IEICE TRANS. INF. & SYST., VOL.E100–D, NO.5 MAY 2017

LETTER

Unsupervised Image Steganalysis Method Using Self-Learning Ensemble Discriminant Clustering*

Bing CAO[†], **Guorui FENG**[†a)], **Zhaoxia YIN**[††], ***Nonmembers***, ***and*** **Lingyan FAN**[†††], ***Member***

SUMMARY Image steganography is a technique of embedding secret message into a digital image to securely send the information. In contrast, steganalysis focuses on detecting the presence of secret messages hidden by steganography. The modern approach in steganalysis is based on supervised learning where the training set must include the steganographic and natural image features. But if a new method of steganography is proposed, and the detector still trained on existing methods will generally lead to the serious detection accuracy drop due to the mismatch between training and detecting steganographic method. In this paper, we just attempt to process unsupervised learning problem and propose a detection model called self-learning ensemble discriminant clustering (SEDC), which aims at taking full advantage of the statistical property of the natural and testing images to estimate the optimal projection vector. This method can adaptively select the most discriminative subspace and then use K-means clustering to generate the ultimate class labels. Experimental results on J-UNIWARD and nsF5 steganographic methods with three feature extraction methods such as CC-JRM, DCTR, GFR show that the proposed scheme can effectively classification better than blind speculation.

key words: *image steganalysis, statistical property, clustering, unsupervised learning*

1. Introduction

Steganography, which delivers secret messages through digital media but arouses no suspicion to others, will give a great threat to public security because individuals obviously divert some certain uncontrolled contents for the specified purpose. Steganalysis which is anti-steganographic becomes more active and urgent. Traditionally, steganalysis can be divided into two steps: feature extraction and classification. In real-world network, JPEG is the most popular storage format in Internet which has been used for decades. Thus we only consider JPEG steganalysis in this paper. As for feature extraction, by far the most mature methods are cc-JRM (JPEG rich model with Cartesian-calibration) [1], DCTR (discrete cosine transform residual) [2] and GFR (Gabor filter residual) [3]. For feature classification, the LDA (linear discriminant analysis) ensemble classifiers [4] is the most frequently used. LDA is a supervised process and mainly requires the training and testing feature sets have the same statistical distribution, and then training classifier parameters using statistical features to separate the steganographic images (stego images) and natural images (cover images).

Learning based on statistical features is the most active and efficient method to detect JPEG steganography. A group of images can be captured while it may be contains both cover and stego images. If we do not know anything about a new image steganographic method, at this time, classification algorithms will suffer from performance degradation with this situation when we just training classifier using the existing steganographic method. Our main work is to process this problem: a large number of images mixed by cover and stego images are captured from Internet, and splits them into two categories in spite we have no knowledge about the steganographic methods. This is called as stego-free image steganalysis. To our knowledge, this is the first work to address stego-free image steganalysis problem using LDA ensemble classifiers. In this letter, we present the detection performance of the proposed scheme with three state-of-the-art feature extraction schemes, i.e. the CC-JRM [1], DCTR [2] and GFR [3] over two selected popular steganographic algorithms of J-UNIWARD [5] and nsF5 [6] in the JPEG image.

2. Self-Learning Ensemble Discriminant Clustering Steganalysis Scheme

Generally, the traditional approach first trains classifier parameters using the training features (the cover and the stego image feature) by LDA [10] to obtain the optimal projection vector which as much as possible separates each of two features, then cluster these sets by K-means in the feature subspace [8]. We still adapt this framework.

2.1 The Principle of LDA and K-Means

Consider a set of input data vector consisting of n data points $\{\mathbf{x}_i\}_{i=1}^{n} \in R^m$. Denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ as the data matrix whose ith vector is given by $\mathbf{x}_i$. Because image steganalysis is a binary classification problem, $\mathbf{m}_j = \sum_{\mathbf{x}_i \in \mathfrak{C}_j} \mathbf{x}_i / n_j, (j = 1, 2)$ is the mean of the jth cluster $\mathfrak{C}_j$, where n_j is the sample size of the jth cluster $\mathfrak{C}_j$ and $\mathbf{m}$ as the mean of $\mathbf{X}$. The total scatter matrices, between-cluster scatter, and within-cluster scatter are defined as follows:

Manuscript received January 12, 2017.
Manuscript publicized February 18, 2017.

[†]The authors are with School of Communication and Information Engineering, Shanghai University, China.
[††]The author is with School of Computer Science and Technology, Anhui University, China.
[†††]The author is with Micro-Electronics Research Institute, Hangzhou Dianzi University, China.

*This work is supported by the National Science Foundation of China under Grant No. (U1536109, 61373151, 61502009), Zhejiang Provincial Science & Technology Innovation Team focused fund under Grant No. 2013TD03.

a) E-mail: fgr2082@aliyun.com

DOI: 10.1587/transinf.2017EDL8011

$$\mathbf{S}_t = \sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \tag{1}$$

$$\mathbf{S}_b = \sum_{k=1}^{2} n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T \tag{2}$$

$$\mathbf{S}_w = \sum_{k=1}^{2}\sum_{i\in C_k}(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T \tag{3}$$

It can be shown that $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$. LDA objective function [8] is:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \tag{4}$$

It has been proved that the optimal $\mathbf{w}$ can be represented as $\mathbf{w} = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$. While the K-means clustering objective function is

$$\min J_k = Tr(\mathbf{S}_w) = Tr(\mathbf{S}_t - \mathbf{S}_b) \tag{5}$$

where $Tr(\mathbf{S}_w)$ means the trace of the matrix $\mathbf{S}_w$. Because of $\mathbf{S}_t$ is a constant, the K-means clustering minimizes $\mathbf{S}_w$ or maximizes $\mathbf{S}_b$ [7]. So LDA and K-means clustering both minimizes $\mathbf{S}_w$ and maximizes $\mathbf{S}_b$. A coherent framework is proposed to integrate LDA and K-means clustering [8]: we use LDA to do subspace selection and use K-means clustering to generate class labels.

2.2 Self-Learning Ensemble Discriminant Clustering

In the real application, we can capture batch images which may contain cover and stego images from Internet while we do not know the steganographic method. As can be seen from the previous section, the ultimate aim of LDA is training the optimal projection vector $\mathbf{w}$, after that we use K-means clustering to generate class labels. Our goal is to find the most discriminative subspace in an unsupervised manner. In this section, we present an approximate evaluation method to find the approximate $\widetilde{\mathbf{w}}_o$ close to the true $\mathbf{w}$.

Consider a set of data vectors $\mathbf{X} = [\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_t]$ as the known cover feature data, and $\mathbf{Y} = [\overline{\mathbf{y}}_1, \ldots, \overline{\mathbf{y}}_N]$ as the sets to be detected batch unlabeled images' feature data from Internet. We can extract natural and stego image features captured from these images. Denote $\overline{\mathbf{m}}_1 = \sum_{i=1}^{t}\overline{\mathbf{x}}_i/t$ as the mean of the cover cluster, t is the number of known cover images. Moreover, $\widetilde{\mathbf{m}} = \sum_{j=1}^{N}\widetilde{\mathbf{y}}_j/N$ denotes the mean of the unlabeled feature, where N is the number of unlabeled images. We define $\widetilde{\mathbf{m}}_j = \sum_{\widetilde{\mathbf{x}}_j\in\mathfrak{C}_j}\widetilde{\mathbf{x}}_j/N_j, (j = 1, 2)$. N_1 and N_2 denote the number of cover and stego images to be detected (*unknown*). We just assume that the input cover images and the candidate cover images have the same statistical property i.e. $\overline{\mathbf{m}}_1 = \widetilde{\mathbf{m}}_1$. For the means:

$$\widetilde{\mathbf{m}} = \frac{N_1\widetilde{\mathbf{m}}_1 + N_2\widetilde{\mathbf{m}}_2}{N} \Rightarrow \widetilde{\mathbf{m}}_2 = \frac{N\widetilde{\mathbf{m}} - N_1\widetilde{\mathbf{m}}_1}{N_2} \tag{6}$$

then

$$\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2 \approx \overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2 = \overline{\mathbf{m}}_1 - \frac{N\widetilde{\mathbf{m}} - N_1\widetilde{\mathbf{m}}_1}{N_2}$$

$$\approx \frac{N_2\overline{\mathbf{m}}_1 + N_1\overline{\mathbf{m}}_1 - N\widetilde{\mathbf{m}}}{N_2} = \frac{N}{N_2}(\overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}}) \tag{7}$$

So the vectors $\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2$ and $\overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2$ have the same direction. According to Sect. 2.1, $\widetilde{\mathbf{S}}_t = \widetilde{\mathbf{S}}_b + \widetilde{\mathbf{S}}_w$, so $\widetilde{\mathbf{S}}_w = \widetilde{\mathbf{S}}_t - \widetilde{\mathbf{S}}_b$. For

$$\widetilde{\mathbf{S}}_t = \sum_{j=1}^{N}(\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{m}})(\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{m}})^T \tag{8}$$

We just take (7) into the Eq. (9), then

$$\widetilde{\mathbf{S}}_b = \frac{NN_1}{N_2}(\overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}})(\overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}})^T \tag{10}$$

For large number of images, $n = \frac{N_1}{N_2}$, that's to say, $\widetilde{\mathbf{S}}_b \approx N \times n \times (\overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}})(\overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}})^T$, here

$$\widetilde{\mathbf{w}}_o = \widetilde{\mathbf{S}}_w^{-1}(\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2) = (\widetilde{\mathbf{S}}_t - \widetilde{\mathbf{S}}_b)^{-1}(\overline{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2) \tag{11}$$

The above-mentioned algorithm is presented in **Algorithm 1**.

Algorithm 1 Self-learning discriminant clustering

Cover feature $\overline{\mathbf{X}}$ and unlabeled feature $\widetilde{\mathbf{Y}}$
1: Compute the mean of the cover cluster $\overline{\mathbf{m}}_1$ and the mean of the unlabeled feature cluster $\widetilde{\mathbf{m}}$;
2: Get $\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2$ with $\overline{\mathbf{m}}_1, \widetilde{\mathbf{m}}$ by Eq. (7);
3: Calculate $\widetilde{\mathbf{S}}_t, \widetilde{\mathbf{S}}_b$ and compute projection $\widetilde{\mathbf{w}}_o$ by Eqs. (8-11);
4: Run K-means: obtain the cluster label vector.

For the ensemble processing, each weak classifier of the ensemble classifier [4] applies to **Algorithm 1** and obtains the cluster labels. After collecting all cluster labels, the final classifier result is formed by combining them using majority-voting strategy.

3. Experimental Verification

In our experience, cover grayscale images sized of 512×512 are conducted on the standard database BOSSbase 1.01 [9]. All images in this database are compressed with quality factors (QF=75) and (QF=95). Stego images are created using the nsF5 [6] and J-UNIWARD [5] methods with various embedding rates. For each steganographic method with a certain embedding rate tested in this section, we use three JPEG-phase-aware steganalysis feature sets: CC-JRM [1], DCTR [2] and GFR [3]. All the results are from the average of ten times. Seldom works have been done to address this type of image steganalysis problem before. In our work, we use the detection error P_E to evaluate the detection performance, which can be defined as the total detection right probabilities:

$$P_E = \frac{N_1 acc1 + N_2 acc0}{N} \tag{12}$$

$$\begin{aligned}\widetilde{\mathbf{S}}_b &= N_1(\widetilde{\mathbf{m}}_1-\widetilde{\mathbf{m}})(\widetilde{\mathbf{m}}_1-\widetilde{\mathbf{m}})^T+N_2(\widetilde{\mathbf{m}}_2-\widetilde{\mathbf{m}})(\widetilde{\mathbf{m}}_2-\widetilde{\mathbf{m}})^T = N_1\left(\widetilde{\mathbf{m}}_1-\frac{N_1\widetilde{\mathbf{m}}_1+N_2\widetilde{\mathbf{m}}_2}{N}\right)\left(\widetilde{\mathbf{m}}_1-\frac{N_1\widetilde{\mathbf{m}}_1+N_2\widetilde{\mathbf{m}}_2}{N}\right)^T \\ &+N_2\left(\widetilde{\mathbf{m}}_2-\frac{N_1\widetilde{\mathbf{m}}_1+N_2\widetilde{\mathbf{m}}_2}{N}\right)\left(\widetilde{\mathbf{m}}_2-\frac{N_1\widetilde{\mathbf{m}}_1+N_2\widetilde{\mathbf{m}}_2}{N}\right)^T = \frac{N_1N_2}{N}(\widetilde{\mathbf{m}}_1-\widetilde{\mathbf{m}}_2)(\widetilde{\mathbf{m}}_1-\widetilde{\mathbf{m}}_2)^T\end{aligned} \tag{9}$$

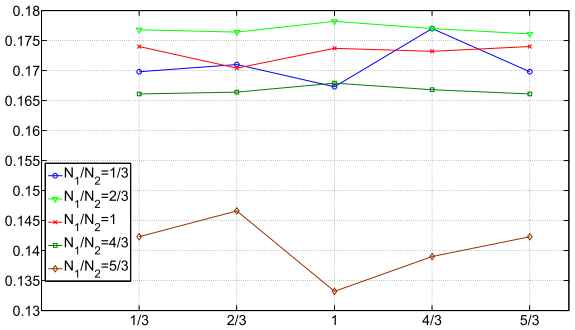

Fig. 1 Detection error P_E of different estimate values of $\tilde{n}$ with ratio of cover and stego images against nsF5 with 0.2 bpnzac payload

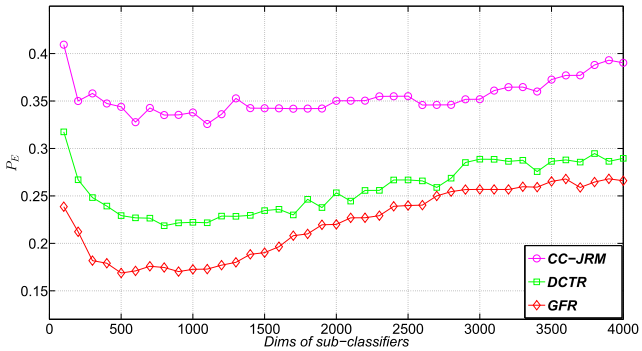

Fig. 2 Detection error P_E of CC-JRM, DCTR,GFR against JUNIWARD (Payload = 0.4 bpnzac) with different dims of sub-classifiers

where $acc1$ means the accuracy on the cover images (true covers/covers) and $acc0$ means the accuracy on the stego images (true stegos/stegos).

As can be seen in the previous section, the proposed algorithm SEDA in Eq. (10) has a parameter $\tilde{n}$ while we can not know its value but need estimate it to calculate the optimal vector. In Fig. 1, against nsF5 of 0.2 bpnzac (bits per non-zero AC coefficient) payload with QF=75. We just select randomly 8400 images ($N = 8400$). X-axis is the estimated value of $\tilde{n}$ and Y-axis shows the detection error in which case that the actual value of $\tilde{n}$ defined as the ratio of the number of cover images (N_1) and stego images (N_2), i.e. the curve $N_1/N_2 = 1/3$ denotes that the test sets contains cover images $N_1 = 2100$, stego images $N_2 = 6300$ while the total number of test sets is 8400. From the large number of experience, in the same case of N_1/N_2, the estimated value of $\tilde{n}$ is not obvious influence on experimental results, i.e. the curve $N_1/N_2 = 1/3$, by setting different n values, the value of P_E hardly change. Thus, in the next experiment, we set $n = 1$.

In the configuration of the experiments, the number of the sub-classifiers L is always 35. As shown in Fig. 2, when three steganalysis schemes including CC-JRM, DCTR, and GFR are tested against JUNIWARD of 0.4 bpnzac payload, the optimal dimension of the sub-classifiers to CC-JRM, DCTR and GFR is 1100, 900 and 800 respectively. The detection performance is good enough to stabilize while the dimension is not too high.

Tables 1-2, $N_1 = N_2 = 5000$, show that the proposed scheme deal with three state-of-the-art steganalysis schemes including CC-JRM, DCTR and GFR when the various payloads of J-UNIWARD and nsF5 with QF=75 and QF=95. With the higher embedding rate, the detection is easier, especially against nsF5. Unfortunately, the proposed method not performs effectively against J-UNIWARD with the low payload. But it should be noticed that the traditional supervised method also ineffective under the same conditions.

Table 1 Detection error P_E for proposed method of CC-JRM, DCTR and GFR against J-UNIWARD of different payloads with $QF = 75$ and $QF = 95$.

QF	Feature	Payload (bpnzac)				
		0.1	0.2	0.3	0.4	0.5
75	CC-JRM	0.5308	0.4786	0.4139	0.3367	0.2662
	DCTR	0.5226	0.4399	0.3208	0.2414	0.1567
	GFR	0.4665	0.3807	0.2436	0.1807	0.1136
95	CC-JRM	0.5409	0.5402	0.5176	0.4730	0.4277
	DCTR	0.5306	0.5297	0.4887	0.4340	0.3502
	GFR	0.5170	0.5070	0.4583	0.3888	0.3059

Table 2 Detection error P_E for proposed method of CC-JRM, DCTR and GFR against nsF5 of different payloads with $QF = 75$ and $QF = 95$.

QF	Feature	Payload (bpnzac)				
		0.05	0.1	0.15	0.2	0.3
75	CC-JRM	0.4574	0.2552	0.1764	0.1050	0.0488
	DCTR	0.4397	0.3159	0.1913	0.1310	0.0395
	GFR	0.4739	0.3657	0.2613	0.1772	0.0767
95	CC-JRM	0.3953	0.2128	0.1228	0.0571	0.0264
	DCTR	0.4566	0.2967	0.1612	0.0743	0.0257
	GFR	0.4980	0.3621	0.2740	0.1889	0.0956

To consider this condition, if we do not know anything about those steganography, what we can do is only to guess. Blind speculate will gain a detection right probabilities of 50%. But if it is treated in the proposed algorithm SEDA, the detection right probabilities is higher than 50% as shown in above tables. Thus, the performance of SEDA is useful when dealing with the problem of a new steganographic algorithm is proposed while we can not know anything about it's property.

4. Conclusion

In this letter, an unsupervised learning method which integrates LDA and K-means clustering into a join framework is

devised to solve stego-free image steganalysis. This method can solve the blind image steganalysis problem. In the proposed method, LDA and K-means are employed to do subspace selection and clustering to generate class labels respectively. We just have no knowledge of the statistical property, thus adopt the approximate evaluation to realize unsupervised learning. The key point in this framework is how to evaluate the optimal projection vector approximately and utilize the K-means clustering for feature classification. Experimental results show that the proposed method can effectively detect the satae-of-the-art steganographic algorithms J-UNIWARD and nsF5. When the target algorithm is J-UNIWARD with payload of 0.5 bpnzac against GFR, the detection error rate to the two quality factor is 11.36% and 30.59% respectively, and the target algorithm is nsF5 with payload of 0.3 bpnzac against the same situation, the detection error rate is 7.67% and 9.56% respectively. This fully shows that our method is effective for this stego-free steganalysis.

References

[1] J. Kodovský, J. Fridrich, N.D. Memon, A.M. Alattar, and E.J.D. Iii, "Steganalysis of JPEG images using rich models," Proceedings of SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics XIV, vol.8303, p.83030A, San Francisco, CA, Jan. 23–25, 2012.

[2] V. Holub and J. Fridrich, "Low-complexity features for JPEG steganalysis using undecimated DCT," IEEE Trans. Inf. Forensics Security, vol.10, no.2, pp.219–228, 2015.

[3] X.F. Song, F.L. Liu, C.F. Yang, X.Y. Luo, and Y. Zhang, "Steganalysis of adaptive JPEG steganography using 2D Gabor filters," Proc. 3rd ACM Workshop on Information Hiding and Multimedia Security, pp.15–23, 2015.

[4] J. Kodovský, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," IEEE Trans. Inf. Forensics and Security, vol.7, no.2, pp.432–444, 2012.

[5] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," EURASIP Journal on Information Security, vol.2014, no.1, pp.1–13, 2014.

[6] J. Fridrich, T. Pevný and J. Kodovský, "Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities," Proc. 9th ACM workshop on Multimedia and security, Dallas, TX, pp.3–14, Sept. 20–21, 2007.

[7] A. Wu, G. Feng, X. Zhang, and Y. Ren. "Unbalanced JPEG image steganalysis via multiview data match," J. Vis. Commun. Image R., vol.34, pp.103–107, 2016.

[8] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and K-means clustering," Proc. 24th International Conference on Machine Learning (ICML), pp.521–528, 2007.

[9] [Online]. Available: http://exile.felk.cvut.cz/boss/BOSSFinal/index.php?mode=VIEW&tmpl=materials

[10] C.M. Bishop, Pattern recognition and machine learning, Springer, 2006.