Cross-Validation and Blind Feature Analysis of 25 Percent Embedding on JPEG Image Format using SVM

Deepa D.Shankar, Prabhat Kumar Upadhyay

Abstract: ***This paper provides a result assessment of traditional JPEG picture extraction function steganalysis compared to a cross-validation picture. Four distinct algorithms are used as steganographic systems in the spatial and transform domain. They are LSB Matching, LSB Replacement, Pixel Value Differencing and F5.A 25 percentage of embedding with text embedding information is considered in this paper. The characteristics regarded for evaluation are the First Order, Second Order, Extended DCT characteristics, and Markov characteristics. Support Vector Machine is the classifier used here. In statistical recovery, six distinct kernels and four distinct sampling techniques are used for evaluation.***

Keywords— crossvalidation, sampling, kernels, features, steganalysis, Support Vector Machines.

I. INTRODUCTION

Steganalysis is the science of discovering in a medium the existence of a concealed payload. The presence of the hidden message will not be recognized by a third party who inspects the picture [1]. In this paper we aim to suggest an analytical survey of the best way to locate a concealed payload in a medium. The selection of JPEG as a medium in this research is due to its widespread internet popularity.

Steganalysis is basically classified as Targeted and Blind. A specific steganographic algorithm is intended for targeted steganalysis [2][3]. Therefore, this analysis generates the biggest outcomes for the algorithm in question. Blind steganalysis is used to differentiate pictures with unidentified steganographic algorithms. Therefore, a statistical identification scheme is done and classification must be performed. To minimize the above errors, the steganalytic scheme requires a rational and sensible tracking frequency.

Various Steganographic algorithms considered in the research. LSB replacement is achieved by replacing the LSBs with message bits of individual coefficients [4]. The change is made to randomly raise or reduce the bits. F5 uses the embedding matrix idea. This algorithm's choice was focused on the fact that PRNG was used by the algorithm to attain confidentiality such as F5.The algorithm is implemented straight to the space domain images and applied to the transformed data in the transform domain. A histogram is produced using the coefficients frequency. Previous study focused on the features of the steganalysis picture [5]. To evaluate the result, the document utilizes a mixture of first-order, second-order characteristics and Markov characteristics.

Support Vector Machines (SVM) has been taken into account here as a classifier since prior literature states that SVM is an outstanding instrument for anticipating in broad real-life areas such as big data. The following sections discuss execution problems and experimental results. The kernel features are used to operate with SVM. Each kernel has different sampling techniques, such as linear, shuffle, stratified and automatic.

II. IMPLEMENTATION

The architecture of the steganalytic system is as follows:

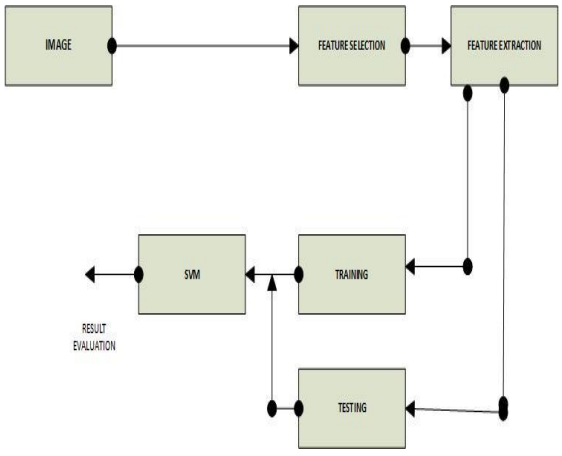

Fig: 2. System Architecture of the steganalysis system

The images used in the paper have a built-in payload of 25 which is regarded as reduced embedding. A Discrete Cosine Transform (DCT) converts the images and derives the proper features. The JPEG pictures are converted using DCT. The features are then enhanced to improve the performance of the algorithm.

Revised Manuscript Received on September 22, 2019.

Deepa D.Shankar, Research Scholar, Banasthali Vidyapith, Rajasthan, India, sudee99@gmail.com

Dr. Prabhat Kumar Upadhyay, Birla Institute of Technology Offshore Campus Ras al Khaimah, UAE, uprabhatbit@gmail.com

Feature Extraction

The research's objective is to perform a relative cross-validation survey to achieve a decent 25 percent tiny embedding rating. The system depends on extraction of relevant features. Standard deviation with individual histogram and dual histogram is regarded as first-order characteristics. Low DCT methods create the individual histogram. With the individual histogram, the global histogram can be used to visualize the coefficient of value distribution. In the characteristics of the second order, blocking, variance, and co-occurrence may occur. This is integrated as the pictures demonstrate inter-block dependencies. The functionality of the initial DCT characteristics is 23[4]. This is denoted as:

Final = | f (Si) – f (Ci) |

However, it could be expanded to 193 functional. Another series of intrablock features are the Markov features. Markov's initial features would have 324 features, which owing to large dimensionality has been lowered to 81.The paper therefore sees interblock dependencies and intrablock dependencies in order to eliminate the disadvantages generated by both characteristics An image with embedding is shown in the DCT coefficient array $d_{k(i, j)}$, where i and j are coefficients, and k is the block. The global histogram is as follows:

g is the total number of blocks and d is the fixed value of the coefficient. It is possible to represent the variance[6] as below

Where I_r and I_c are block indices vectors

Blockiness can be shown as

Where M and N are the dimensions of the image. The allocation of probabilities of adjacent DCT coefficient pairs is regarded as co-occurrence. It is represented as

The Markov characteristic establishes the distinction as a Markov process between the actual numbers of the neighboring DCT coefficients. These characteristics are calculated using four transition probability matrices. Markov's initial features will add up to 324. This will increase the size. It is taken to decrease the average of four 81 dimensionality characteristics.

B. Cross-Validation

A database of images is generally divided into a set of training and testing. This is done by assigning the image randomly, thus avoiding any bias. There is no rule to prove that the training image set and the test image set must be the same. In fact, in the real world, the training set comparatively smaller than the internet content that needs to be tested. This may contribute to powerful performance differences. Multiple training and improvisation are therefore carried out,known as k-fold cross validation. The system's strength can be evaluated by assessing the statistics on detection performance. The significance of the cross validation used in this paper is k = 10.

***C.* Classification**

Extraction of the feature follows the classification stage. SVM used in this paper as a classifier would produce an optimal hyperplane for classification. The hyperplane can be designed to provide the maximum vector support distance, known as the margin.

D. Cross-Validation

A database of images is usually split into a training and test set. This is achieved by allocating the picture evenly to prevent any bias. In fact, the training set relatively smaller in the real world than the internet content that needs to be tested. This may contribute to powerful performance differences. Multiple training and improvisation are therefore carried out, known as k-fold cross validation. The system's strength can be evaluated by assessing the statistics on detection performance. The value of cross validation used in this paper is 10.

E. Classification

Extraction of the feature follows the classification stage.SVM used in this paper as a classifier would produce an optimal hyperplane for classification. Choosing SVM as it works well with features of high dimensions. Different kernels can be considered in SVM classification.

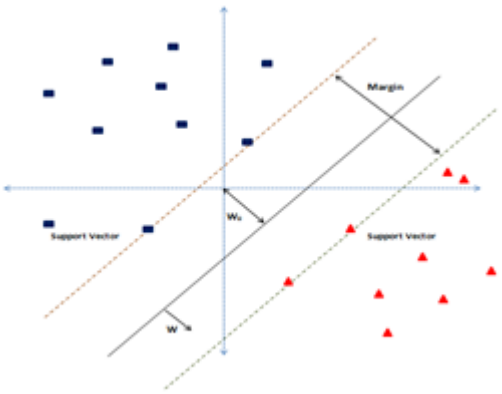

Fig: 1. Classification using Support Vector Machine

The expression of kernel radial is

Where gamma is, the gamma kernel parameter is defined. The adaptive gamma parameter performs an important part in the kernel's efficiency. The expression of dot kernel is

i.e.this is the x and y inner product. The expression of polynomial kernel is

where d is the degree of the polynomial and it is specified by the kernel degree parameter. The polynomial kernels are suitable for problems with normalization of all training data. The ANOVA kernel is defined by

where g is gamma and d is degree. The Epanechnikov kernel is defined by the function

for u between -1 and 1 and zero for u outside that range. The expression of quadratic kernel is

III. EXPERIMENTAL RESULTS

- ***Image Repository***

Any study's efficiency assessment is based on the importance of the data sets used in the research. The study uses a collection of 2300 images. All pictures are each in JPEG size. The picture is packed up to 256 X 256 size. UCID and INRIA's normal pictures are used for the dataset. The 1500 image UCID database [7] is used to create the teaching dataset and the 800 image INRIA [8] is used as sample information

- ***Training Phase***

The training data set contains 1,500 photos and 274 features each. During this stage, SVM is supplied with a collection of information either marked as cover or stego. This information is used by the classifier to lower its error matrix the quantity of false negatives and false negatives.

- ***Testing Phase***

Only after the SVM has been trained can testing be performed. 800 pictures were used, each extracting the characteristics. To prevent overfitting, the test information should be distinct from the training data. In addition, the above idea is used because the assessment is always carried out with invisible information in actual life.

IV. RESULT

For separate kernels and separate sampling, SVMs are known to be flexible and therefore regarded in this research for classification. The efficiency vector is calculated on the basis of parameters such as precision, classification error and Kappa for each kernel and sample.

A.RESULTS WITHOUT CROSS-VALIDATION

The results are as follows for non-cross-validated images.

TABLE I.RESULT DETAILS WITH LSB REPLACEMENT

	Linear	Shuffle	Stratified	Automatic
Dot	25.66	58.32	56.29	56.29
Radial	13.73	13.85	14.54	14.54
Polynomial	30.15	59.26	61.83	61.83
Multiquadric	29.78	48.72	48.98	48.98
Epanechnikov	22.57	13.96	14.56	14.56
ANOVA	49.63	57.68	58.37	58.37

From the above findings, excellent outcomes in stratified and automatic sampling are provided by the polynomial kernel. ANOVA will give the next better result.

TABLE II. RESULT DETAILS WITH LSB MATCHING

	Linear	Shuffle	Stratified	Automatic
Dot	25.66	58.32	56.29	56.29
Radial	13.73	13.85	14.54	14.54
Polynomial	30.15	59.26	61.83	61.83
Multiquadric	29.78	48.72	48.98	48.98
Epanechnikov	22.57	13.96	14.56	14.56
ANOVA	49.63	57.68	58.37	58.37

Dot From the above findings, excellent outcomes in stratified and automatic sampling are provided by the polynomial kernel. ANOVA will give the next better result.

TABLE II. RESULT DETAILS WITH LSB MATCHING

	Linear	Shuffle	Stratified	Automatic
Dot	54.38	63.94	59.72	59.72
Radial	23.52	15.73	16.73	16.73
Polynomial	31.37	47.58	49.67	49.67
Multiquadric	31.8	39.64	41.48	41.48
Epanechnikov	22.6	18.63	17.52	17.52
ANOVA	39.75	42.93	45.79	45.79

Dot kernel provides an agreeable outcome with shuffled sampling for LSB Matching.

TABLE III. RESULT DETAILS WITH PVD

	Linear	Shuffle	Stratified	Automatic
Dot	15.79	65.51	58.62	58.62
Radial	5.64	26.71	28.92	28.92
Polynomial	8.64	54.83	56.84	56.84
Multiquadric	12.68	47.74	49.46	49.46
Epanechnikov	8.43	29.75	31.83	31.83
ANOVA	16.52	17.47	17.74	17.74

The above findings show that the Dot kernel has a higher shuffle sampling result than any other kernel or test considered in the paper.

TABLE IV. RESULT DETAILS WITH F5

	Linear	Shuffle	Stratified	Auto matic
Dot	57.95	78.84	76.63	76.63
Radial	46.63	55.28	55.72	55.72
Polynomial	57.37	75.93	73.97	73.97
Multiquadric	43.88	52.47	53.26	53.26
Epanechnikov	56.16	54.37	56.25	56.25
ANOVA	58.48	75.25	91.79	91.79

For F5, ANOVA with stratified and automatic sampling provides the highest outcomes of 89.56.

V. RESULTS WITH CROSS-VALIDATION

The findings for the cross-validated pictures are as follows. The findings are obtained with distinct kernels and sampling for all four steganographic algorithms.

TABLE V. RESULT DETAILS WITH LSB REPLACEMENT

	Linear	Shuffle	Stratified	Auto matic
Dot	45.73	69.39	69.38	69.38
Radial	43.88	28.32	29.63	29.63
Polynomial	51.51	69.69	73.68	73.68
Multiquadric	43.88	60.38	61.73	61.73
Epanechnikov	43.88	49.43	51.48	51.48
ANOVA	43.88	63.83	65.87	65.87

The above findings indicate that with stratified and automatic sampling, the polynomial kernel provides the greatest outcomes.

TABLE VI. RESULT DETAILS WITH LSB MATCHING

	Linear	Shuffle	Stratified	Auto matic
Dot	54.84	65.96	61.43	61.43
Radial	43.88	32.74	29.06	29.06
Polynomial	33.59	58.26	58.23	58.23
Multiquadric	42.84	51.53	52.56	52.56
Epanechnikov	38.26	42.46	41.64	41.64
ANOVA	46.52	56.32	51.34	51.34

Like the earlier outcome, with the dot kernel and shuffled sampling, LSB Matching produces excellent outcomes.

TABLE VII. RESULT DETAILS WITH PVD

	Linear	Shuffle	Stratified	Automatic
Dot	45.75	68.47	64.11	64.11
Radial	45.75	31.53	32.54	32.54
Polynomial	31.26	61.35	58.37	58.37
Multiquadric	45.75	51.36	51.25	51.25
Epanechnikov	45.75	31.63	32.63	32.63
ANOVA	45.75	56.39	54.34	54.34

For PVD, the highest outcomes with shuffled sampling are provided in the Dot kernel.

TABLE VIII. RESULT DETAILS WITH F5

	Linear	Shuffle	Stratified	Automatic
Dot	68.42	77.43	78.56	78.56
Radial	25.86	58.41	59.15	59.15
Polynomial	72.54	75.38	74.62	74.62
Multiquadric	32.65	49.78	51.48	51.48
Epanechnikov	43.69	58.79	63.43	63.43
ANOVA	72.22	73.32	89.56	89.56

THE ANOVA KERNEL OFFERS THE LARGEST STRATIFIED AND IMMEDIATE OUTCOMES OF SAMPLING RELATIVE TO THE OTHER F5 RESEARCH KERNELS.

V. RESULT AND CONCLUSION

Four steganographic systems were used in the experiment to embed the payload message. The embedding rate of picture payload was 25. The results with SVM classifier were achieved by different kernels and separate sampling. When the dataset was not cross-validated, very tiny percentages of information had been given by the radial kernel and epanechnikov kernel. The percentage rise is significant upon verification. Linear sampling has always given a lower proportion of identification. The dot kernel provides the spatial domain a good result. For LSB Matching and PVD, it has yielded excellent outcomes. However, ANOVA provides the test field a stronger outcome with F5.

REFERENCES

1) B.Li,H.Junhui,H.Jiwu," A Survey on Image Steganography and Steganalysis", *Journal of Information Hiding and Multimedia Signal Processing*, Vol2 No.2,2011.
2) Z.Xia,X.Wang,X.Sun,"Steganalysis of LSB matching using differences between nonadjacent pixels", *Multimedia tools and Applications*,Issue 4 ,Volume 75,1947-1962,2016.
3) D. Neeta, K. Snehal and D. Jacobs, "Implementation of LSB Steganography and Its Evaluation for Various Bits,"1st International Conference on Digital Information Management, Bangalore, 173-178,2007.
4) L.Chyuan,L.Chen-Hao,"LSB steganographic method based on reversible histogram transformation function for resisting statistical steganalysis", *Information Sciences*, vol 188, 346-358,2012.
5) H.Zhongwei,L.Wei,". Digital image splicing detection based on Markov features in DCT and DWT domain", Pattern Recognition,No:12,Vol 45,,4292-4299,2012.
6) Deepa D.Shankar, T.Gireeshkumar, K.Praveen, R.Jithin, Ashji S.Raj,".Block Dependency Feature Based Classification Scheme for Uncalibrated Image Steganalysis", *Lecture Notes on Computer Science*, 189-195,2012.
7) G. Schaefer, "An uncompressed benchmark image dataset for colour imaging," 2010 IEEE International Conference on Image Processing, Hong Kong, 3537-3540,2010.
8) G. Overett, L. Petersson, N. Brewer, L. Andersson and N. Pettersson, "A new pedestrian dataset for supervised learning," IEEE Intelligent Vehicles Symposium, Eindhoven,373-378,2008.