Neurocomputing 351 (2019) 205–216

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Steganalysis on Internet images via domain adaptive classifier

Yong Yang[a], Xiangwei Kong[b,*], Bo Wang[a], Ke Ren[a], Yanqing Guo[a]

[a] *School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China*
[b] *Department of Data Science and Engineering Management, Zhejiang University, Hangzhou 310058, China*

ARTICLE INFO

Article history:
Received 20 May 2018
Revised 22 February 2019
Accepted 10 April 2019
Available online 16 April 2019

Communicated by Dr. Ivor Tsang

Keywords:
Steganalysis
Domain adaptation
Mismatch
Internet images

ABSTRACT

In recent years, various steganalysis algorithms have been proposed and achieved satisfactory performance. However, these conventional methods are not effective for mismatched steganalysis. In real world, there are millions of images captured by different cameras and users transmitted on the Internet every day. The steganalysis on Internet images will encounter steganographic algorithm mismatch (SAM) and cover source mismatch (CSM). Therefore, the steganalysis on the Internet is essentially to solve the mismatch problem. This paper proposes a method to solve the mismatched steganalysis on the Internet images by domain adaptation classifier. It makes the distribution between training and testing sets more similar to obtain better detection performance. We integrate joint distribution adaptation and geometric structure as regularization terms to a standard supervised classifier. Specifically, joint distribution adaptation contains marginal and conditional distributions. And considering the characteristics of steganalysis on the Internet images, we add the conditional regularization in the geometric structure to the existing algorithms. Experimental results (include SAM and CSM) on Internet images show that our method has a better performance than state-of-the-art methods.

1. Introduction

Steganography is a technology that embeds secret messages in digital media and transmits it in open channels [1,2]. No one knows the existence of secret information except the sender and the receiver. In contrast to steganography, steganalysis aims to detect the presence of secret messages in digital media [3,4]. These two technologies are moving forward together. Conventional steganalysis methods include two steps: extracting features and training classifier. In recent years, many steganalysis features have achieved good performance in laboratory environments, such as PEV274 [5], DCTR [6], JRM [7]. These recently proposed steganalysis features trend to become more complex and have higher dimensions. In order to obtain lower computational complexity with the high-dimensional features, the applications of ensemble classifier [8] are becoming more and more popular. However, the good performances are obtained on matched steganalysis under laboratory conditions.

The rapid development of Internet has greatly enriched people's material and spiritual life, and billions of people have personal computers or smart phones. Images are widely used because they contain large amounts of information. In our lives, everyone can upload and download images at anytime and anywhere by using their smart devices. In addition, there are millions of images are posted on the Internet by different users every day, and their sizes, imaging equipments, qualities, contents are quite different. For instance, social media networks flickr and instagram have millions of users sharing images. Furthermore, the development of personal cameras, cell phones and image modification software is convenient for people's daily life, but criminals can also use these tools to transfer secret messages more simply. In addition, many news media have reported that criminals employ steganography in many terrorist attacks and crimes [9], such as September 11 attacks, islamic terrorist organization. Therefore, steganalysis on the Internet images is essential. However, when we move conventional steganalysis algorithms into real-world, the different distributions of training and testing sets can cause significant performance degradation [10]. We call the phenomenon mismatch [11–16].

Specifically, the problem of mismatched steganalysis is caused by the differences between the training and testing sets. In general, the differences occur mainly in the two processes of generating cover and stego images [11]. Cover source of images includes the imaging equipment, size, quality factor, compression history, upload or download record and so on. The Internet images contain all the cover sources mentioned above. Similarly, in the process

[*] Corresponding author.
E-mail addresses: kongxw@dlut.edu.cn, kongxiangwei@zju.edu.cn (X. Kong).

https://doi.org/10.1016/j.neucom.2019.04.025
0925-2312/© 2019 Published by Elsevier B.V.

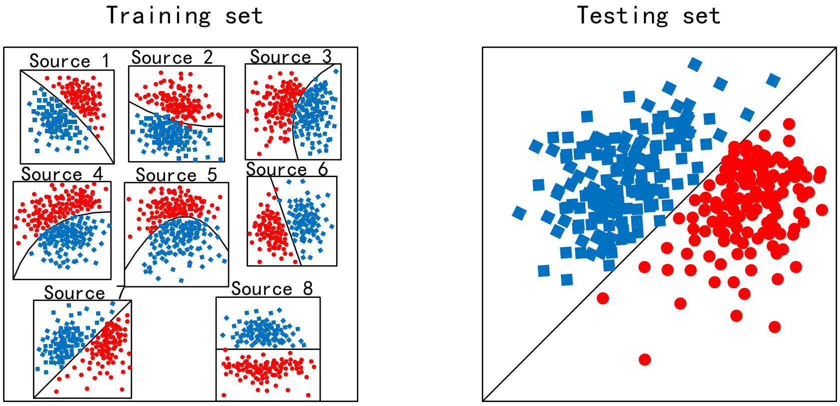


Fig. 1. Expanding diversity of training set. There are many types of distribution in training set, and always one can match testing set.

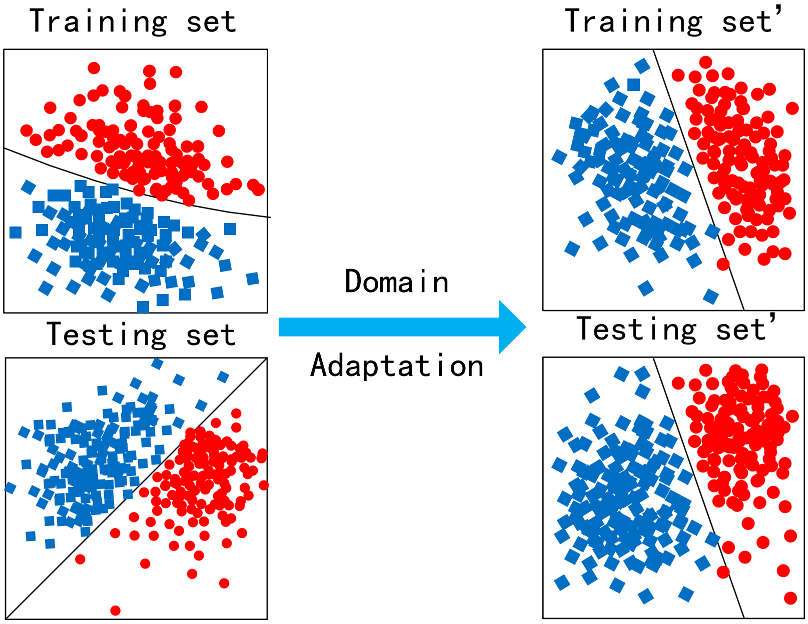


Fig. 2. The different distributions of original features in training and testing sets can be similar by domain adaptation.

of generating stego images, steganographic algorithm and payload size also lead to mismatch. The main purpose of this paper is to propose a method for tackling SAM (steganographic algorithm mismatch) and CSM (cover source mismatch) on the Internet images.

Recently, the literatures have shown an increasing interest in the mismatched steganalysis. The existing mismatch steganalysis methods can be roughly divided into two categories: supervised and unsupervised. The main issue of supervised method is how to make the classification model obtained in training set performs well in a different testing set. Specifically, there are two strategies for supervised methods. One is expanding the diversity [11,12,17] of training set to make it applicable to more testing sets, the otherwhich is known as domain adaptation [13–15] reduces difference of distributions between training and testing sets. The existing unsupervised methods are outlier detection [16,18,19], which treat the stego images and steganographers as rare outliers.

(1) Expanding diversity: Fig. 1. indicates that it matches the testing set by expending the diversity of training set. Ker et al. [17] proposed a mishmash method for mitigating the model mismatch mess. Fridrich et al. [11] proposed two algorithms for cover source mismatch, Mixture and Closest. Mixture extends image sources of training set, and Closest chooses the training set from many pre-training sets, which is the most similar to the testing set. Xu et al. [12] proposed large representative training set for intraclass variation.

(2) Domain adaptation: Transfer learning is an effective technology that uses rich labeled data in source domain to obtain a classifier for the target domain where labeled data is sparse. It has been widely used in image classification, object recognition, cross-domain recommendation and etc. [20–26]. It is a coincidence that the application scenario of cross domain transfer learning is similar to mismatched steganalysis. Kong et al. [13–15] introduced transfer learning to mismatch steganalysis by sharing feature representation between training and testing sets. Daniel et al. [27] sought a latent space by using manifold alignment. As shown in the Fig. 2., domain adaptation transforms the original steganalysis feature, such as PEV-274, to a new feature representation, as a result the distributions of training and testing sets become similar.

(3) Outlier detection: This strategy is accompanied by moving steganalysis into real-world, because there are few people who know how to use the steganography technology on the Internet. Fig. 3. shows that stego image and steganographer deviate from cover images and normal users. Ker et al. [18] proposed a new steganalysis paradigm that steganographer (social media network user) is the outlier. They adopted local outlier factor (LOF) [28] to

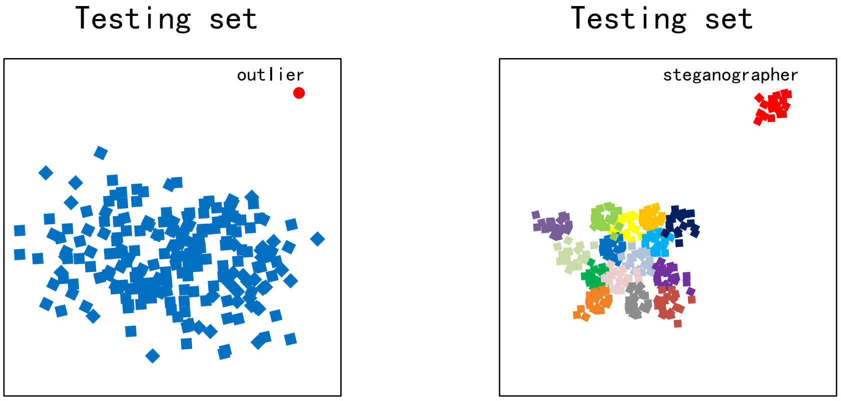


Fig. 3. Outlier detection is an unsupervised strategy, where the stego image and steganographer are treated as outliers, and mismatch is non-existent.

evaluate the degree of actors deviation from the majority. The user who is the farthest from the majority is considered as the steganographer. Li et al. [19] also used unsupervised clustering ensembles to find steganographer. The advantage of this strategy is that it is completely unaffected by mismatch because it does not require training set.

All of these methods alleviate the influence of mismatched steganalysis on detection performance degradation to some extent. However, these strategies still have limitations. First, we cannot indefinitely extend the diversity of training set to make it contain all types of the testing sets. Moreover unsupervised methods can achieve great performance only when the steganographic embedding rate is high. It is inevitable that cover images may also deviate from the majority. Therefore the method based on outlier detection is unstable.

Our work belongs to domain adaptation, and we propose a method to obtain an adaptive classifier that includes joint distribution adaptation and graph Laplacian regularization. Joint distribution adaptation contains marginal distribution and conditional distribution. We also extend the knowledge of marginal and conditional distribution into graph regularization. In addition, we construct the conditional distribution based graph regularization with the testing label obtained by the adaptive classifier to continuously boost the detection accuracy rate.

The remainder of the paper is organized as follows. We review the related work and summarize contributions of our work in Section 2. In Section 3, we propose a joint distribution based adaptive classifier on Internet images steganalysis. We conduct experiments on real-world datasets and compare with previous methods in Section 4. Finally, we conclude this paper in Section 5.

2. Related work

In this section, we discuss the existing domain adaptation methods and their applications in mismatched steganalysis in detail. Conventional steganalysis consists of extracting features and classifiers. We consider the main reason for the existence of mismatched steganalysis is that there is a huge difference between the distributions of training and testing features. In order to solve this problem, we can extract better features or learn better classifier. The two parts of steganalysis exactly correspond to subspace transfer learning and domain adaptive classifier. Specifically, subspace transfer learning makes a new feature representation, and the better classification results can be obtained in the new feature space. In addition, domain adaptive classifier improves detection accuracy by constructing an adaptive classifier.

Transfer learning aims to learn an effective classifier by sufficient labeled source data for unlabeled target data. It is an important issue how to reduce the difference of distributions between source and target data and preserve the properties of original data simultaneously [29]. According to Pan and Yang [30], we roughly separate existing methods into two categories: subspace transfer learning and domain adaptive classifier.

(1) Subspace transfer learning: These methods sought a latent sharing subspace to reduce the distribution difference between both domains. Some methods introduced Principal Component Analysis (PCA) to reconstruct feature representation of original data. In order to reduce the distribution difference, the first thing is to measure the distance of distributions in different domains properly, e.g., Maximum Mean Discrepancy [20,29,31] and Bregman divergence [32]. However, it is not enough to obtain an effective feature representation. Therefore, the great performance of subspace also needs to preserve the important properties of original data.

(2) Domain adaptive classifier: These methods aim to construct a domain adaptive classifier [21–25] by integrating the domain adaptation principles as regularization terms directly. Cao et al. [2] proposed to implement multiple kernel learning for domain adaptation. Similar to subspace transfer learning, the principles of domain adaptation also include reducing the distribution difference and preserving latent properties and so on.

Most of the existing transfer learning approaches in steganalysis belong to subspace transfer learning. Feng et al. [15] aligned mean and variance to reduce statistical differences of features. Zeng et al. [13] considered preserving properties of training set and reducing distribution difference. Long et al. [21] proposed a general framework ARTL (Adaptation Regularization Transfer Learning) that learns an adaptive classifier by simultaneously minimizing structural risk, joint distribution adaptation and geometric structures of marginal distribution.

Our work belongs to domain adaptive classifier. Specifically, we improved the method ARTL (Adaptation Regularization Transfer Learning) by considering the conditional distribution in manifold consistency learning. The contributions of this paper are as follows:

(a) We aim to steganalysis for Internet images [10,18,19,33], which is more practical and necessary.

(b) Similar to ARTL [21], we learn an adaptive classifier to detect the stegos. The joint distribution adaptation includes marginal and conditional distribution. But the ARTL only preserves the marginal geometric structure in graph Laplacian, so we add the conditional distribution in graph Laplacian regularization.

(c) Experimental results include steganographic algorithms mismatch and cover source mismatch on Internet images. Our proposed method has better performance than state-of-the-art methods.

Table 1
Notations and descriptions used in this paper.

Notation	Description	Notation	Description
n_s, n_t	#samples of X_s, X_t	$\mathbf{X}$	Data matrix
d	#feature dimension	$\mathbf{y}$	Label vector
C	#classes	$\mathbf{K}$	Kernel matrix
λ	Shrinkage parameter	$\boldsymbol{\alpha}$	Classifier parameters
β	MMD parameter	$\mathbf{M}$	MMD matrix
γ	Graph parameter	$\mathbf{L}$	Graph Laplacian matrix

3. Proposed method

In this section, we will introduce Joint Distribution based Adaptive Classifier (JDAC) in detail. Firstly, we define the problem and summarize the frequently used notation in Table 1. Given the training set $\mathbf{X}_s = [\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, \ldots, \mathbf{x}_{s_{n_s}}]^T \in \mathbb{R}^{n_s \times d}$, which is composed of n_s samples and d-dimensional features. $y_s \in R^{n_s \times 1}$ is the label vector of training set and $y_{s_i} \in \{-1, 1\}$ represent the cover and stego respectively. Similarly, the testing feature set is $\mathbf{X}_t = [\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \ldots, \mathbf{x}_{t_{n_s}}]^T \in \mathbb{R}^{n_t \times d}$ and the testing label vector $y_t \in R^{n_t \times 1}$ is unknown. The task of this paper is to use sufficient labeled training data to learn an effective adaptive classifier for unlabeled testing data.

JDAC integrates joint distribution adaptation and geometric structures as regularization terms to a standard classifier. Similar to many literatures, Maximum Mean Discrepancy (MMD) [31] is used to measure the distance between cross domain distributions. The marginal and conditional distribution adaptations are constructed by utilizing the true labels of training data and the pseudo labels of testing data effectively. In addition, we approve the manifold assumption in cross domain task: if two points $\mathbf{x_1}, \mathbf{x_2}$ are close in the intrinsic geometric structure of marginal distributions P_s and P_t, then the conditional distributions $P(y|\mathbf{x}_1.)$ and $P(y|\mathbf{x}_2.)$ will be also similar. Furthermore, we develop the initial manifold assumption by considering the geometric structure distance of two points in conditional distribution. In other words, we reasonably use the pseudo labels of testing data when constructing the conditional graph Laplacian regularization. Therefore, the objective function contains three parts: *Prediction Losses, Joint Distribution Adaptation* and *Graph Laplacian Regularization*. It is worth noting that graph Laplacian regularization is based on joint distribution.

3.1. Prediction losses

The goal is to obtain an adaptive classifier $f(\mathbf{x})$ that can effectively classify testing data by using the labeled training data. Therefore, we firstly introduce the standard classifier prediction function

$$f(\mathbf{x}) = \boldsymbol{\alpha}^T \phi(\mathbf{x}) \tag{1}$$

where $\boldsymbol{\alpha}$ is $n_s \times 1$ classifier parameters vector, and $\phi(\cdot)$ is kernel-induced mapping feature to *Reproducing Kernel Hilbert Space* (RKHS). According to structural risk minimization [23,34], the squared loss function is adopted. Then, the prediction losses function in training data can be represented as:

$$\sum_{i=1}^{n_s} (f(\mathbf{x}_{s_i}) - y_{s_i})^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{2}$$

where $\|\cdot\|_{\mathcal{H}}^2$ is the ℓ_2 norm in RKHS, and λ is shrinkage parameter. Then the classifier function (1) is taken into formula (2). And considering that the following formulas involve both the training and testing data, for expressing conveniently, we extend the source data to entire data in Eq. (2) by introducing coefficient matrix $\mathbf{E}$. Specifically, $\mathbf{E} \in R^{n \times n}$ is a diagonal matrix that ignores the unknown testing labels with the element $\mathbf{E}_{ii} = 1$ when $\mathbf{x}_i \in \mathbf{X}_s$, and $\mathbf{E}_{ii} = 0$ otherwise. So (2) can be rewritten as:

$$\sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 \mathbf{E}_{ii} + \lambda \|f\|_{\mathcal{H}}^2$$
$$= \left\| \left(\mathbf{y}^T - \boldsymbol{\alpha}^T \mathbf{K}\right)\mathbf{E} \right\|_F^2 + \lambda tr\left(\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}\right) \tag{3}$$

where $n = n_s + n_t$ is number of training and testing samples, and label vector $\mathbf{y} = [y_1, y_2, \ldots, y_{n_s+n_t}]^T \in R^{n \times 1}$ includes labels of the entire data. It does not matter what the label of testing data is in (3), because the coefficient matrix can filter it. We map features matrix $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t]^T$ into RKHS and construct kernel matrix $\mathbf{K} = \Phi(\mathbf{X})\Phi(\mathbf{X})^T \in R^{n \times n}$.

Noting that JDAC is constructed only as a standard classifier on labeled training data so far. In order to make $f(\mathbf{x})$ more adaptive and effective on testing data, joint distribution adaptation and graph Laplacian regularization will be introduced.

3.2. Joint distribution adaptation

In the previous subsection, a standard classifier is obtained, but it does not necessarily achieve great performance on testing data. So the next step is to integrate the knowledge of transfer learning to previous predicted function. According to Long et al. [20], we adopt joint distribution adaptation that contains marginal distribution and conditional distribution.

Marginal distribution adaptation: It is difficult to directly describe the marginal distribution of two data sets. Similar to [14,20,29], we adopt a basic assumption of Maximum Mean Discrepancy (MMD): if the mean values of two distributions with sufficient samples are equal, the two distributions are similar. Then the distance between training and testing data is measured as:

$$MMD_0{}^2 = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(\mathbf{x}_i) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2 \tag{4}$$

where $\phi(\cdot)$ is kernel-induce mapping feature to RKHS and $\|\cdot\|_{\mathcal{H}}^2$ is the ℓ_2 norm in RKHS. In order to combine prediction function f with marginal distribution adaptation, the Eq. (1) is incorporated into formula (4) and obtain *predicted MMD*$_0$:

$$PMMD_0{}^2 = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} f(\mathbf{x}_i) - \frac{1}{n_t} \sum_{j=1}^{n_t} f\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2$$
$$= tr\left(\boldsymbol{\alpha}^T \mathbf{K} \mathbf{M}_0 \mathbf{K} \boldsymbol{\alpha}\right) \tag{5}$$

where $tr(\cdot)$ is trace and $\mathbf{K} = \Phi(\mathbf{X})\Phi(\mathbf{X})^T$ is kernel matrix. $\mathbf{M}_0$ is coefficient matrix that can be computed as:

$$(\mathbf{M}_0)_{ij} = \begin{cases} \frac{1}{n_s n_s}, & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_s \\ \frac{1}{n_t n_t}, & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t \\ \frac{-1}{n_s n_t}, & otherwise, \end{cases} \tag{6}$$

Minimize Eq. (5) to reduce marginal distributions differences between training and testing sets.

Conditional distribution adaptation: The conditional distribution cannot be directly represented, since there are no labeled data in testing set. In this paper, we adopt a standard classifier to obtain pseudo labels of testing samples. Then conditional distributions of difference sets are evaluated with true training labels and pseudo testing labels. Similar to MMD_0, the class-conditional distribution is defined as

$$MMD_c{}^2 = \left\| \frac{1}{n_s^c} \sum_{i=1}^{n_s^c} \phi(\mathbf{x}_i) - \frac{1}{n_t^c} \sum_{j=1}^{n_t^c} \phi\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2 \tag{7}$$

where n_s^c and n_t^c are the number of training and testing samples belonging to class c. Since steganalysis is binary classification problems, $c \in \{-1, 1\}$ respectively represent cover and stego images.

Following to Eq. (5), the classifier is incorporated into (7) and the predicted conditional distribution *MMD* can be computed as:

$$PMMD_c{}^2 = \left\| \frac{1}{n_s^c} \sum_{i=1}^{n_s^c} f(\mathbf{x}_i) - \frac{1}{n_t^c} \sum_{j=1}^{n_t^c} f\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2 = tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{M}_c\mathbf{K}\boldsymbol{\alpha}\right) \tag{8}$$

where $\mathbf{M}_c$ is coefficient matrix and can be represented as:

$$(\mathbf{M}_c)_{ij} = \begin{cases} \frac{1}{n_s^c \cdot n_s^c}, & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_s^c \\ \frac{1}{n_t^c \cdot n_t^c}, & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t^c \\ \frac{-1}{n_s^c \cdot n_t^c}, & \begin{cases} \mathbf{x}_i \in \mathbf{X}_s^c, \mathbf{x}_j \in \mathbf{X}_t^c \\ \mathbf{x}_i \in \mathbf{X}_t^c, \mathbf{x}_j \in \mathbf{X}_s^c \end{cases} \\ 0, & otherwise, \end{cases} \tag{9}$$

The conditional distributions difference of training and testing sets can be reduced by minimizing Eq. (8).

This paper combines Eqs. (5) and (8) to reduce difference in marginal and conditional distributions between training and testing set. So joint distribution adaptation can be computed as

$$tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{M}_0\mathbf{K}\boldsymbol{\alpha}\right) + tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{M}_c\mathbf{K}\boldsymbol{\alpha}\right) = tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{M}\mathbf{K}\boldsymbol{\alpha}\right) \tag{10}$$

where $\mathbf{M} = \mathbf{M}_0 + \mathbf{M}_c$ is joint distributions matrix, and $\mathbf{K} = \Phi(\mathbf{X})\Phi(\mathbf{X})^T$ is kernel matrix.

3.3. Graph Laplacian regularization

As a supplement to joint distribution adaptation, our work incorporate graph Laplacian regularization of joint distributions in both domains to predict loss function. The geometric structures of distributions imply label information of testing data. According to manifold assumption, the geometric structures of marginal distributions P_s, P_t between two points $\mathbf{x}_1$, $\mathbf{x}_2$ can approximately reflect conditional distributions $P(y|\mathbf{x}_1.)$ and $P(y|\mathbf{x}_2.)$. One of our important contributions in this paper is that we develop manifold assumption from marginal distribution to joint distribution. The graph Laplacian regularization contains marginal and conditional geometric structures.

Similar to the previous subsection, since the testing set has non-labeled samples, marginal graph Laplacian regularization is computed as:

$$\|f\|_{G_0}^2 \triangleq tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}_0\mathbf{K}\boldsymbol{\alpha}\right) \tag{11}$$

where G_0 is affinity graph with marginal distribution in training and testing sets, and $\mathbf{L}_0$ is the normalized graph Laplacian matrix. $\mathbf{W}_0$ is graph affinity matrix [23] which can be computed as:

$$(\mathbf{W}_0)_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_s \; or \; \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t \\ 0, & otherwise, \end{cases} \tag{12}$$

where σ is bandwidth parameter, and the elements are calculated by Gaussian function.

Then the graph Laplacian regularization of conditional distribution is designed by utilizing the pseudo labels of testing data. And the pseudo testing labels can be obtained by a standard classifier on training data. The conditional graph Laplacian regularization is computed as:

$$\|f\|_{G_c}^2 \triangleq tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}_c\mathbf{K}\boldsymbol{\alpha}\right) \tag{13}$$

where $c \in \{-1, 1\}$ respectively represent cover and stego images. G_c is conditional affinity graph and it associates with information of labels in both domains. Similar to Eq. (12), the conditional graph affinity matrix can be computed as:

$$(\mathbf{W}_c)_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}}, & \begin{cases} \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_s^c \\ \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t^c \\ \mathbf{x}_i \in \mathbf{X}_s^c, \mathbf{x}_j \in \mathbf{X}_t^c \\ \mathbf{x}_i \in \mathbf{X}_t^c, \mathbf{x}_j \in \mathbf{X}_s^c \end{cases} \\ 0, & otherwise, \end{cases} \tag{14}$$

This formula means that the edge weights of two samples are calculated only when the labels of them are identical.

Combining Eqs. (11) and (13), joint distribution based graph Laplacian regularization approximates the manifold in training and testing distribution. And it can be rewritten as:

$$tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}_0\mathbf{K}\boldsymbol{\alpha}\right) + tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}_c\mathbf{K}\boldsymbol{\alpha}\right) = tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha}\right) \tag{15}$$

where $\mathbf{L}_0$ and $\mathbf{L}_c$ are the normalized graph Laplacian matrix with marginal distribution and conditional distribution, respectively. Furthermore, they can be computed as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is diagonal matrix with each elements $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{W}_{ij}$.

3.4. Optimization algorithms

Finally, the joint distribution based adaptive classifier is obtained by incorporating prediction losses, joint distribution and graph Laplacian regularization. Combining Eqs. (3), (10) and (15), the final objective function is as follows:

$$\begin{aligned} &\arg\min_{\boldsymbol{\alpha}} \left\| \left(\mathbf{y}^T - \boldsymbol{\alpha}^T \mathbf{K}\right)\mathbf{E} \right\|_F^2 + \lambda tr\left(\boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha}\right) \\ &\quad + \beta tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{M}\mathbf{K}\boldsymbol{\alpha}\right) + \gamma tr\left(\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha}\right) \end{aligned} \tag{16}$$

where λ, β and γ are positive regularization parameters, respectively, shrinkage parameter, MMD parameter and graph Laplacian parameter. The pseudo labels are easily obtained by a standard classifier, e.g., Support Vector Machine (SVM). However it also can be acquired by an adaptive classifier, modifying formula (16) to marginal distribution based adaptive classifier. And setting the derivative of Eq. (16) as 0, the marginal distribution based adaptive classifier parameters are as follows:

$$\boldsymbol{\alpha}_0 = (\lambda\mathbf{I} + (\mathbf{E} + \beta\mathbf{M}_0 + \gamma\mathbf{L}_0)\mathbf{K})^{-1}\mathbf{E}\mathbf{y} \tag{17}$$

where $\mathbf{M}_0$, $\mathbf{L}_0$ are MMD parameter and graph parameter of marginal distributions. By substituting Eq. (17) into prediction function (1), a simple adaptive classifier is learned to obtain first-time pseudo labels of testing data. It is worth noting that the outputs, pseudo labels of testing data can be used as inputs to boost the classifier more accurate. Then the derivative of formula (16) is set as 0, the joint distribution based adaptive classifier parameter can be computed as:

$$\boldsymbol{\alpha} = (\lambda\mathbf{I} + (\mathbf{E} + \beta\mathbf{M} + \gamma\mathbf{L})\mathbf{K})^{-1}\mathbf{E}\mathbf{y} \tag{18}$$

The final classifier can be obtained by incorporating Eq. (18) into (1). This paper iterates the pseudo testing labels to make it steady. The overall algorithm of this paper is given in Algorithm 1.

4. Experiments

In this section, we conduct extensive experiments on two real-world data sets, MIRFlickr 1M [35], Amazon and WebVision [36] (Fig. 4.), to evaluate the performance of our proposed JDAC. Our work aims to moving steganalysis from laboratory into real world, so image dataset are chosen from the Internet, such as, image sharing and shopping websites. This paper addresses the problems of steganalysis with steganographic algorithm mismatch (SAM) and cover source mismatch (CSM).

Fig. 4. The experimental dataset, MIRFlickr 1M, Amazon and WebVision.

Algorithm 1 JDAC: joint distribution based adaptive classifier.

```
Input:
    Training set $\mathbf{X}_s$, testing set $\mathbf{X}_t$, training set labels $\mathbf{y}_s$; $\lambda$, $\beta$, $\gamma$; $k = 0$, $MaxIters$
Output:  Adaptive Classifier $f$
Compute kernel matrix $\mathbf{K}$ by $\mathbf{K}=\Phi(\mathbf{X})\Phi(\mathbf{X})^T$ with kernel function $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$;
Construct MMD matrix $\mathbf{M}_0$ by Equation (5) and graph Laplacian matrix $\mathbf{L}_0$ by Equation (11);
Compute adaptive classifier parameters $\alpha_0$ by Equation (18) and obtain pseudo labels of testing data;
repeat
    $k = k + 1$;
    Construct MMD matrix $\mathbf{M}$ by Equation (10) and graph Laplacian matrix $\mathbf{L}$ by Equation (15);
    Compute adaptive classifier $f$ by Equation (17) and obtain pseudo labels of testing data;
until $\alpha$ is converged or $k > MaxIters$
return  Adaptive Classifier $f$.
```

4.1. Data set

MirFlickr 1M [35] contains one million images which are collected from YAHOO's Flickr that is an image social network website. The MIR Flickr collect high-quality photographic images from thousands of Flickr users, made available under the Creative Commons license. The dataset has a wide variety of categories which contains landscapes, characters, animals, buildings, objects and so on. It also provides information of image tags, licenses and EXIF. And most of the image quality factors (QF) in MirFlickr dataset are 96, but there are also some other quality factors in the image. The image sizes in this database are less than 500 × 500, we consider that it may be because the constructor of the dataset may resize the images when they collected these images. More detailed introduction and download of MirFlickr can be found on http://press.liacs.nl/mirflickr/.

In addition, we obtain 5000 images of the goods by downloading on shopping website (https://www.amazon.cn). We crawled the images from the mentioned website. The dataset contains many types of goods, including food, clothes, books, medicines, electronic products and so on. But in order to collect pictures quickly and conveniently, when we download images from the website, we set the quality factor of the images to 85, and the size of the images is limited to less than 500 × 500.

In order to verify the robustness of our method for more different image sizes, we add a new WebVision dataset [36]. The WebVision dataset contains more than 2.4 million of images crawled from the Flickr website and Google Images search. Images in this database are naturally collected from two domains, Flickr and Google, so we implemented the experiment with cover source mismatch on this dataset. The quality factors of images from WebVision are very complex, the largest is 100 and the smallest is 20. And it contains a wide variety of image sizes, the largest of which is 4752 × 3168 and the smallest is 80 × 80. More detailed introduction and download of WebVision can be found on http://www.vision.ee.ethz.ch/webvision/2017/index.html.

Table 2
Detection accuracy (%) on MIRFlickr 1M with 20% embedding payloads.

	SVM	GTCA	RDFT	ARTL	JDAC_marginal	JDAC
F5 vs OutGuess	56.4	62.0	66.9	76.3	82.8	**83.4**
F5 vs MBS	77.8	90.8	83.6	88.6	90.5	**91.9**
F5 vs Jsteg	68.4	**97.8**	92.9	93.1	92.0	93.5
F5 vs nsF5	75.3	72.9	**82.2**	67.9	74.7	71.4
OutGuess vs F5	61.4	75.0	65.4	87.1	86.7	**88.1**
OutGuess vs MBS	91.2	**95.3**	92.3	93.0	90.0	91.3
OutGuess vs Jsteg	93.9	**98.0**	96.6	88.7	85.3	86.3
OutGuess vs nsF5	55.1	54.8	60.0	70.4	74.4	**74.9**
MBS vs F5	59.3	76.2	67.4	90.7	92.6	**93.9**
MBS vs OutGuess	70.6	80.2	79.4	83.7	87.3	**87.9**
MBS vs Jsteg	89.6	**98.2**	95.1	96.9	95.5	96.4
MBS vs nsF5	53.9	58.1	61.3	66.9	**74.7**	72.6
Jsteg vs F5	50.1	69.3	50.0	76.1	91.4	**93.6**
Jsteg vs OutGuess	50.3	64.3	50.0	71.1	**83.3**	**83.3**
Jsteg vs MBS	50.7	82.0	50.0	92.8	94.7	**95.7**
Jsteg vs nsF5	50.1	54.2	50.0	61.0	**70.9**	65.2
nsF5 vs F5	**94.8**	92.2	**94.8**	85.3	87.9	88.7
nsF5 vs OutGuess	67.3	69.2	69.7	75.9	**76.7**	75.6
nsF5 vs MBS	85.5	**93.3**	83.5	78.9	80.4	80.0
nsF5 vs Jsteg	89.5	**96.3**	90.4	80.0	78.6	79.3
Average	69.56	79.01	74.08	81.22	84.52	**84.65**

4.2. Experimental settings

A total of 1300 images are respectively randomly selected from 4 datasets as experimental images, which are MIRFlickr 1M, Amazon datasets and webvision' s flickr and webvision' s google. In fact, it cannot know whether these images from the Internet have some images that have been hidden secret messages. There is an important assumption that the images are all covers. This assumption is difficult to prove, but we all know that there are very few people who know how to use steganography technology, so stego image is negligible. The same assumption has been adopted in previous work on the Internet image steganalysis [19,33]. In order to generate stego images, we adopt the commonly used steganographic algorithms that are F5 [37], nsF5 [38], MBS [39], OutGuess [40] and Jsteg [41], respectively. Moreover, we set the payload to 20% of maximum embedding capacity. These 1000 images are randomly divided into training and testing sets and both sets contain cover and stego images.

It is worth noting that our work does not try to detect stego images which are generated by J-UNIWARD [42] algorithm. It is sure that matched is better than mismatched steganalysis. But the performance of J-UNIWARD steganalysis is terrible under match condition. Therefore we consider that how to improve the detection accuracy of J-UNIWARD in matched condition is more important. And there is no point in conducting experiments on J-UNIWARD to prove the effectiveness of our proposed JDAC.

In addition, PEV-274 is adopted as steganalysis feature, because its dimension is lower than JRM and the calculation is simpler. PEV-274 consists of 193-dimensional DCT features and 81-dimensional average calibrated Markov features. The classifier is our proposed JDAC, which involves three trade-off parameters, namely, shrinkage parameter, MMD parameter and graph Laplacian parameter. In the experiment, we have been set up a verification set for each dataset, which contains 300 images different from the training set and the testing set. And we optimized the parameters of our proposed JDAC and the other 4 compared methods, according to the results of verification set. According to parameter sensitivity analysis experiment, we set the parameters $\lambda = 1$, $\beta = 100$, $\gamma = 1$, with the highest detection accuracy.

We consider that only marginal distribution is insufficient to describe distribution differences between source and target domains. Therefore, JDAC adopts the joint distribution adaptation and also uses the graph Laplacian regularization based on joint distributions. In order to verify the validity of conditional distribution, we added a compared method (JDAC_marginal) which adopts only marginal distribution. JDAC and JDAC_marginal are both our methods. Our method is compared with several state-of-the-art algorithms:

SVM [43]: Support Vector Machine is a baseline classifier. We train it on labeled training data with original PEV-274 feature.

GTCA+SVM [14]: Generalized Transfer Component Analysis is an improvement version on TCA by introducing the alignment of mean and variance between training and testing data. GTCA adopts SVM as classifier, since it calculates conditional adaptation with the posterior probability obtained by SVM.

RDFT+SVM [13]: Robust Discriminative Feature Transformation is a method of subspace transfer learning. This method constructs a latent feature representation by reducing the distribution difference and keeping the classification ability of training set. Following [13], SVM is trained on feature subspace of labeled training data.

ARTL [21]: Adaptation Regularization Transfer Learning belongs to constructing an adaptive classifier.

4.3. Experimental results

The experiments evaluate the performance of our proposed JDAC_marginal, JDAC and the other four algorithms by showing average total accuracy P_A under equal priors achieved on testing set, $P_A = \frac{1}{2}ave(p_c + p_s)$, where p_c is the classification accuracy rate of cover images, and p_s is the detection accuracy rate of stego images. In order to ensure the effectiveness of our algorithm, we repeated all experiment results 5 times and took the average as final results.

The experiments of mismatched steganalysis are composed by two parts, which are Steganographic Algorithms Mismatch (SAM) and Cover Source Mismatch (CSM), respectively.

4.3.1. Steganographic algorithms mismatch (SAM)

In this experiment, 1300 images are randomly selected in MIRFlickr 1M as the training, verification and testing sets. SAM means that the methods to generate stego images of training and testing sets are different. The experiment is a mutual detection of 5 steganographic algorithms, namely, F5, OutGuess, MBS, Jsteg, nsF5. The detection accuracies of our proposed JDAC and 4 compared algorithms on SAM are shown in Table 2, where *F5 vs OutGuess* means that the algorithm of generating stego images in training set is *F5* and in testing set is *OutGuess*.

Table 3
Detection accuracy (%) of CSM, between MIRFlickr 1M and Amazon with 20% embedding payloads.

		SVM	GTAC	RDFT	ARTL	JDAC_marginal	JDAC
	F5	75.4	73.7	79.4	87.9	82.0	**94.7**
MRIFlickr 1M	*OutGuess*	86.9	**94.8**	90.9	84.5	81.3	86.9
vs	*MBS*	88.3	88.0	**88.8**	71.9	81.2	87.8
Amazon	*Jsteg*	98.8	**99.7**	99.1	98.1	90.6	99.4
	nsF5	65.1	58.4	66.1	63.5	**66.8**	66.3
	F5	65.0	81.1	74.7	62.1	86.3	**91.2**
Amazon	*OutGuess*	52.3	69.7	64.6	59.0	77.7	**80.2**
vs	*MBS*	83.6	79.7	88.0	70.2	84.0	**94.3**
MRIFlickr 1M	*Jsteg*	67.5	96.1	92.8	98.1	86.9	**98.3**
	nsF5	68.8	65.7	71.9	75.6	77.9	**78.3**
Average		75.17	80.69	81.63	77.09	81.47	**87.74**

Table 4
Detection accuracy (%) of CSM, between WebVision's Flickr and WebVision's Google with 20% embedding payloads.

		SVM	GTAC	RDFT	ARTL	JDAC_marginal	JDAC
	F5	73.2	61.6	71.7	**76.4**	72.6	75.0
WebVision's Flickr	*OutGuess*	81.2	69.9	79.4	85.3	85.0	**87.4**
vs	*MBS*	83.8	70.9	84.3	85.7	87.0	**88.2**
WebVision's Google	*Jsteg*	93.7	85.2	88.4	91.1	95.3	**95.5**
	nsF5	62.1	54.5	61.9	63.5	61.3	**64.0**
	F5	69.0	53.1	68.5	70.9	72.8	**75.4**
WebVision's Google	*OutGuess*	89.6	77.5	88.2	88.5	90.9	**92.2**
vs	*MBS*	89.0	80.8	89.9	87.4	92.0	**92.5**
WebVision's Flickr	*Jsteg*	95.7	92.2	93.5	92.7	**95.8**	95.7
	nsF5	42.4	43.6	55.3	63.1	63.2	**65.7**
Average		77.97	68.93	78.11	80.46	81.59	**83.16**

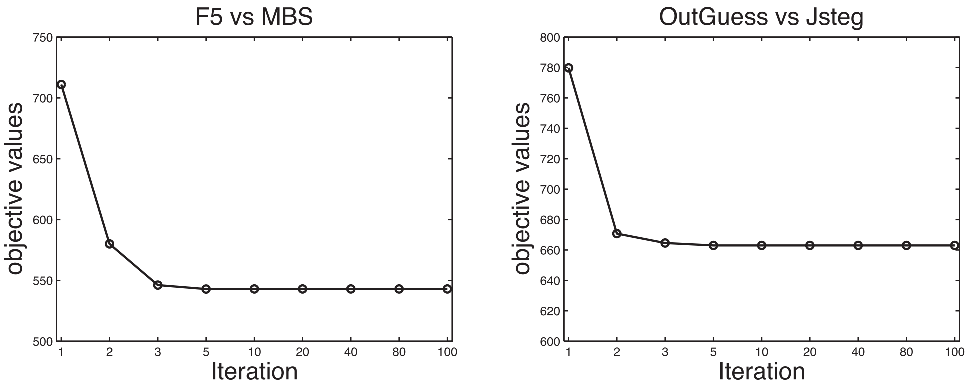


Fig. 5. The convergence process of JDAC on two detection tasks with Steganographic Algorithms Mismatch.

We can see that SVM is the baseline of this experiment, and the average detection accuracy of SVM is only 69.56%. In fact, this method is completely mismatched steganalysis, which means that the classification model learnt from the source domain is directly used for the classification task of the target domain. In most tasks, it will achieve poor detection accuracy. We can also observe that RDFT performs best on 2 tasks, and the average accuracy is about 4.5% higher that baseline. On the *F5 vs nsF5* task, the detection accuracy of RDFT is 10.8% higher than our proposed JDAC. This is reasonable since the constraint term which keeps the classification ability of training data preserves the important property. In these tasks, the detection accuracy of GTCA is over 5% higher than RDFT. And we can see that GTCA performs best on 6 tasks.

Second, it can be observed that the algorithms of domain adaptive classifier are better than subspace transfer learning in these experiments. ARTL is only worse than our proposed JDAC_marginal and JDAC, and the goal of the three methods is to construct an adaptive classifier. And the average detection accuracy of ARTL is 2% higher than GTCA. Our proposed JDAC has an improvement based on ARTL, because we consider the conditional graph Laplacian regularization. JDAC increases the accuracy by about 3 % on average. We can see that the accuracy of JDAC is higher than ARTL on 15/20 tasks, and JDAC performs best on 9 tasks. Compared with JDAC_marginal, JDAC achieves higher detection rate and performs better on 15/20 tasks. This shows that the joint distribution adaptation can effectively improve the detection performance by using pseudo labels to calculate conditional distribution in mismatched steganalysis.

4.3.2. Cover source mismatch (CSM)

Our algorithm is also effective for Cover Source Mismatched steganalysis. We set training and testing sets from different databases, where *MIRFlickr 1M vs Amazon* represents the training images chosen from *MIRFlickr 1M*, and testing from *Amazon*. Moreover, we conduct the experiments on 5 steganographic algorithms and control the methods of constructing the stego images are identical. The detection accuracy rate of CSM is shown in Tables 3 and 4.

It can be seen in Table 3 that the methods of subspace transfer learning, GTCA and RDFT, are better than the baseline SVM by 5 %, and 3.5 % higher than ARTL. In addition, GTCA performs best on 3 tasks. And on the task of *MIRFlickr 1M vs Amazon* with *OutGuess*, GTCA increases the accuracy by about 4 % than other methods. Secondly, the adaptive classifier ARTL is worse than GTCA and RDFT

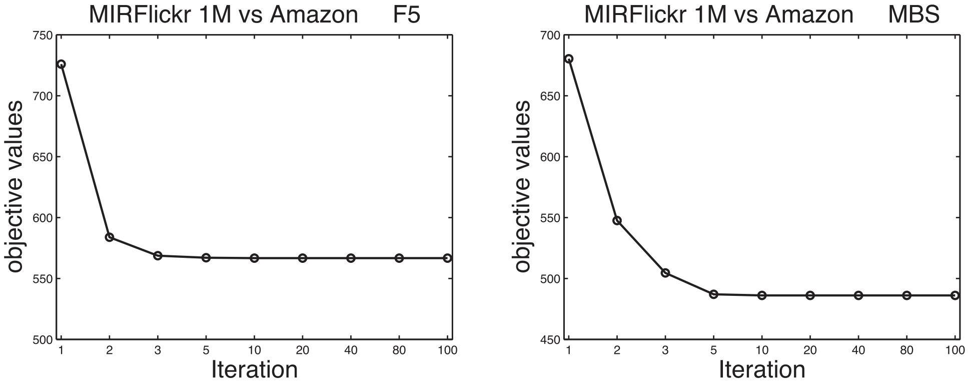
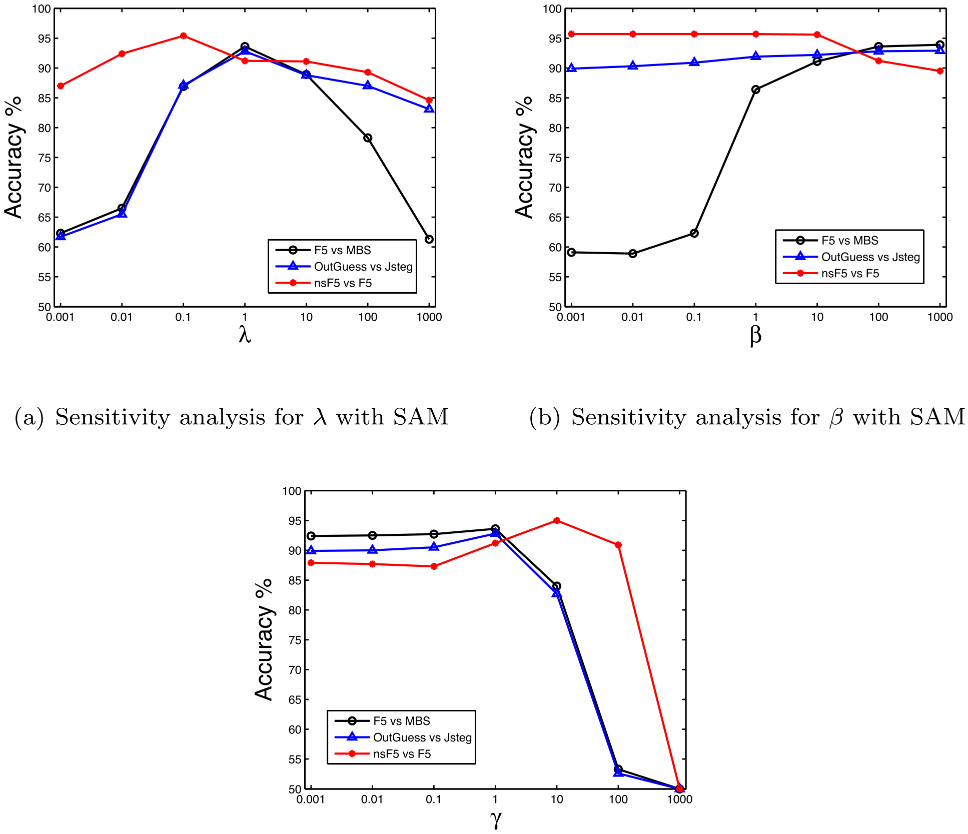


Fig. 6. The convergence process of JDAC on two detection tasks with Cover Source Mismatch.



(a) Sensitivity analysis for λ with SAM

(b) Sensitivity analysis for β with SAM

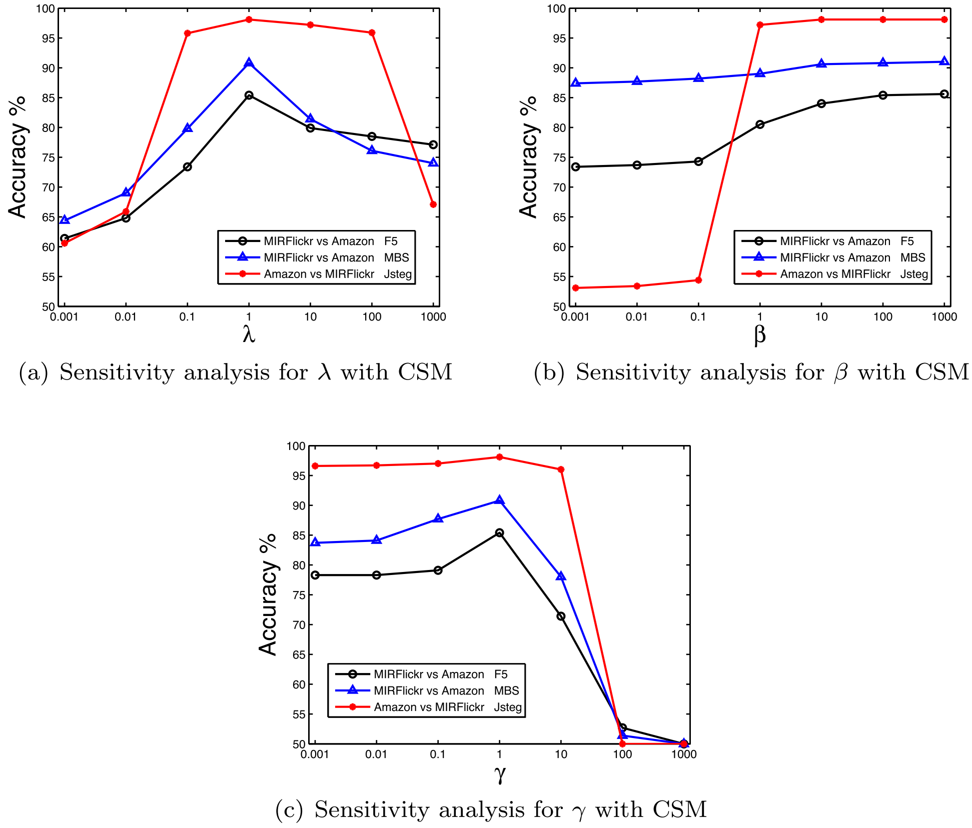
(c) Sensitivity analysis for γ with SAM

Fig. 7. Parameter analysis for JDAC on 3 tasks with SAM.

on average by 5%, and it performs best on 1 task. In the Table 3, it can be seen that the detection accuracy on average of our proposed method JDAC is the highest, and it increases accuracy over 5% higher than other methods. JDAC achieves the best accuracy in 5/10 tasks.

Images in WebVision database are naturally collected from two domains, Flickr and Google, so we implemented the experiment with cover source mismatch on this dataset, and the results are shown in the Table 3. The experiment is a mutual detection between the 2 domains, where *WebVision's Flickr vs WebVision's Google* represents the training images chosen from *WebVision's Flickr*, and testing from *WebVision's Google*. Compared with other methods, it can be seen that JDAC also achieves the best performance on WebVision, and the best detection accuracy in 8/10 tasks. Table 3 shows that the algorithms of domain adaptive classifier are better than subspace transfer learning in this dataset. Our proposed JDAC increases accuracy by 2.7% on average than ARTL. And JDAC_margial achieves the best result on 1/10 task, and is only 0.1% higher than JDAC.

The experimental results in Tables 3 and 4 also show that joint distribution can effectively improve the performance of JDAC with Cover Source Mismatch(CSM). Compared with JDAC_marginal, JDAC increases the average accuracy by about 6% in Table 3, and increases the accuracy by about 1.6% in Table 3. In addition, several values are below 50%. For example, in Task of *WebVision's Flickr vs WebVision's Google* with *nsF5*, we can see that the detection



(a) Sensitivity analysis for λ with CSM

(b) Sensitivity analysis for β with CSM

(c) Sensitivity analysis for γ with CSM

Fig. 8. Parameter analysis for JDAC on 3 tasks with CSM.

accuracy of SVM is 42.4%. These cases occur when the steanography method is nsf5. The main reason is that nsF5 is an improved version of F5 and has higher security. And our experiments are based on cross-domain steganalysis, so a well-trained classification model in the source domain has the possibility that it cannot solve the classification task of the target domain, which leads to the opposite classification results, so that the accuracy is less than 50%.

4.4. Convergence

We illustrate the convergence of our proposed adaptive classifier algorithm JDAC in this subsection and respectively verify the convergence of JDAC on SAM and CSM. The objective function value of Eq. (16) is plotted as the number of iterations increases in Figs. 5 and 6. Due to the limitation of space, we only show 2 tasks to each mismatch. Fig. 5 shows the tasks of *F5 vs MBS* and *OutGuess vs Jsteg*, Fig. 6 shows the tasks of *MIRFlickr 1M vs Amazon F5* and *MBS*. We keep the experimental setting the same as pervious subsection. It can be found that the value of the objective function converges after the 10 iterations.

4.5. Parameter sensitivity

In the experiment, we have set up a verification set for each dataset, which contains 300 images different from the training set and the testing set. And we optimized the parameters of our proposed JDAC_marginal, JDAC and the other 4 compared methods, according to the results of verification set. Our proposed joint distribution based adaptive classifier (JDAC) contains three trade-off parameters. This subsection demonstrates that our proposed JDAC performs well in a large range of parameters. Similar to pervious subsection, the parameters analyses are divided into SAM and CSM. Limited to space, we only illustrate the experiment results on the tasks of *F5 vs MBS*, *OG vs Jsteg* and *nsF5 vs F5* on SAM, and show 3 tasks to analyse parameter with CSM. In addition, the results of parameter sensitivity analysis are similar on other mismatched tasks.

Experiments are conducted with a set of different α values and keep the values of the other two parameters fixed, $\lambda \in [0.001, 1000]$. Each experiment runs 5 times and the average accuracy are shown in Figs. 7 and 8 (a). It shows that JDAC is bit sensitive to shrinkage parameter λ, and the JDAC performs best when $\lambda = 1$.

The MMD parameter β affects the joint distribution adaptation between training and testing sets. Similarly, we set the $\beta \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, and keep the values fixed for the other parameters. From Figs. 7 and 8 (b), we can observe that our proposed JDAC is not sensitive in a wide range to $\beta \in [10, 1000]$. And our method is insensitive to graph Laplacian parameter γ and it achieve a satisfactory performance in a wide range $\gamma \in [0.001, 1]$.

5. Conclusion

In this paper, we proposed a novel joint distribution based adaptation classifier (JDAC) for mismatched steganalysis on Internet images. This paper is an attempt to move steganalysis from laboratory to real-world. That is more practical to tackle the steganalysis on the Internet. And JDAC aims to obtain an effective classifier on sufficient labeled training data for unlabeled testing samples. JDAC integrates the joint distribution adaptation and geometrical structure as a regularization term to a standard classifier. In addition, we design the conditional geometrical structure, which extends the graph Laplacian regularization from marginal to joint distribution. It reduces the distributions difference and preserves the important properties of original data. Extensive experiments show that our proposed method performs better on steganographic algorithm mismatch (SAM) and cover source mismatch (CSM) than several state-of-the-art related methods.

References

[1] J. Wang, J. Ni, X. Zhang, Y.Q. Shi, Rate and distortion optimization for reversible data hiding using multiple histogram shifting, IEEE Trans. Cybern. 47 (2) (2017) 315.

[2] X. Cao, L. Du, X. Wei, D. Meng, X. Guo, High capacity reversible data hiding in encrypted images by patch-level sparse representation, IEEE Trans. Cybern. 46 (5) (2016) 1132–1143.

[3] W. Tang, H. Li, W. Luo, J. Huang, Adaptive steganalysis based on embedding probabilities of pixels, IEEE Trans. Inf. Foren. Secur. 11 (4) (2016) 734–745.

[4] Y. Ren, J. Yang, J. Wang, L. Wang, AMR steganalysis based on second-order difference of pitch delay, IEEE Trans. Inf. Foren. Secur. 12 (6) (2017) 1345–1357.

[5] T. Pevny, J. Fridrich, Merging Markov and DCT features for multi-class JPEG steganalysis, in: SPIE International Society for Optics and Photonics, San Jose, CA, USA, 2007, pp. 650503–650503-13.

[6] V. Holub, J. Fridrich, Low-complexity features for JPEG steganalysis using undecimated DCT, IEEE Trans. Inf. Foren. Secur. 10 (2) (2015) 219–228.

[7] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Trans. Inf. Foren. Secur. 7 (3) (2012) 868–882.

[8] J. Kodovsky, J. Fridrich, V. Holub, Ensemble classifiers for steganalysis of digital media, IEEE Trans. Inf. Foren. Secur. 7 (2) (2012) 432–444.

[9] J. Kodovsky, Steganalysis of Digital Images Using Rich Image Representations and Ensemble Classifiers, Dissertations and theses, Gradworks(2012).

[10] A.D. Ker, P. Bas, R. Bhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, T. Pevn, Moving steganography and steganalysis from the laboratory into the real world, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2013, pp. 45–58.

[11] J. Kodovsk, V. Sedighi, J. Fridrich, Study of cover source mismatch in steganalysis and ways to mitigate its impact, Proc. SPIE Int. Soc. Opt. Eng. 9028 (2) (2014) 96–101.

[12] X. Xu, J. Dong, W. Wang, T. Tan, Robust steganalysis based on training set construction and ensemble classifiers weighting, in: Proceedings of the IEEE International Conference on Image Processing, 2015, pp. 1498–1502.

[13] L. Zeng, X. Kong, M. Li, Y. Guo, Jpeg quantization table mismatched steganalysis via robust discriminative feature transformation, in: Proceedings of the Media Watermarking, Security, and Forensics, 2015, p. 94090U.

[14] X. Li, X. Kong, B. Wang, Y. Guo, X. You, Generalized transfer component analysis for mismatched jpeg steganalysis, in: Proceedings of the IEEE International Conference on Image Processing, 2014, pp. 4432–4436.

[15] X. Kong, C. Feng, M. Li, Y. Guo, Iterative multi-order feature alignment for jpeg mismatched steganalysis, Neurocomputing 214 (2016) 458–470.

[16] Y. Dong, T. Zhang, X. Hou, C. Xu, A new steganalysis paradigm based on image retrieval of similar image-inherent statistical properties and outlier detection, in: Proceedings of the International Conference on Wireless Communications & Signal Processing, 2015, pp. 1–5.

[17] A.D. Ker, T. Pevn, A mishmash of methods for mitigating the model mismatch mess, in: Proceedings of the Media Watermarking, Security, and Forensics, 2014, pp. 79–85.

[18] A.D. Ker, T. Pevny, The Steganographer is the Outlier: Realistic Large-Scale Steganalysis, IEEE Press, 2014.

[19] F. Li, K. Wu, J. Lei, M. Wen, Z. Bi, C. Gu, Steganalysis over large-scale social networks with high-order joint features and clustering ensembles, IEEE Trans. Inf. Foren. Secur. 11 (2) (2017) 344–357.

[20] M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, Transfer feature learning with joint distribution adaptation, in: Proceedings of the IEEE International Conference on Computer Vision, 2014, pp. 2200–2207.

[21] M. Long, J. Wang, G. Ding, S.J. Pan, P.S. Yu, Adaptation regularization: a general framework for transfer learning, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1076–1089.

[22] M. Xiao, Y. Guo, Feature space independent semi-supervised domain adaptation via kernel matching., IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 54–66.

[23] T. Yao, Y. Pan, C.W. Ngo, H. Li, T. Mei, Semi-supervised domain adaptation with subspace learning for visual recognition, in: Proceedings of the Computer Vision and Pattern Recognition, 2015, pp. 2142–2150.

[24] L. Duan, D. Xu, W.H. Tsang, Domain adaptation from multiple sources: a domain-dependent regularization approach, IEEE Trans. Neural Netw. Learn. Syst. 23 (3) (2012) 504–518.

[25] L. Duan, I.W. Tsang, D. Xu, Domain transfer multiple kernel learning., IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 465–479.

[26] W. Li, Z. Xu, D. Xu, D. Dai, G.L. Van, Domain generalization and adaptation using low rank exemplar SVMS, IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1114–1127.

[27] D. Lerch-Hostalot, D. Megias, Manifold alignment approach to cover source mismatch in steganalysis, Reunión Española de Criptografía y Seguridad XIV, 2016.

[28] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, LOF: identifying density-based local outliers, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000, pp. 93–104.

[29] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis., IEEE Trans. Neural Netw. 22 (2) (2011) 199–210.

[30] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.

[31] B. Schlkopf, J. Platt, T. Hofmann, A Kernel Method for the Two-Sample-Problem arXiv:0805.2368, (2008) 513–520.

[32] S. Si, D. Tao, B. Geng, Bregman divergence-based regularization for transfer subspace learning, IEEE Trans. Knowl. Data Eng. 22 (7) (2010) 929–942.

[33] T. Pevn, A.D. Ker, Towards dependable steganalysis, Proceedings of SPIE - The International Society for Optical Engineering 9409 (2015).

[34] M. Belkin, P. Niyogi, Using manifold structure for partially labelled classification, in: Proceedings of the International Conference on Neural Information Processing Systems, 2002, pp. 953–960.

[35] M.J. Huiskes, B. Thomee, M.S. Lew, New trends and ideas in visual concept detection:the MIR flickr retrieval evaluation initiative, in: Proceedings of the International Conference on Multimedia Information Retrieval, 2010, pp. 527–536.

[36] W. Li, L. Wang, W. Li, E. Agustsson, L.V. Gool, Webvision Database: Visual Learning and Understanding from Web Data, arXiv:1708.02862, (2017).

[37] A. Westfeld, F5-a steganographic algorithm: high capacity despite better steganalysis, in: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, 2137, 2001, pp. 289–302.

[38] J.J. Fridrich, T. Pevn, J. Kodovsk, Statistically undetectable jpeg steganography: dead ends challenges, and opportunities, in: Proceedings of the Workshop on Multimedia & Security, 2007.

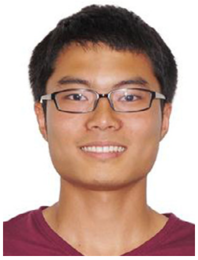
[39] P. Sallee, Model-based steganography, in: Proceedings of the Second International Workshop on Digital Watermarking, IWDW, Seoul, Korea, October 20–22, 2003, Revised Papers, 2003, pp. 154–167.

[40] N. Provos, Defending against statistical steganalysis, in: Proceedings of the Conference on Usenix Security Symposium, 2001 24–24.

[41] D. Upham, Jpeg-jsteg-v4. http://www.funet.fi/pub/crypt/steganography/jpeg-jsteg-v4.diff.gz. 1997.

[42] V. Holub, J. Fridrich, T. Denemark, Universal distortion function for steganography in an arbitrary domain, Eurasip J. Inf. Secur. 2014 (1) (2014) 1.

[43] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, Acm. T. Intel. Syst. Tec. 2 (3) (2011) 1–27.



Yong Yang received his Bachelor degree from Dalian University of Technology, China, in 2016. Currently, he is seeking his Master degree in School of Information and Communication Engineering at Dalian University of Technology, China. His research interests include digital watermarking, image steganography and steganalysis.

Xiangwei Kong received the Ph.D. degree in management science and engineering from Dalian University of Technology, Dalian, China, in 2003. From 2006 to 2007, she was a Visiting Scholar with the Purdue University, USA. From 2014 to 2015, she was a Senior Research Scientist with the New York University, USA. She was a Professor with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. She is a Professor with the Department of Data Science and Engineering Management, Zhejiang University. She has published four edited books and more than 200 research papers in refereed international journals and conferences in the areas of cross-modal retrieval, multimedia information security, knowledge mining, and business intelligence.

Bo Wang received the Ph.D. degree from Dalian University of Technology, China, in 2010. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology. His research interests include image forensics and image steganalysis.

Ke Ren received her Bachelor degree from Dalian University of Technology, China, in 2017. Currently, she is seeking her Master degree in School of Information and Communication Engineering at Dalian University of Technology, China. Her research interests include JPEG images steganography and steganalysis.

Yanqing Guo is with School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. Yanqing Guo (M’ 13) received the B.S. and Ph.D. degrees in electronic engineering from the Dalian University of Technology of China, Dalian, China, in 2002 and 2009, respectively. He is currently a Professor with the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interests include machine learning, computer vision, and multimedia security.