A Novel Bayesian Classifier with Smaller Eigenvalues Reset by Threshold Based on Given Database

Guorong Xuan[1], Xiuming Zhu[1], Yun Q. Shi[2], Peiqi Chai[1], Xia Cui[1], Jue Li[1],
[1]Dept. of Computer Science, Tongji University, Shanghai, China
[2]Dept. of ECE, New Jersey Institute of Technology, Newark, New Jersey, USA
grxuan@public1.sta.net.cn,

Abstract

A novel Bayesian classifier with smaller eigenvalues reset by threshold based on database is proposed in this paper. The threshold is used to substitute eigenvalues of scatter matrices which are smaller than the threshold to minimize the classification error rate with a given database, thus improving the performance of Bayesian classifier. Several experiments have shown its effectiveness. The error rates of both handwritten number recognition with MNIST database and Bengali handwritten digit recognition are small by using the proposed method. The steganalyszing JPEG images using this proposed classifier performs well.

Keywords: Improved Bayesian classifier, threshold, eigenvalues, handwritten digit recognition，steganalysis

1. Introduction

Research on improving classifier has always been a hot point in pattern recognition. Given probability distribution, Bayesian Classifier is well known to be the optimum for classification [1]. In addition, its discriminant function is rather simple.Generally, Gaussian distribution is used to simulate the real distribution of samples, and is only related to mean and covariance. The Baysian classifier .minimize the classification error ratio. However, if the probability distribution is not known, how to achieve smaller error rate in classification is a practical issue that needs to be resolved. Furthermore, in the case of high dimensionality or the case of small sample database, the covariance matrix is often not full ranked, i.e., there are zero and/or small eigenvalues, it is difficult to obtain discriminant function. This has limited utilization of Baysian classifier.

In [2], the principal component null space method was proposed. It keeps the zero variance subspace. For some applications, it does achieve good performance. In fact, null space method can be viewed as zero variance re-setting from this proposed methodology

In this paper[1], a novel Bayesian classifier with smaller eigenvalues resetting by threshold which is determined based on the given database is proposed. It is similar to

[1] This research is supported partly by National Natural Science Foundation of China (NSFC) on the project (90304017).

the conventional Baysian classifier based on Gaussian distribution. However, it uses a threshold to reset the eigenvalues which are smaller than this threshold to make the classification error rate for the given database smallest.

2. Baysian Classifier

Bayesian classifier can be also expressed as follows:

$$i = \arg\max_i(g_i(x)) \quad (1)$$

$$g_i(x) = -(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \ln|\Sigma_i| \quad (2)$$

where x is sample with dimensionality d, $g_i(x)$ is the discriminant function, μ_i is the mean vector for class i, Σ_i is the covariance matrix for the class i. The minimum error rate classification can be achieved by use of the maximizing the discriminant functions under Gaussian distribution and equal prior probability [1, p.36]:

Bayesian classifier can also be also expressed d×d dimensional eigenvalue matrix Λ_i and d×d dimensional eigen vector matrix U_i as follows:

$$g_i(x) = -(x-\mu_i)^t \left(U_i^t \Lambda_i U_i\right)^{-1}(x-\mu_i) - \ln|\Lambda_i| \quad (3)$$

The $\Sigma_i = U_i^t \Lambda_i U_i$,where eigenvalue matrix Λ_i :

$$\Lambda_i = \begin{bmatrix} \lambda_{i1} & & & & & \\ & \ddots & & & & \\ & & \lambda_{ik} & & & \\ & & & \lambda_{i(k+1)} & & \\ & & & & \ddots & \\ & & & & & \lambda_{id} \end{bmatrix} \quad (4)$$

and $$\lambda_{i1} > \cdots > \lambda_{ik} > \lambda_{i(k+1)} > \cdots > \lambda_{id} \quad (5)$$

When eigenvalues $\lambda_{i(k+1)} \cdots \lambda_{id}$ are very small or equal to zero, it is impossible to calculate $g_i(x)$, because Σ_i^{-1} does not exist. Hence, traditional Bayesian classifier cannot be used under such cases.

In [2], the null sub-space method, named Principle Component Null Space Analysis (PCNSA), is proposed to solve the problems that the matrices are not fully ranked. It utilizes the Euclidean distance in 'null space' of each class instead of discriminant function. In the PCNSA a thresholds is used by a constant 10^{-4} time class maximum eigenvalue, but it cannot be used for different conditions. In paper [3] a thresholds is used under minimum error rate classification for a given database, but it still utilizes the Euclidean distance.

An improving Bayesian classifier used in this paper instead of the Euclidean distance used in both [2] and [3], and the better results are obtained.

3. Proposed Bayes Classifier with Smaller Eigenvalues Resetting

3.1. Concept

In this paper, a method of Bayes classifier with eigenvalues smaller than a threshold resetted with the threshold is proposed, which is an improvement from the traditional Bayes Classifier under normal distribution assumption. That is, assume all of eigen values of the covariance matrix are listed in a non-increasing order: $\lambda_{i1},\cdots,\lambda_{ik}$, $\lambda_{i(k+1)},\cdots,\lambda_{id}$. Then, λ_0 with $\lambda_0 > \lambda_{i(k+1)}\cdots\lambda_{id}$ is selected to replace $\lambda_{i(k+1)},\cdots,\lambda_{id}$, and λ_0 is selected so that the classification error rate is the smallest. Some points about the proposed method are listed below.

1) The hypothesis of approximate normal distribution in large sample set is sometimes reasonable, because the distribution of large sample set is often clustered around the mean vectors.
2) The expression is close to traditional Bayesian Classifier, but the calculation difficulty of not fully ranked matrices would be avoided.
3) The threshold λ_0 is keeping larger than $\lambda_{i(k+1)}\cdots\lambda_{id}$ and smaller than $\lambda_{i1}\cdots\lambda_{ik}$. It should not be near zero, and sometimes is a large number.

3.2. Bayes Classifier with Smaller Eigenvalues Reset

A small threshold λ_0 is utilized to substitute eigenvalues $\lambda_{i(k+1)}\cdots\lambda_{id}$ of eigenvalue matrix Λ_i of covariance matrix Σ_i which are smaller than λ_0.

$$\lambda_0 > \lambda_{i(k+1)}\cdots\lambda_{id} \tag{6}$$

Bayes classifier with smaller eigenvalues resetting can be expressed by classification formula（7）and discriminant:

$$i = \arg\max_i(g_{Bi}(x)) \tag{7}$$

$$g_{Bi}(x) = -(x-\mu_i)^t\left(U_i^t\Lambda_{Bi}U_i\right)^{-1}(x-\mu_i) - \ln\left|\Lambda_{Bi}\right| \tag{8}$$

where

$$\Lambda_{Bi} = \begin{bmatrix} \lambda_{i1} & & & & & \\ & \ddots & & & & \\ & & \lambda_{ik} & & & \\ & & & \lambda_0 & & \\ & & & & \ddots & \\ & & & & & \lambda_0 \end{bmatrix} \quad (9)$$

The threshold λ_0 of Bayes classifier with smaller eigenvalues resetting is carefully selected to make sure that the pattern recognition rate is maximum for a given database.

$$\lambda_0 = \arg\min_{\lambda_0 \, in \, D_{Bi}} (error_{database}) \quad (10)$$

3.3. Discussion about Threshold λ_0

As mentioned above, the value of λ_0 is mainly determined by the given database. There are several cases can be considered.

Case 1: $$\lambda_0 \le \min(\lambda_{id}) \quad (11)$$

λ_0 is smaller than all eigenvalues of all covariance matrices. In this case, $\Lambda_{Bi} = \Lambda_i$ the proposed Bayesian classifier is the same as traditional Bayesian classifier.

Case 2:

$$\lambda_0 \ge \max(\lambda_{i1}) \quad (12)$$

λ_0 is bigger than all eigenvalues of all covariance matrices.

$$g_{Bi}(x) = -\lambda_0 (x-\mu_i)^t (x-\mu_i) - d \cdot \ln|\lambda_0| \quad (13)$$

In this case, the proposed Bayesian classifier is only related to mean values and is in fact the same as Euclidean distance classifier.

Case 3: $$\max(\lambda_{i1}) > \lambda_0 > \min(\lambda_{id}) \quad (14)$$

This is actually the most of cases, and the core of the proposed Bayesian classifier, which is supposed and later shown in this paper to perform better than the traditional Bayesian classifier.

3.4. Discussion about Essence

(1) If the normal distribution assumption is valid for a given database, the

Bayesian classifier is optimal. When the eigenvalues are close to zero, Bayesian classifier cannot be used. Here, a small threshold is introduced and Bayesian classifier can be used.

(2) If the assumption of Gaussian distribution is not valid, the traditional Bayesian classifier is no longer optimal. The threshold, which is selected according to minimization of classification error rate, makes the proposed method to act like a 'regulator' to reduce the error rate.

(3) The proposed Bayesian classifier with smaller eigenvalues resetting by a threshold determined based on the given database divides all eigenvalues into two parts. In the first part, the eigenvalues are smaller than the threshold. It is known that these small eigenvalues contribute more to correct classification [3]. After their resetting by using the selected thereshold, the Euclidean distance is actually used in this part. In the second part, the eigenvalues are bigger than the threshold and contribute less to correct classification than those smaller eigenvalues. They are used in the same way as used in Baysian classifier.

(4) The proposed small eigenvalues resetting with threshold determeined by the give database is different from principle component null space analysis (PCNSA) [2], classwise non-principal component analysis (CNPCA) [3], Fisher classifier, Mahalanobis distance based classifier, Euclidean distance based classifier, Baysian classifier after lowering dimensionality via principle component analysis (PCA) [1].

3.5. Classification Rejection Rate

From the classification formula（7）and discriminant formula（8）, we use two discriminants to obtain the classification rejection rate. Consider a test sample. If its corresponding feature vector makes the discriminant $g_{Bi1}(x)$ to reach the maximum among all classes, then this sample is classified to belonging to class i_1. Similarly, if the feature vector makes the discriminant $g_{Bi2}(x)$ the maximum then the sample is classified to class i_2.

Now assume, among all of the classes, the $g_{Bi1}(x)$ and $g_{Bi2}(x)$ are two largest values, and the difference of these two largest values is smaller than a pre-defined threshold value R as shown follow:

$$\left|g_{Bi2}(x)-g_{Bi1}(x)\right| \le R. \qquad (15)$$

We then claim that the input feature vector, hence, the corresponding sample is rejected for classification. That is, the two discriminants which assume the largest values determine the so-called classification rejection rate "R". Obviously, the larger the classification rejection value R, the larger the reject rate, the smaller the classification error rate.

4 Experimental Works

4.1 Experiment I: Classification of Handwritten Numbers on MNIST Database

The MNIST handwriting number database [4] is a well-known and commonly used database for performance evaluation of various pattern recognition algorithms. That is, it can be used an objective way to evaluate the performance of a classification algorithm. Different classification methods can be compared with each objectively as well. Some samples in the MNIST database are shown in Figure 1. From these samples, it is observed that the between-class distances are rather clustered and the within-class distances are rather scattered. In addition, there are 60,000 samples. Table 1 shows the different recognition error rate according to different thresholds λ_0. The smallest error rate can be obtained when λ_0= 5400. Table 2 shows the performance comparison among the proposed method and other commonly used recognition algorithms.

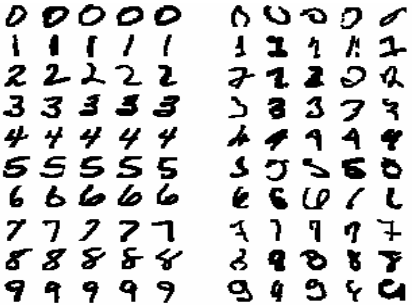

Figure 1 Some samples in MNIST database.

Table 1 Experiment results on the MNIST database

Threshold λ_0	Classification Error Rate	Comment
1000	0.0460	
4000	0.0366	
5400	0.0361	minimum
6000	0.0367	
7000	0.0363	
9000	0.0368	
11000	0.0372	

Table 2 Performance comparison among different recognition algorithms

Recognition Algorithm	Classification Error Rate
Proposed method	0.0361
Principal component null space analysis	0.0473
Bayesian classifier after PCA preprocessing	0.0622
Euclidean distance	0.1700

Because the covariance matrices are not full ranked for all of the classes, in order to use the traditional Bayesian classifier, the PCA is introduced to lower down the dimensionality to 500 first. The traditional Bayesian classifier is then applied. From Table 2, it is observed that the proposed method performs better than all of other classification algorithms. Furthermore, the scheme's execution time is very low.

4.2. Experiment II: Classification of Handwritten Bengali Numbers in Postcode

For the purpose of automatic post-mail classification, thousands of actually used envelops were collected and scanned. The single numbers in the envelops were segmented to form a database consisting of 24,876 separate hand-written Bengali numbers. The size of separate hand-written numbers is normalized to 28x28 pixels. Therefore, the feature dimensionality is d=28x28=784. The total number of classes are 10 (from 0, 1, …, up to 9). In this experiment, in order to lower classification error rate, the concept of rejecting classification described in Section 3.5 is used. The utilized rejection rule is defined as follows. That is, if $g_{i1}(x), g_{i2}(x)$ are the two largest discriminants among 10 classes then they are used in Formula (15). The *R* is a predefined value. As discussed in Section 3.5, the larger the *R* value, the larger the reject rate, and the smaller classification error rate.

Table 3 Bengali digits

Arabia digits	0	1	2	3	4	5	6	7	8	9
Bengali digits	০	১	২	৩	৪	৫	৬	৭	৮	৯

Table 4 Recognition results for Bengali handwriting postcodes

Classification rejection rate	0%	3.44%	7.9%	9.76%	18.27%
Total number of errors	176	101	46	35	12
Error detection rate	0.0361	0.0207	0.0094	0.0072	0.0025

Table 5 Eigenvalues compared with threshold λ_0=20000
(total dimensionality d=784）

Digits	0	1	2	3	4	5	6	7	8	9
Number of eigenvalues equal to or larger than threshold λ_0	731	729	725	727	705	716	722	728	727	714
Number of eigenvalues smaller than threshold λ_0	53	55	59	57	79	68	62	56	57	70

The Bengali numbers from 0, 1, up to 9 are shown in Table 3. The numbers of samples used for training and for testing are 20000 and 4876, respectively. The best threshold value is λ_0=20000. Table 4 shows the different recognition results associated with different classification rejection rates. The numbers of eigenvalues that are smaller or larger than threshold λ_0 are shown in Table 5.

Table 6 Handwritten Bengali digits classification

with zero classification rejection rate

digit	0	1	2	3	4	5	6	7	8	9	correct	error
0	1646	0	0	3	0	7	0	0	0	0	1646	10
1	9	957	1	0	1	0	0	0	0	12	957	23
2	0	12	689	0	3	3	0	0	0	2	689	20
3	11	0	0	302	0	1	9	0	0	1	302	22
4	0	0	0	0	217	0	0	0	0	1	217	1
5	1	0	0	1	2	224	7	0	0	0	224	11
6	1	4	8	13	1	2	217	0	3	1	217	33
7	3	1	0	0	1	1	0	234	0	1	234	7
8	2	0	0	1	3	0	2	0	120	1	120	9
9	1	33	2	0	0	0	2	1	1	94	94	40
total	1674	1007	700	320	228	238	237	235	124	113	4700	176

In Table 6, the test results for all of 4876 test samples are listed there. It is observed the classification error rate is 3.61%.

These recognition results are satisfactory and the algorithm has been used in the practical recognition systems now.

4.3. Experiment III: JPEG Image Steganalysis

Steganalysis can be treated as a two- or multi-class pattern recognition problem. It is noticed that the between-class distance is small, while the within-class distance is large. Hence, our proposed new Baysian classifier can play role in steganalysis as well.

The advanced JPEG steganographic techniques such as OutGuess [5], F5 [6] and MB (Model Based steganography) [7] modify the LSB (least significant bits) of JPEG AC coefficients to embed data into JPEG images, while keeping the histogram remaining unchanged. To defeat these steganographic schemes, Markov process has been used in order to utilize the 2^{nd} order statistics in steganalysis [8,9]. In this subsection, we present a new scheme, which uses the bidirectional Markov model to form 600 dimensional feature vectors for steganalysis. Then, we utilize our proposed Baysian classifier with small eigenvalues reset with theorehold. Experimental results have shown that the success of the proposed Baysian classifier in JPEG steganalysis.

4.3.1. Extraction of 600 Features

For each JPEG image, we can apply entropy decoding to obtain JPEG quantized 8x8 block discrete cosine transform (DCT) coefficients (referred to as JPEG coefficients). For each 8x8 block, we scan the JPEG coefficients in three different ways: zigzag [10], vertical and horizontal as shown in Figure 2. That is only the 21 JPEG coefficients are scanned because most of high-frequency JPEG coefficients are zero after the JPEG quantization. For each of these three scans, i.e., 1-D sequence of JPEG coefficients, we apply so called one-step [11] bidirectional [12] Markov model, and the so called probability transition matrix is used to characterize the Markov model [11]. In order to reduce the dimensionality of the probability transition matrixes, we use a thresholding technique so that all JPEG coefficients with their absolute values larger than the threshold, T, are forced to equal to T. In this way, the dimensionality of the transition matrix is (2T+1)x(2T+1). In our experimental works,

we choose T=7. Hence the probability transition matrix is of 15x15. Each 8x8 block of JPEG coefficients generates one transition matrix. Assume that we have total M 8x8 blocks. We average all of these M transition matrix, resulting one average matrix of (2T+1)x(2T+1). Since we use bidirectional Markov model, the matrix is symmetric. Hence, we only choose the 120 elements of the transitional matrix as features for steganalysis, as shown in Figure 3. Consequently, the three scans result in 120x3=360 features. In addition, we arrange the 1-D sequence resulted from the zigzag scan from all 8x8 blocks in two different ways, thus generating additional 120x2=240 features. Finally, we have totally 360+240=600 features.

$$\begin{bmatrix} 0 & 1 & 5 & 6 & 14 & 15 \\ 2 & 4 & 7 & 13 & 16 & \\ 3 & 8 & 12 & 17 & & \\ 9 & 11 & 18 & & & \\ 10 & 19 & & & & \\ 20 & & & & & \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 10 & 9 & 8 & 7 & 6 & \\ 11 & 12 & 13 & 14 & & \\ 17 & 16 & 15 & & & \\ 18 & 19 & & & & \\ 20 & & & & & \end{bmatrix} \quad \begin{bmatrix} 0 & 10 & 11 & 17 & 18 & 20 \\ 1 & 9 & 12 & 16 & 19 & \\ 2 & 8 & 13 & 15 & & \\ 3 & 7 & 14 & & & \\ 4 & 6 & & & & \\ 5 & & & & & \end{bmatrix}$$

(a) Zigzag (b) Horizontal (c) Vertical

Figure 2 Three different scanning ways

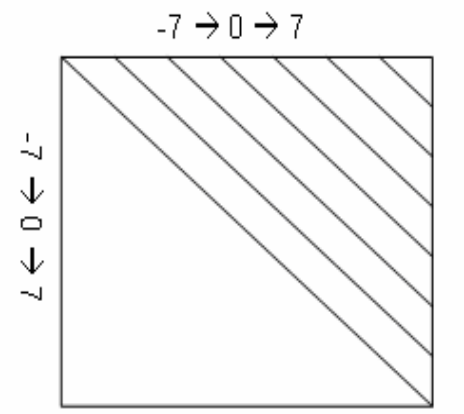

Figure 3 Probability transition matrix

4.3.2. Selection of Threshold in Proposed Baysian Classifier for Steganalysis

We used all of 1096 CoralDraw images [13] in our experimental works. 896 images are used for training and the remaining 200 images are used for testing. Each CorelDraw image is of size either 512x768 or 768x512. In order to reduce the affect caused by double JPEG compression [8], we JPEG compress the CorelDraw BMP images with Q-75 and use as the cover images. We apply OutGuess, F5, and MB without (MB1) and with deblocking (MB2) to CorelDraw BMP images with Q-75 to embed data and use these images as stego images. In experiments, we embed into

each CorelDraw image 1kB（1024 Bytes），2kB，and 4kB，respectively, corresponding to 0.021 bits per pixel (bpp), 0.041bpp and 0.083bpp, respectively. We randomly run 10 experiments. In each experiment, we randomly selected 896 image pair (cover and stego) for training and 200 image pairs for testing. The reported classification results are the arithmetic average of 10 experimental results.

Because most of eigenvalues in the bidirectional Markov transfer matrix are very small and many of them are equal to zero, a threshold can be used to substitute these small eigenvalues, indicating that the proposed Bayesian classifier lends itself to this application well.

Figure 4 displays the success rates in detecting stego images versus the adjusted threshold λ_0 in steganalyzing F5. It is observed that, when $\lambda_0=10^{-4}$, the detection rate is the best. This is also true for detecting OutGuess, MB1, and MB2. Hence, 10^{-4} is chosen for λ_0 in the experiments.

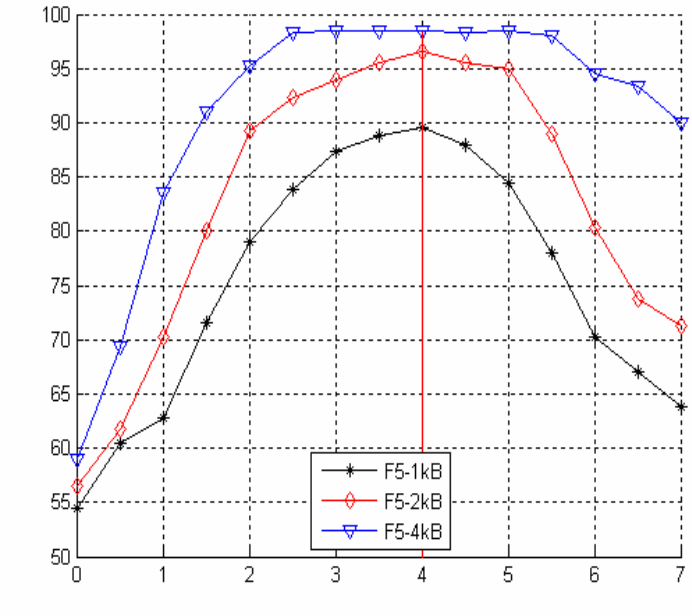

Figure 4 Adjusting Threshold in F5

（The horizontal axis is minus logarithm of threshold and the vertical is the detection ratio）

Table 7 shows the performance comparison between the proposed method and Fridrich's method [8] on the CorelDraw image database. From Table 6, we can see that the proposed method outperforms the prior art.

Table 7 Comparisons between different steganalysis methods

Data hiding Methods	Amount of hidden data (Bytes)	Detection rate (Fridrich's [8]method) %	Detection rate (proposed method) %
F5	1k	74	89
	2k	87	96
	4k	96	99
Outguess	1k	89	97
	2k	97	100
	4k	98	100
MB1	1k	66	93
	2k	86	98
	4k	90	100
MB2	1k	62	94
	2k	76	99
	4k	84	100

6. Conclusion

(1) Under Gaussian distribution, Bayesian classifier is known as optimal. However, when there is zero eigenvalue in covariance matrix, it is impossible to obtain the inverse covariance matrix, especially under high dimensional cases. In this paper, a threshold is proposed to substitute smaller eigenvalues. On the one hand, the "modified" Bayesian classifier can no longer be limited by high dimensionaity. On the other hand, the threshold can be selected carefully in order to minimize the recognition error rate in respect of the given database.

(2) The proposed Bayesian classifier does not needs PCA preprocessing. It is more effective than the traditional Bayesian classifier with the PCA preprocessing.

(3) This novel Bayesian classifier is suitable for the classification situation in which the amount of samples is large and the within-class distance is big. In many situations, when the assumption of Gaussian distribution is not valid, the threshold can act as a 'regulator' to adjust the differences between the real distribution and the assumed distribution. This characteristic also contributes to better classification performance.

(4) Three experiments have proven the effectiveness of the novel Bayesian classifier. Experiments on the MNIST database demonstrate that it is better than all other linear classifiers. It has also been utilized in practical Bengali handwritten postcode recognition system because the recognition result has met the requirements. Finally, the experiments on JPEG image steganalysis have shown that the proposed Bayesian classifier performs well for steganalysis.

7. Acknowledgement

This paper is partly supported by Chinese National Nature and Science Foundation Council. Thanks go to Professor Xuefeng Tong at Department of Computer Science, Tongji University, and Director Professor Zhikang Dai, Professor Yue Lv and Dr. Shujing Lv at Shanghai Research Institute of China Post Bureau for their kind help and valuable comments.

References

[1] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd Edition, John Wiley & Sons, 2001.
[2] N. Vaswani, R. Chellappa, "Classification probability analysis of principal component null space analysis," International Conference of Pattern Recognition (ICPR), 2004
[3] G. Xuan, Y. Q. Shi, P. Chai, X. Zhu, Q. Yao, C. Huang and D. Fu, "A novel pattern classification scheme: Classwise non-principal component analysis (CNPCA)," International Conference of Pattern Recognition (ICPR), August 20 – 24, 2006, Hong Kong, China.
[4] The MNIST Database of handwritten digits (http://yann.lecun.com/exdb/mnist/)
[5] N. Provos, "Defending against statistical steganalysis," 10th USENIX Security Symposium, Washington DC, USA, 2001.
[6] A. Westfeld, "F5 a steganographic algorithm: high capacity despite better steganalysis," 4th International Workshop on Information Hiding, Pittsburgh, PA, USA, 2001
[7] P. Sallee, "Model-based methods for steganography and steganalysis," International Journal of Image and Graphics, 5(1): 167-190, 2005
[8] J. Fridrich, "Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes," 6th Information Hiding Workshop, Toronto, ON, Canada, 2004
[9] Shi, Y. Q., Chen, C., Chen, W.: A Markov Process Based Approach to Effective Attacking JPEG Steganography. *Information Hiding Workshop*, Old Town Alexandria, VA, USA, 2006.
[10] Y. Q. Shi and H. Sun, "Image and video compression for multimedia engineering: fundamentals, algorithms and standards", CRC press, 1999
[11] A. Leon-Garcia, "Probability and random processes for electrical engineering", 2nd Edition, Addison-Wesley Publishing Company, 1994
[12]http://linuxbrit.co.uk/rbot/wiki/Ideas,
http://www.informs-cs.org/wsc00papers/064.PDF
[13] CorelDraw Software, www.corel.com.