Performance Evaluation of Exponential Discriminant Analysis with Feature Selection for Steganalysis

Gaurav Rajput, R.K. Agrawal, and Namita Aggarwal

Jawaharlal Nehru University, New Delhi – 110 067, India
E-mail: gauravrajput1@gmail.com

ABSTRACT

The performance of supervised learning-based seganalysis depends on the choice of both classifier and features which represent the image. Features extracted from images may contain irrelevant and redundant features which makes them inefficient for machine learning. Relevant features not only decrease the processing time to train a classifier but also provide better generalisation. Linear discriminant classifier which is commonly used for classification may not be able to classify in better way non-linearly separable data. Recently, exponential discriminant analysis, a variant of linear discriminant analysis (LDA), is proposed which transforms the scatter matrices to a new space by distance diffusion mapping. This provides exponential discriminant analysis (EDA) much more discriminant power to classify non-linearly separable data and helps in improving classification accuracy in comparison to LDA. In this paper, the performance of EDA in conjunction with feature selection methods has been investigated. For feature selection, Kullback divergence, Chernoff distance measures and linear regression measures are used to determine relevant features from higher-order statistics of images. The performance is evaluated in terms classification error and computation time. Experimental results show that exponential discriminate analysis in conjunction with linear regression significantly performs better in terms of both classification error and compilation time of training classifier.

Keywords: Feature selection, Chernoff distance measure, kullback divergence, linear regression, steganalysis, exponential discriminant analysis

1. INTRODUCTION

Steganography is the science of embedding hidden message in one of the cover multimedia, i.e. texts, images, audio, video files. The goal of steganalysis is to detect hidden information from observed data with little or no information about the steganography algorithm. Steganalysis has drawn attention of research community in last few years since detection of hidden messages can lead to the prevention of devastating security incidents. Steganalysis[1-6] is a very challenging field because of the scarcity of information about the specific characteristics of the cover multimedia that can be exploited to hide information and detect the same. Detection of hidden message becomes more difficult as the approaches adopted for steganalysis also sometimes depend on the underlying steganography algorithm(s) used.

Several steganalysis approaches[6-9] have been proposed which can broadly be classified into four categories: Supervised learning-based steganalysis[10,11], blind identification-based steganalysis[7], parametric statistical steganalysis[9,12,13] and hybrid techniques[7]. Supervised learning-based steganalysis techniques involve two phases: (a) training phase and (b) testing phase. In the training phase, examples of both stego and non-stego are provided to a statistical classifier. The learning classifier determines the best classification rule using these examples. In the testing phase, unknown images are given as input to the trained classifier to decide whether image contains a secret message or not. Blind identification methods[7] pose

the steganalysis problem as a system identification problem. Some statistical properties such as independence of host and secret message etc. are exploited. The embedding algorithm is represented as a channel and the goal is to invert this channel to identify the hidden message. Parametric statistical steganalysis[9,12,13] approaches tend to assume a certain parametric statistical model for the cover image, stego image and the hidden message. Steganalysis is formulated as a hypothesis testing problem, namely, $H0$: no message present (null hypothesis) and $H1$: message present (alternate hypothesis). A statistical detection algorithm is then designed to test between the two hypotheses. Hybrid techniques[7] involve more than one of the above approaches.

Fridrich[14], *et al.* propose a technique for estimating the unaltered histogram to find the number of changes and length of the secret message. The process involves cropping the JPEG image by four columns and then applying a quantization table to re-compress the image. The resulting discrete cosine transform (DCT) coefficient histogram will be a close estimate of the original. Avcibas[10], *et al.* use a set of image quality metrics (IQMs) to build up a discriminator algorithm that differentiates cover images from stego images. IQMs are ranked based on their F-scores using analysis of variance (ANOVA) statistical test to identify the embedded message. The success of the approach lies in the recognition of IQMs that are susceptible to steganography and whose changes as a result of message embedding can be measured in better way. To increase the

probability of a successful detection of hidden message, several IQMs are normally employed to measure the distortions at different levels of sensitivity. The mean square values for the human visual system (HVS)-weighted errors demonstrate more sensitivity to pure blur; while the gradient measure responds to changes in the texture and the image periphery[10]. The message embedding steganography algorithms differ in the changes brought to the different IQMs. Farid[1,5], *et al.* proposed the use of higher order statistics in the generic steganalysis techniques *vis-à-vis* the first-order statistics (such as the histogram DCT coefficients) employed by the specific steganalysis techniques. Steganalysis techniques fail that employs the changes in the first-order statistics for detecting the presence of hidden messages if a steganography algorithm keeps the first-order statistics intact. Farid, *et al.* propose the use of quadratic mirror filters (QMF) to decompose an image into sub-bands and then evaluate higher-order statistics metrics such as the mean, variance, kurtosis and skewness to each of the sub-bands obtained. However, the resultant features may contain noisy, irrelevant or redundant features which make them inefficient for machine learning. In fact, the presence of irrelevant and redundant features may deteriorate the performance of the classifier and requires high computation time and other resources for training and testing the data. Another important component of supervised based Steganalysis is the choice of classifier which distinguishes two different types of images. Many classification techniques have been proposed by machine learning, data mining and pattern recognition community. Each one of them is associated with some advantages and disadvantages. Linear discriminant analysis (LDA) is one of the commonly used classifier which is simple to implement, understand and computationally less expensive. However, it is not able to classify well non-linearly separable data. Recently, exponential discriminant analysis (EDA)[15] is proposed which is a variant of LDA. Exponential discriminant analysis is based on matrix approach that maps the data in such a way that the margin between different classes is much more than the margin between different classes in the original space. This provides EDA much more discriminant power to classify non-linearly separable data and helps in improving classification accuracy in comparison to LDA.

Performance of EDA in conjunction with feature selection methods were investigated in this paper. For feature selection, Kullback divergence, Chernoff distance measures, and linear regression measures are used to determine relevant features from higher-order statistics of images. The performance is evaluated in terms of classification accuracy and computation time.

2. FEATURE EXTRACTION USING HIGHER ORDER IMAGE STATISTICS

Although the human eye is most often not able to detect the presence of embedded message, but it may nevertheless changes the statistics of an image. The resultant distortions cause due to embedding in the cover image can be analysed by comparing the statistical properties of both cover and stegoimages[8,13]. Several techniques are available to detect such changes based on first order statistical distributions of intensity or transform coefficients[13,16].

However, the shortcoming of this investigation is that simple counter-measures that match first order statistics are likely to spoil detection of embedding message in cover image. The research works[1,5] has pointed out that steganalysis-based on higher-order statistical models may distinguish stegoimages and cover images. It has been observed across a large number of natural images that there exist strong higher-order statistical regularities within a wavelet-like decomposition. The embedding of a message may significantly change the statistics of image and thus becomes measures to detect stegoimages.

The decomposition of image is possible by using separable QMFs. Thus, the frequency space is divided into multiple scales and orientations. This can be accomplished by applying separable lowpass and highpass filters along the image axes generating a vertical, horizontal, diagonal, and lowpass subband. The diagonal, horizontal, and vertical subband at scale $i = 1, 2, ..., n$ are represented as $D_i(x, y, H_i(x, y)$, and $V_i(x, y)$ respectively. Subsequent scales are obtained by recursively filtering the lowpass subband. The research works[1,5] pointed out that using above decomposition the statistical model containing the mean, variance, skewness, and kurtosis of the subband coefficients for each orientation and scales can be obtained for $i = 1$ to n. This characterises the basic coefficient distributions statistically. The second set of statistics is based on the errors in an optimal linear predictor of coefficient magnitude. It is pointed out[1,5] that the subband coefficients of the image are correlated to their spatial, orientation and scale neighbours. Taking this into account, $V_i(x, y)$, a vertical band at scale i, can be represented in terms of neighbouring pixels in spatial domain as:

$$V_i(x, y) = w_1 V_i(x - 1, y) + w_2 V_i(x + 1, y)$$
$$+ w_3 V_i(x, y - 1) + w_4 V_i(x, y + 1)$$
$$+ w_5 V_{i+1}(x/2, y/2) + w_6 D_i(x, y) + w_7 D_{i+1}(x/2, y/2) \qquad (1)$$

where w_k denotes scalar weighting values. In more compact form, it can be expressed as:

$$V = Qw \qquad (2)$$

where $w = (w_1, w_2, ..., w_7)^t$, the vector V contains the coefficient magnitudes of $V_i(x, y)$ strung into a column vector and the columns of the matrix Q contain the neighbouring coefficient magnitudes as specified in (1) also strung out in column vectors.

To determine coefficients magnitudes, quadratic error function is defined[1,5] as

$$E(w) = [V - Qw]^2 \qquad (3)$$

This error function $E(w)$ can be minimised by differentiating Eqn (3) and substituting it equal to zero yields:

$$w = (Q^TQ)^{-1}Q^TV \qquad (4)$$

From the above, order statistics such as the mean, variance, skewness, and kurtosis can be evaluated. Similarly, the above procedure can be repeated to get the subbands $H_i(x, y), D_i(x, y)$. Since, there are four statistics and linear predictor for three subband are computed for $(n - 1)$ levels, we have total $12(n-1)$ error statistics and $12(n-1)$ coefficient statistics. Thus, these $24(n-1)$ statistics altogether will form a feature vector of the image.

3. FEATURE SELECTION TECHNIQUES

Feature selection is used to remove noisy, irrelevant, and

redundant features. There are two major approaches to feature selection: filter and wrapper approach[11,17,18]. Most filter methods employ statistical characteristics of data for feature selection which requires less computation time. It independently measures the importance of features without involving any classifier. Since, the filter approach does not take into account the learning bias introduced by the final learning algorithm, it may not be able to select the most relevant set of features for the learning algorithm. On the other hand, wrapper methods tend to find features better suited to the predetermined learning algorithm resulting in better performance. But, it tends to be computationally more expensive since the classifier must be trained for each candidate subset.

Feature ranking approaches have been widely and commonly investigated for feature selection[19-21]. However, disadvantage associated with feature ranking methods is that they ignore the correlation present among the features because of their univariate approach. Hence the selected features subset may possess low discriminatory capacity and increased redundancy. A forward/backward feature selection method or its combinations are used to remove redundancy and select relevant and non redundant features with a suitable measure. Among the most widely used filter methods[4] for feature selection, there are techniques based on statistical separability measures which allow one to select a suitable subset of features by assigning the degree of interclass separability associated with each subset considered. In particular, Kullback divergence, Chernoff distance measures, and linear regression are commonly employed by research community[20,22,23]. To obtain a quantitative measure of how separable are two classes, a distance measure can be easily extracted from some parameters of the data. A very important aspect of probabilistic distance measures is that a number of these criteria can be analytically simplified in the case when the class conditional p.d.f.s $p(X_i \mid C_i)$ follows multivariate normal distribution. For multivariate normal distribution for two classes, KD and CD measures are given as follows[20,23]:

$$J_d = \frac{1}{2}(\mu_2 - \mu_1)^t \left((\Sigma_1)^{-1} + (\Sigma_2)^{-1}\right)(\mu_2 - \mu_1)$$
$$+ \frac{1}{2}tr\left((\Sigma_1)^{-1}\Sigma_2 + (\Sigma_2)^{-1}\Sigma_1\right) \tag{5}$$

$$J_c = \frac{1}{2}\beta(1-\beta)(\mu_2 - \mu_1)^t \left[(1-\beta)\Sigma_2 + \beta\,\Sigma_1\right]^{-1}$$
$$(\mu_2 - \mu_1) + \frac{1}{2}\log\frac{|(1-\beta)\Sigma_1 + \beta\Sigma_2|}{|\Sigma_1|^{1-\beta}|\Sigma_2|^{\beta}} \tag{6}$$

where μ_k^i is a mean vector and $\sum_k^i$ is a covariance matrix of k-dimensional data for class C_i, $i = 1, 2$.

The regression analysis considers the relations between the selected features which minimise redundancy. While using regression analysis for data a multiple regression model is considered because there can be many features which could affect the presence or absence of stegoimage. A multiple regression model with a target variable y and multiple variables X is given by[22]:

$$y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \xi_i,$$

$$i = 1, 2, ..., n \tag{7}$$

where β_0, β_1, ..., β_n are constants estimated by observed values of X and class label y and is estimated by normal distribution having mean zero and a variance σ^2.

The sum of squares error (SSE) which is sum of the squared residuals is given by

$$SSE = \sum_{i=0}^{n}(y_i - y_i^p) \tag{8}$$

where y and y^p are observed and predicted values, respectively.

A large value of SSE signifies that the regression is predicted poorly. The total sum of squares is given by

$$SSTO = \sum_{i=0}^{n}(y_i - \bar{y}) \tag{9}$$

where $\bar{y}$ is the average of y_i. In a regression model, the choice of features which best explains the class label depends on the value of R^2 and is given by

$$R^2 = 1 - \frac{SSE}{SSTO} \tag{10}$$

4. EXPONENTIAL DISCRIMINANT ANALYSIS

Linear discriminant analysis based on statistical method is most commonly and widely used for feature extraction and pattern classification. The objective of LDA is to determine an optimal projection matrix W by maximising the ratio of between-class scatter S_b to within-class scatter S_w as follows:

$$\arg \max_{W} \left|\frac{W^T S_b W}{W^T S_W W}\right| \tag{11}$$

where $S_b = \left(\mu_k^2 - \mu_k^1\right)\left(\mu_k^2 - \mu_k^1\right)^k$

$$S_w = \sum_{i=1}^{c}\sum_{k}^{i}$$

It is shown that a vector W that maximises $J(.)$ must satisfy

$$S_b\,W = \lambda S_w\,W \tag{12}$$

for some constant λ, which is a generalised eigenvalue problem. If S_w is non singular, we can obtain a conventional eigenvalue problem by rewriting Eqn. (12) as:

$$S_w^{-1}S_b W = \lambda W \tag{13}$$

The classical LDA criterion can be rewritten as:

$$J(W) = \arg \max_{W} \left|\frac{W^T(\Phi_b^T \Lambda_b)W}{W^T(\Phi_W^T \Lambda_W)W}\right| \tag{14}$$

where $\Phi_b = (\phi_{b1}, \phi_{b2}, ..., \phi_{bn})$ and $\Phi_w = \text{diag}(\phi_{w1}, \phi_{w2}, ..., \phi_{wn})$ are the eigenvector matrix of between class scatter S_b and with-in class scatter S_w that corresponds to eigenvalues $\Lambda_b = (\lambda_{b1}, \lambda_{b2}, ..., \lambda_{bn})$ and $\Lambda_w = \text{diag}(\lambda_{w1}, \lambda_{w2}, ..., \lambda_{wn})$ respectively.

Generally, the matrix S_w is not a full rank matrix in the under-sampled case. However, the discriminant information corresponding to the zero eigenvalues of S_w have the most discriminative power. But, LDA can extracts discriminant information only from the principal subspace of S_w. Zhang[15], *et al.* have proposed another approach based on the matrix exponential and is known as exponential discriminant analysis (EDA)[15]. To extract this kind of discriminative information,

they had replaced the eigenvalues of S_w i.e., λ_{wi}, by $\exp(\lambda_{wi})$ and λ_{bi} i.e., the eigenvalue of S_b, by $\exp(\lambda_{bi})$ and represent

$$\exp(\Lambda_b) = \text{diag}(e^{\lambda b1}, e^{\lambda b2}, \ldots, e^{\lambda bn}) \text{ and}$$
$$\exp(\Lambda_w) = \text{diag}(e^{\lambda w1}, e^{\lambda w2}, \ldots, e^{\lambda wn}) \tag{15}$$

Then the LDA criterion is modified as below:

$$J(W) = \arg\ \frac{\max}{W}\left|\frac{W^T(\Phi_b \exp(\Lambda_b)\Phi)W}{W^T(\Phi_w \exp(\Lambda_w)\Phi)W}\right|$$
$$= \arg\ \frac{\max}{W}\left|\frac{W^T \exp(S_b)W}{W^T \exp(S_w)W}\right| \tag{16}$$

The matrix $\exp(S_w)$ is always a full-rank matrix; and hence the discriminative information which was contained in the null space of S_w can be extracted using Eqn (16), even though the small sample-size problem is involved.

As it is known that orthogonality is of utmost importance to discriminant analysis, because redundant features can be combined back to the same number of variables through orthogonal transformation of the measurement space. The benefit of employing orthogonal transformation is that the correlations among candidate features are decomposed so that the significance of individual features can independently be evaluated. Hence they have defined the EDA criterion by enforcing the projection matrix W in Eqn (16) to be orthogonal as follows:

In EDA, similar to the kernel method, there exists a non-linear mapping function Φ such that the scatter matrices are mapped in to a new space, i.e.,

$$J(W) = \frac{\arg\ \max}{W^T W = 1}\left|\frac{W^T \exp(S_b)W}{W^T \exp(S_w)W}\right| \tag{17}$$

$$\Phi : R^n \rightarrow F \tag{18}$$

Such that

$$S_b \rightarrow \Phi(S_b) = \exp(S_b)$$
$$S_w \rightarrow \Phi(S_w) = \exp(S_w)$$

It is shown[15] that with the help of the mapping Φ, there is a difference in diffusion scale between the within- and between-class distances, and the diffusion scale to the between–class distance is larger than the within-class distance. Hence, the margin between different classes under such mapping is much more than the margin between different classes in the original space. This provides EDA more discriminant power which helps in improving classification accuracy in comparison to LDA.

5. EXPERIMENTAL SETUP AND RESULTS

To investigate the efficacy of the EDA for classification of stegoimages and non-stego images, we prepared a database of 1500 natural images taken from different sources i.e., www.1000pictures.com, www.1000wallpapers.co. All the images were originally in JPEG format. The original image resolutions were ranging from 800 × 600 to 1600 × 1200. We first resized each one of these images to 640 × 480 pixels images and embedded message images of six different resolutions 256 × 256, 128 × 128, 64 × 64, 32 × 32, 16 × 16, and 8 × 8 into cover image using OUTGUESS[13].

1000 non-stego images and 1000 stego images were created. Features are extracted from each one of the grey images using Haar wavelet. Each image is represented in terms of 72 statistics from four levels wavelet decomposition. To remove redundancy from the selected pool of features, three feature selection measures are investigated[23]: Kullback divergence measure, Chernoff distance measure and linear regression. For Chernoff distance measure, features are selected using 3 different values ranging from 0.1 to 0.9 with an increment of 0.4. We have used naive bayes classifier (Naivebc), Fisher linear classifier (Fisherc), and exponential discriminant classifier (EDA to evaluate the performance of the feature selection methods. The average classification error is computed using ten cross-validations. All the simulations are done using matlab. Tables 1-3 show the minimum classification error achieved with different classifiers along with the number of features for different measures. For chernoff distance measure, the minimum classification error achieved for optimal value of β is also shown in Tables 1-3. The best results in each category are indicated in bold.

We observe the following from Tables 1-3:
(a) The performance of exponential discriminant analysis is significantly better in comparison to naive bayesian and fisher discriminant analysis in terms of classification error for all sizes of embedding used in experiments.
(b) The minimum classification error is achieved with linear regression for all classifiers and for different size of embedding.
(c) The number of features required to obtain minimum classification error is significantly smaller using linear regression in comparison to baseline, kullback divergence measure and chernoff distance measure using all classifiers and different size of embedding.

Table 1. Comparative results of classification error and minimum number of features for 8 × 8, 16 × 16

Size		Baseline ErrAll	JD		JC		Linear	
			Features		Features, β		Features	
8×8	Naivebc	0.5605	64	0.4705	16, 0.9	0.443	2	0.436
	Fisherc	0.324	67	0.3045	46, 0.9	0.30	24	0.2905
	EDA	0.4049	27	0.2979	15, 0.9	0.2941	14	0.2723
16×16	Naivebc	0.539	27	0.466	21, 0.9	0.435	4	0.386
	Fisherc	0.316	34	0.3045	38, 0.5	0.2955	20	0.291
	EDA	0.3957	23	0.2885	25, 0.5	0.2724	14	0.262

**Table 2.** Comparative results of classification error and minimum number of features for sizes 32 × 32 and 64 × 64

Size		Baseline ErrAll	JD		JC		Linear	
			Features		Features, β		Features	
32×32	Naivebc	0.5385	19	0.4675	13,0.9	0.4375	2	0.4275
	Fisherc	0.321	51	0.3085	36,0.9	0.3015	21	0.293
	EDA	0.4267	14	0.2912	14,0.1	0.2803	13	0.2753
64×64	Naivebc	0.5325	6	0.4725	15,0.5	0.442	2	0.433
	Fisherc	0.326	25	0.313	37,0.1	0.302	19	0.2925
	EDA	0.4457	49	0.284	16,0.1	0.2907	13	0.2636

Table 3. Comparative results of classification error and minimum number of features for size 128 × 128 and 256 × 256

Size		Baseline ErrAll	JD		JC		Linear	
			Features		Features, β		Features	
128×128	Naivebc	0.5085	25	0.49	29,0.5	0.443	2	0.436
	Fisherc	0.3355	69	0.313	45,0.5	0.3075	22	0.296
	EDA	0.4117	60	0.2997	14,0.9	0.2854	14	0.2686
256×256	Naivebc	0.436	64	0.433	22,0.5	0.4125	4	0.386
	Fisherc	0.336	67	0.277	46,0.9	0.273	20	0.2965
	EDA	0.4017	46	0.2774	29,0.9	0.2606	17	0.2523

(d) The number of features required to obtain minimum classification error is significantly smaller using linear regression in comparison to other feature selection methods. Hence, the computation time by all the learning methods will be significantly less with features selected using linear regression method.

(e) For each feature selection method, the number of features which achieves minimum classification error with EDA is less than Fisher discriminant analysis and more than naive Bayesian classifier. However, the classification error is significantly less with EDA in comparison to both Fisher discriminant analysis and naive Bayesian classifier.

Figure 1 shows the variation in computation time of different classifiers using different feature measures to achieve minimum classification accuracy for embedding size 8 x 8. The computation time required by each classifier for different measures without feature selection (baseline) is also shown in Fig 1. It can be observed from Fig 1 that the computation
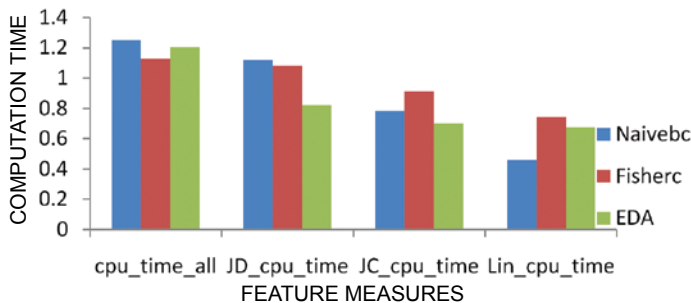
Figure 1. Variation of computation time for embedding size 8 × 8.

time required will be significantly less with linear regression in comparison to baseline, and less in comparison to kullback divergence measure and chernoff distance measure for a given classifier. Similar observations are also made for other embedding size.

6. CONCLUSIONS

The performance of supervised learning-based steganalysis depends on the choice of features to represent the image and classification method used to distinguish stego from non stego image. Features based on higher order statistics are extracted from stego and non-stego images using wavelet decompositions. Feature extracted from image may contain irrelevant and redundant features which makes them inefficient for machine learning. LDA is commonly and widely used for classification. However, LDA is not able to classify well non-linearly seaparable data. Recently, Exponential discriminant analysis is proposed which maps the data in such a way that the margin between different classes is much more than the margin between different classes in the original space. This provides EDA much more discriminant power to classify non-linearly separable data and helps in improving classification accuracy in comparison to classical LDA.

Performance of exponential discriminant analysis in conjunction with different feature selection methods were investigated in this paper. For feature selection, Kullback Divergence, Chernoff distance measures and linear regression measures are used to determine relevant features from higher-order statistics of images. The performance of steganalysis is compared and evaluated in terms of classification error and computation time of training classifier. Experimental

results show that the performance of exponential discriminant analysis is significantly better in comparison to naive bayesian and fisher discriminant analysis in terms of classification error for all sizes of embedding used in experiments. The minimum classification error is achieved with linear regression for all classifiers and for different size of embedding. Also, the number of features required to obtain minimum classification error is significantly smaller using linear regression in comparison to baseline, kullback divergence measure and chernoff distance measure using all classifiers and different size of embedding. Hence, exponential discriminant analysis in conjunction with linear regression outperforms other combination in terms of both classification error and computation time.

REFERENCES

1. Farid, H. Detecting hidden messages using higher-order statistical models. *In* Proceedings of the IEEE International Conference on Image Processing, New York 2002, **2**, pp. 905-08.

2. Johanson, N. & Jajodia, S. Exploring steganography: Seeing the unseen. *IEEE Computer*, 1998, 26-34.

3. Kahn, D. The history of steganography. *In* Anderson, R. (ed.) IH 1996. LNCS, 1174. Springer, Heidelberg, 1996.

4. Lie, W.N. & Lin, G.S. A Feature-based classification technique for blind image steganalysis. *Proc. IEEE Trans. Multimedia,* 2005, 7(6).

5. Lyu, S. & Farid, H. Steganalysis using higher-order image statistics. *IEEE Trans. Inf. Forensics Security,* 2006, **1**(1), 111-19.

6. Mehrabi, M.A.; Faez, K. & Bayesteh, A.R. Image steganalysis on statistical moments of wavelet subband histograms in different frequencies and support vector machine. *In* Proceedings of the 3rd International Conference on Natural Computation, August 2007. pp. 24-27.

7. Chandramouli, R. A mathematical framework for active steganalysis. *ACM Multimedia Systems*, 2003, **9**(3), 303-11.

8. Pevn´y, T. & Fridrich, J. Determining stego algorithm for JPEG images. *IEEE Proc. Iinf. Security,* 2006, **153**(3).

9. Westfeld, A. & Pfitzmann, A. Attacks on steganographic systems. *In*: Pfitzmann, A. (ed.) IH 1999. LNCS, **1768**, Springer, Heidelberg, 2000. pp. 61–75.

10. Avcibas, I.; Memon, N. & Sankur, B. Steganalysis using image quality metrics. *IEEE Trans. Image Proc.*, 2003, **12**(2), 221-29.

11. Guyon, I. & Elisseeff, A. An Introduction to Variable and feature Selection. *J. Machine Learning Res.*, 2003, **3**, 1157-182.

12. Harmsen, J. & Pearlman, W. Steganalysis of additive noise modelable information hiding. *In* Proceedings of SPIE Electronic Imaging, 2003.

13. Provos, N. & Honeyman, P. Detecting steganographic content on the internet. ISOC NDSS'02, San Diego, CA, February, 2002. August 2001, CITI Techreport.

14. Fridrich, J.; Du, R. & Long, M. Steganalysis of lsb encoding in color images, IEEE ICME, 2000.

15. Zhang, T; Fang, B.; Tang, Y.; Shang, Z. & Xu, B. Generalised discriminant analysis: A matrix exponential approach. *IEEE Trans. Systems, Man, Cybernetics- Pt B,* 2010, **40**(1), 186-97.

16. Wang, Y. Optimised Feature Extraction for Learning-Based Image Steganalysis. *IEEE Trans. Infor. Forensics Security*, 2007, **1**.

17. Kohavi, R. & John, G. Wrapper for feature subset selection. *Artificial Intelligence*, 1997, **1-2**, 273-24.

18. Ruiz, R.; Riquelmea, José C. & Aguilar-Ruizb, Jesús S. Incremental wrapper based gene selection from microarray data for cancer classification. *Pattern Recognition,* 2006, **39**(12), 2383-392.

19. Guyon, I.; Weston, Jason; Barnhill, Stephen & Vapnik, Vladimir. Gene selection for cancer classification using support vector machine. *Machine Learning*, 2003, **46**, 263-68.

20. Pierre, A.D. & Kittler, J. Pattern recognition: A statistical approach. PHI, 1982.

21. Tibsrani, R.; Hastie, Trevor; Narasimhan, B. & Chu, Gilbert. Diagnosis of multiple cancer types by shrunken centriods of gene expression. *Proc. Natl. Acad. Sci USA.*, 2002, **99**, 6567-572.

22. Han-Saem, P.; Yoo, Si-Ho & Cho, Sung-Bae Forward selection Method with regression analysis for optimal gene selection in cancer classification. *Int. J. Comp. Math.*, 2007, **84**(5), 653-68.

23. Rajput, G.K. & Agrawal R. K. Evaluation of feature selection measures for steganalysis. PReMI 2009, LNCS 5909 Springer, 2009. pp. 432-439.

Contributors

Mr Gaurav Rajput received MSc (Math) from Aligarh Muslim University and MTech from Jawaharlal Nehru University, New Delhi, Currently he is pursuing PhD from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. Her current area of research are Steganography and Steganalysis.

Dr R.K. Agrawal obtained MTech (Computer Application) from Indian Institute of Technology Delhi, New Delhi and PhD (Computational Physics) from University of Delhi, Delhi. Presently, he is working as an Associate Professor at the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. His current research areas are: Classification, feature extraction and selection for pattern recognition problems in domains of image processing, security, and bioinformatics.

Ms Namita Aggarwal received BSc (Math) from University of Delhi, Delhi and MCA from Delhi University, Delhi. Currently she is pursuing PhD from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. Her current area of research is pattern recognition.