work in [7] proposes a noise-related method based on double dictionary transform learning, which includes analysis dictionary and synthetic dictionary.

In this paper, we propose a novel method for one-class text learning, termed as dictionary-based transfer learning for one-class text classification (DTOC), which extends our previous on-line one-class method [3] to construct the off-line one-class method. Specially, for the source task and the target task, we first select the one-class as the target class, and the other classes as the nontarget classes. And then we build a DTOC model based on transfer learning and one-class learning. In all, the main contributions of the paper are be listed as follows:

- We first design a novel method for one-class text classification, and then propose a dictionary learning based on transfer learning (DTOC). For the target task, the one-class text classifier is learned by the knowledge of the source task. In addition, dictionary learning reduces the effect of uncertain data.
- We then propose an iterative framework to solve the proposed DTOC model. The interactive framework updates each variable by fixing other variables.
- Extensive experiments show that our method has higher competitive classification accuracy and less sensitive to label noise than existing state-of-the-art methods.

2. METHOD

Suppose there are 2 tasks. Let $X_1 = [x_{11},\ x_{12}, \dots, x_{1n}] \in \mathbb{R}^{r\times n}$ represents the source task, while $X_2 = [x_{21},\ x_{22}, \dots, x_{2n}] \in \mathbb{R}^{r\times n}$ represents the target task. Inspired by the work in [8], Let $D_t = [d_{t1},\ d_{t2}, \dots, d_{tk}] \in \mathbb{R}^{n\times k}$ represents synthesis dictionary. The synthesis dictionary represents the input data by using a linear combination of synthesis. $S_t = [S_{t1},\ S_{t2}, \dots, S_{tn}] \in \mathbb{R}^{k\times n}$ is the coding coefficient, which converts data X_t into low-rank representation. $P_t = [P_{t1}, P_{t2}, \dots, P_{tn}] \in \mathbb{R}^{k\times n}$ is an analysis dictionary, which can bridge input data with the approximated coding coefficients. Then we have

$$\min_{S,D,P} \|X_t - D_t S_t\|_F^2 + \tau \left[\|P_t X_t - S_t\|_F^2 + \|S_t\|_{2,1} \right]$$
$$\text{s.t.}\ \ \|d_v\|_2^2 \le 1, \mathrm{v} \in \{1, \dots, \mathrm{k}\}. \qquad (1)$$

where $\|d_v\|_2^2 \le 1$ is the constraint of dictionary atom, which ensures the stability of DTOC calculation.

According to dictionary learning, $\|X_t - D_t S_t\|_F^2$ is the sparse code extraction term. The learned coding coefficient S_t can be directly used as a feature for classification. To represent distinct information between two tasks, we hope the coding coefficient S_t only represents X_t well, otherwise, $D_t \bar{S}_t \approx 0$. $\bar{S}_t$ is the complementary matrix of S_t, and $S_t = [S_1,\ S_2]$. So the analysis incoherence term is

$$\min_{S,D} \|D_t \bar{S}_t\|_F^2 \qquad (2)$$

In this paper, we learn a synthesis dictionary D_t and a analysis dictionary P_t together, such that this coding coefficient S_t can be approximated as $S_t \approx P_t X_t$. Then, we let the $P_t X_t$ as the feature for classification of X_t. To make most of distinct information between two tasks, we hope $P_t X_t$ represents badly the feature for classification of $\bar{X}_t$, $P_t \bar{X}_t \approx 0$. $\bar{X}_t$ is the complementary matrix of X_t and $X_t = [X_1,\ X_2]$. So the analysis sparse code extraction term is

$$\min_{P} \| P_t \bar{X}_t \|_F^2 \qquad (3)$$

Suppose the source task X_1 has N_1 positive samples, while the target task X_2 has N_2 positive samples. The one-class classifier for X_t is $\mathrm{f}_t(\mathrm{x}) = w_t^T x - \rho_t$. According to transfer dictionary learning, we have

$$w_1 = w_0 + v_1,\ w_2 = w_0 + v_2 \qquad (4)$$

Where w_0 is the weight of global classification hyperplane for two tasks and represents the knowledge between two tasks. v_1 and v_2 represent the displacement based on the global classification hyperplane. $w_0 + v_1$ is the weight of classification hyperplane for the source task, while $w_0 + v_1$ is the weight of classification hyperplane for the target task. Hence, we can obtain objective function of transfer learning for one-class classification:

$$\begin{gathered} \min_{P} \|w_0\|^2 + c_1\|v_1\|^2 + c_2\|v_2\|^2 - \rho_1 - \rho_2 + \sum_{t=1}^{2} \mathrm{c}_t \sum_{i=1}^{N_t} \xi_{ti} \\ \text{s.t.} \quad (w_0 + v_1)^T P_1 x_{1i} \geq \rho_1 - \xi_{1i},\ \mathrm{i} \in \{1, \dots, N_1\} \\ (w_0 + v_2)^T P_2 x_{2j} \geq \rho_2 - \xi_{2j},\ \mathrm{j} \in \{1, \dots, N_2\} \\ \xi_{1i} \geq 0,\ \xi_{2j} \geq 0 \qquad (5) \end{gathered}$$

Where c_1 and c_2 are regularized parameters, $c_1 < c_2$, which means the source task improves the classification performance of the target task. ξ_{1i} and ξ_{2j} are training errors.

Eqs. (2)-(3) are the constraint of Eqs. (1). Eqs. (1)-(3) represent the performance of data representation, while Eqs. (5) represents the performance of classification. By considering Eqs. (1)-(5), we obtain objective function of dictionary-based transfer learning for one-class classification:

$$\begin{aligned} \min_{S,D,W,P} \textstyle\sum_{t=1}^{2} \quad & \{\|X_t - D_t S_t\|_F^2 \\ & +\tau\left[\|P_t X_t - S_t\|_F^2 + \|S_t\|_{2,1}\right]\} \\ & +J(w_0, v_t, P_t, \rho_t, \xi_t) \\ & \text{s.t. } \|d_v\|_2^2 \leqslant 1,\ v \in \{1, \dots, k\} \qquad (6) \\ & (w_0 + v_1)^T P_1 x_{1i} \geq \rho_1 - \xi_1,\ i \in \{1, \dots, N_1\} \\ & (w_0 + v_2)^T P_2 x_{2j} \geq \rho_2 - \xi_2, j \in \{1, \dots, N_2\} \\ & \xi_{1i} \geq 0,\ \xi_{2j} \geq 0 \end{aligned}$$

Where

$$\begin{aligned} J(w_0, v_t, P_t, \rho_t, \xi_t) = \quad & \|w_0\|^2 + c_1\|v_1\|^2 + c_2\|v_2\|^2 \\ & -\rho_1 - \rho_2 + \textstyle\sum_{t=1}^{2}\ c_t \sum_{i=1}^{N_t}\ \xi_{it} \end{aligned} \qquad (7)$$

To facilitate optimization, let $\mathbb{O} = (0, 0, \dots, 0)^T$ and $\mathbb{I} = (1, 1, \dots, 1)^T$ are T- dimensional column vectors. We have

$$\mathrm{w} = \left(w_0, \sqrt{\frac{c_1}{T}}v_1, \sqrt{\frac{c_2}{T}}v_2\right)^T \qquad (8)$$

$$[\mathrm{z}(P_1x_{1i}) \quad \mathrm{z}(P_2x_{2i})] = \begin{pmatrix} P_1x_{1i} & P_2x_{2i} \\ P_1\sqrt{\frac{T}{c_1}}x_{1i} & \mathbb{O} \\ \mathbb{O} & P_2\sqrt{\frac{T}{c_2}}x_{2i} \end{pmatrix}$$

$$= [P_1\mathrm{z}(x_{1i}) \quad P_2\mathrm{z}(x_{2i})] \qquad (9)$$

$$(e_1 \quad e_2) = \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & 0 \end{pmatrix} \qquad (10)$$

By considering Eqs. (8)-(10), we rewrite objective function dictionary-based transfer learning for one-class classification:

$$\min_{S,\mathrm{D},w,P} \sum_{t=1}^{2} \{\|X_t - D_tS_t\|_F^2 + \lambda\|D_t\bar{S}_t\|_F^2$$
$$+\tau\left[\|P_tX_t - S_t\|_F^2 + \|P_t\bar{X}_t\|_F^2 + \|S_t\|_{2,1}\right]\}$$
$$+\|w\|^2 - \rho^Te + \sum_{t=1}^{T} c_t \sum_{i=1}^{N_t} \xi_{ti}$$
$$\mathrm{s.t.}\ \|d_v\|_2^2 \leq 1, \mathrm{v} \in \{1, \dots, \mathrm{k}\},$$
$$w^Tz(P_tx_{ti}) \geq \rho^Te_t - \xi_{ti}, i = 1, \dots, N_t, t = 1,2,$$
$$\xi_{ti} \geq 0 \qquad (11)$$

3. OPTIMIZATION

Inspired by the work in [3], our interactive framework updates each variable by fixing other variables.

3.1 Fixing $\boldsymbol{D}, \boldsymbol{w}, \boldsymbol{P}, \boldsymbol{\xi}$ and Optimize $\boldsymbol{S}$

Let $\|S_t\|_{2,1} = 2\mathrm{tr}\ (S_t^T\Lambda_tS_t)$, $\Lambda_{tuu} = 1/2\|S_t^u\|_2$. To remove terms that are irrelevant to S_t, we have:

$$\wp(S) = \sum_{t=1}^{2} \{\|X_t - D_tS_t\|_F^2 + \tau\|P_tX_t - S_t\|_F^2 + \tau tr\ (S_t^T\Lambda_tS_t)\}$$

to set the derivative $\frac{\partial \oint\ (S)}{S} = 0$, we have

$$S_t^* = (D_t^TD_t + \tau I + \tau\Lambda_t)^{-1}(\tau P_tX_t + D_t^TX_t) \qquad (12)$$
$$\Lambda_{tuu} = 1/2\|S_t^u\|_2. \qquad (13)$$

3.2 Fixing $\boldsymbol{S}, \boldsymbol{w}, \boldsymbol{\rho}, \boldsymbol{\xi}$ and optimize $\boldsymbol{P}$

To remove terms that are irrelevant to P_t, we have

$$L(P) = \sum_{t=1}^{2} \{\tau\|P_tX_t - S_t\|_F^2 - \sum_{i=1}^{N_t} \alpha_{ti} w^T P_t z(x_{ti})^T\}$$

where α_{ti} is nonnegative Lagrange multipliers. to set the derivative $\frac{\partial L_1(P)}{\partial P} = 0$, we have

$$P_t^* = (S_tX_t^T + H_t)(X_tX_t^T + \gamma I)^{-1} \qquad (14)$$

$$\mathrm{H}_t = \frac{1}{2\tau}\sum_{i=1}^{N_t} \alpha_{ti} w z(x_{ti}, t)^T \qquad (15)$$

where $\gamma = 1e - 4$ is a small constant.

3.3 Fixing $\boldsymbol{P}, \boldsymbol{D}, \boldsymbol{S}$ and optimize $\boldsymbol{w}$

To remove terms that are irrelevant to w, we have

$$L_2(\mathrm{w}) = \|w\|^2 - \rho^T e + \sum_{t=1}^{T} c_t \sum_{i=1}^{N_t} \xi_{ti}$$

$$+ \sum_{t=1}^{T}\sum_{i=1}^{N_t} \alpha_{ti}[\rho^T e_t - \xi_{ti} - w^T z(P_t x_{ti})]$$

$$- \sum_{t=1}^{T}\sum_{i=1}^{N_t} \beta_{ti}\xi_{ti} \qquad (16)$$

We obtain the dual form of Eqs. (16):

$$\max - \frac{1}{4}\sum_{t=1}^{2}\sum_{h=1}^{2}\sum_{i=1}^{N_t}\sum_{j=1}^{N_t} \alpha_{ti} z(P_t x_{ti})^T z(P_h x_{hj}) \alpha_{hj}$$

$$\text{s.t.} \sum_{t=1}^{2}\sum_{j=1}^{N_t} \alpha_{ti} e_t = e,$$

$$0 \le \alpha_{ti} \le c_t, i \in \{1, \dots, N_t\}, t \in \{1,2\} \qquad (17)$$

where $\alpha_{ti} \ge 0$ and $\alpha_{hj} \ge 0$ are Lagrange multipliers. We have:

$$w^* = \frac{1}{2}\sum_{t=1}^{2}\sum_{i=1}^{N_t} \alpha_{ti} \mathrm{z}(x_{ti}, t) \qquad (18)$$

3.4 Fixing $\boldsymbol{S}, \boldsymbol{P}, \boldsymbol{w}, \boldsymbol{\xi}$ and optimize $\boldsymbol{D}$

To remove terms that are irrelevant to D_t, we have

$$D^* = \sum_{t=1}^{T} \|X_t - D_t S_t\|_F^2,$$

$$\text{s.t. } \|d_{tv}\|_2^2 \leqslant 1,\ v \in \{1, \dots, k_t\} \qquad (19)$$

We obtain the dual form of Eqs. (19):

$$g(\eta) = \mathrm{i}\ \left\{\|X_t - D_t S_t\|_F^2 + \sum_{i=1}^{k} \eta_{ti}\left(\|d_i\|^2 - 1\right)\right\},$$

where η_{ti} represents the Lagrange multiplier, let $(M_t)_{ii} = \eta_{ti}$, we rewrite Eqs. (19) as follow:

$$L_3(D) = \|X_t - D_t S_t\|_F^2 + \mathrm{tr}\ (D_t^T D_t M_t) + \mathrm{tr}\ (M_t)$$

To set the derivative $\frac{L_3(D)}{D} = 0$, we have

$$D_t^* = X_t S_t^T (S_t S_t^T + M_t)^{-1} \qquad (20)$$

And then we use the optimal P and w to obtain the liner classifier for the target task:

$$f_t\big(z(P_t x_{ti}, t)\big) = w^T z(x_{ti}, t) P_t x_{ti} - \rho_t$$

$$= \frac{1}{2}\left(I_0 + \sqrt{\frac{T}{c_t}} I_0\right)^T \sum_{t=1}^{2} \sum_{i=1}^{N_t} \alpha_{ti} z(x_{ti}, t)^T P_t x_{ti} - \rho_t.$$

Setting $z_t = P_t x$, we can obtain the nonlinear classifier for the target task:

$$f_t\big(\phi(z_t)\big) =$$

$$\frac{1}{2}\left(I_0 + \sqrt{\frac{T}{c_t}} I_0\right)^T \sum_{t=1}^{2} \sum_{i=1}^{N_t} \alpha_{ti} K(z(x_{ti}, t), z_t) - \rho_t.$$

4 EXPERIMENTS

4.1 Dataset

We have experimented with several datasets, which are widely used in one-class learning and transfer learning [9]. The datasets have 20 Newsgroup, Reuters-21578 and Mushroom. According to the work in [3], we reorganize three text datasets into six one-class text sub-dataset in table 1.

TABLE 1 DATASETS

Sub-dataset	Source task	Target task
Dataset 1	comp.sys.mac.hardware	comp.os.ms-windows.misc
Dataset 2	comp.sys.mac.hardware	comp.sys.ibm.pc.hardware
Dataset 3	rec.sport.baseball	rec.sport.hokey
Dataset 4	rec.sport.baseball	rec.autos
Dataset 5	orgs(2).{. . . }	orgs(1).{. . . }
Dataset 6	edible.tapering	edible.enlarging

4.2 Baselines

- The first baseline is robust one-class SVM [10] (ROC), which designs hinge loss function to improve the performance of classification.
- The second baseline is support vector data description [11] (SVDD), which can handle with noise.
- The third baseline is transfer learning for one-class SVM [12] (TLOC), which can solve the problem of limited data.

4.3 Experiment setting

For our method DTOC, the RBF kernel is :

$$K(x_{ti}, x_{hj}) = \exp\left(\frac{-\left\|x_{ti} - x_{hj}\right\|_2^2}{2v^2}\right).$$

The parameter λ and τ are selected from the range {0.001,0.005,0.01,0.05,0.1,0.5,1,5}. The parameter $c_t$are selected from the range {1,10,100,500,1000}.

4.4 Performance comparison

TABLE 2 F-MEASURE VALUES

Sub-dataset	ROC	SVDD	TLOC	DTOC
Dataset 1	72.11	76.41	77.66	84.07
Dataset 2	78.38	80.02	81.62	86.29
Dataset 3	79.36	81.74	84.64	88.18
Dataset 4	71.08	74.81	72.50	75.55
Dataset 5	75.84	80.35	81.77	85.26
Dataset 6	85.16	88.24	89.49	93.19

From Table 2, we can see that our DTOC method attains significantly better improvements over the baselines on most datasets. For example, on dataset 1, the F-measure value of DTOC is 84.07, which is better than ROC (72.11) and SVDD (76.41) . This is because that our DTOC method can improve the classification performance of the target task by transferring knowledge from the source task, while ROC and SVDD only build classifiers based on single task. In addition, the F-measure value of our DTOC method is also is better than TLOC, since dictionary learning can fully represent the potential information of data.

4.5 Performance on different noise levels

In order to evaluate the noise sensitivity of our DTOC method and the baselines, we add noise to the data by referring to the work in [1].

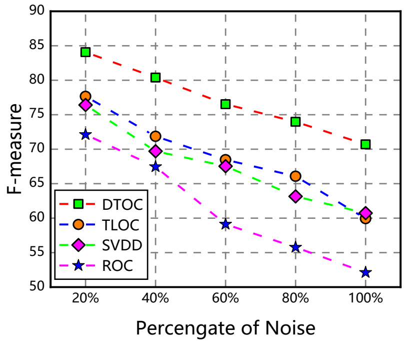

Figure 1. Dataset 1

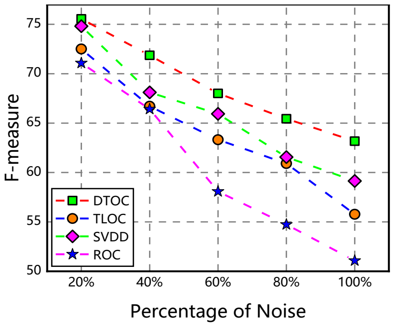

Figure 2. Dataset 4

Figure 1 and figure 2 present the corresponding classification accuracy on the dataset 1 and dataset 4. It can be seen that when the percentage of labeling noise increases from 20% to 100%, the accuracy of our method and the baselines decreases. This is because that the increase of noise makes the distribution of target class and non-target classes insignificant. In addition, the *F*-measure value of DTOC is always better than the baselines. Because our DTOC method is constructed by dictionary learning, which can reduce the impact of noise by sparse encoding.

5 Conclusion

In this paper, we propose a novel method for one-class text learning, termed as dictionary-based transfer learning for one-class text classification (DTOC). We integrate transfer learning and dictionary learning into one-class text model. In order to solve DTOC model, the interactive framework updates each variable by fixing other variables, and updates one-class classifier and dictionary alternatively to obtain the predictive classifier. Finally, extensive experiments show that our method has higher competitive classification accuracy and less sensitive to label noise than existing state-of-the-art methods.

For DTOC model, the shared parameter w_0 represents the knowledge between two tasks in classification term. In the future, we plan to design the shared parameter D_0, which represents the knowledge between two tasks in dictionary term.

References

[1] Kaur J, Saini J R. A study of text classification natural language processing algorithms for Indian languages[J]. VNSGU J Sci Technol, 2015, 4(1): 162-167.

[2] Chahi A, El Merabet Y, Ruichek Y, et al. Off-line text-independent writer identification using local convex micro-structure patterns [C] Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society. 2019: 1-5.

[3] Xie H, Liu B, Xiao Y. Transfer learning-based one-class dictionary learning for recommendation data stream[J]. Information Sciences, 2021, 547: 526-538.

[4] Zhang Y, Jatowt A. Estimating a one-class naive Bayes text classifier[J]. Intelligent Data Analysis, 2020, 24(3): 567-579.

[5] Shravan Kumar B, Ravi V. Text document classi-fication with PCA and one-class SVM[C]Proceedings of the 5th international con-ference on frontiers in intelligent computing: the-ory and applications. Springer, Singapore, 2017: 107-115.

[6] Huang Y, Li W, Yuan F. Speckle noise reduction in sonar image based on adaptive redundant dictionary [J]. Journal of marine science and engineering, 2020, 8(10): 761.

[7] Liao M, Fan X, Li Y, et al. Noise-related face im-age recognition based on double dictionary transform learning[J]. Information Sciences, 2023.

[8] Gu S, Zhang L, Zuo W, et al. Projective dictionary pair learning for pattern classification[J]. Advances in neural information processing systems, 2014, 27.

[9] Li J, Wu W, Xue D. Transfer Naive Bayes algorithm with group probabilities[J]. Applied Intelligence, 2020, 50: 61-73.

[10] Xing H J, Ji M. Robust one-class support vector machine with rescaled hinge loss function[J]. Pattern Recognition, 2018, 84: 152-164.

[11] Yin L, Wang H, Fan W. Active learning based support vector data description method for robust novelty detection[J]. Knowledge-Based Systems, 2018, 153: 40-52.

[12] Xue Y, Beauseroy P. Transfer learning for one class SVM adaptation to limited data distribution change[J]. Pattern Recognition Letters, 2017, 100: 117-123.