Defence Science Journal, Vol. 60, No. 4, July 2010, pp. 420-422

SHORT COMMUNICATION

Classification of Encrypted Text and Encrypted Speech

Rajesh Asthana and Neelam Verma*

Scientific Analysis Group, Metcalfe House, Delhi-110 054
*E-mail: rajasthana_drdo@rediffmail.com, *neelamverma123@gmail.com*

ABSTRACT

The information to be exchanged between two parties can be text data or speech data. This data is encrypted for its security and communicated (to the other end). When an adversary intercepts these encrypted data then in order to recover the actual information, his first step is to identify whether intercepted data is encrypted text or encrypted speech are used. The next step is to get the actual information from encrypted text or encrypted speech. In this paper, pattern recognition techniques are applied for identification of encrypted text and encrypted speech. Some new and modified feature extraction techniques have been used to convert the text and speech data into three-dimensional, four-dimensional, and five-dimensional measurement vectors. These multi-dimensional measurement vectors are converted into two-dimensional vectors using projection pursuit technique based on Sammon's algorithm and Chien's algorithm. The quantified classification performances using minimum distance classifier and maximum likelihood classifier have also been given.

Keywords: Feature extraction, projection pursuit techniques, minimum distance classifier, maximum likelihood classifier, Sammon's algorithm, Chien's algorithm

1. INTRODUCTION

Classification of data belonging to various categories is a pattern recognition problem which is being tackled since the last four decades[1]. The classification process consists of mainly four stages, viz., data acquisition, feature extraction, feature selection, and usage of various classification techniques[2]. Data acquisition is one of the important stages as collection of relevant data for various categories in required format is the stepping stone for its analysis. In the second stage, conversion of data in the form of real valued measurement vector is done. Various feature extraction techniques have been devised[3] and depending on the problem, development of new feature extraction techniques are still an area of research. The third stage deals with the problem of curse of dimensionality where emphasis is given on selection of prominent features that are sufficient for optimum classification[4]. The classification techniques can be statistical, neural, SVM-based, Fuzzy or some hybrid technique[5].

The process of identification of intercepted data into two categories, i.e., encrypted text or encrypted speech, goes through all the above-mentioned four stages. In secure communication, data/information is encrypted using various encryption algorithms available in the literature[6]. When a person intercepts the encrypted data, then his first hunch is to identify whether this data is encrypted text or encrypted speech. Once this is achieved then one can work to get the exact information embedded in encrypted speech or encrypted text. For speech data two coding schemes linear predictive coding (LPC) and continuously variable slope delta (CVSD) have been used for its conversion to bit stream where as in text ASCII coding scheme is used for its conversion to bit stream. AES-128 encryption algorithm was consideredfor encryption of both text and speech data. The features extracted from these bit stream data were converted to two-dimensional data through projection pursuit techniques based on Sammon's nonlinear mapping algorithm and Chien's nonlinear mapping algorithm. Finally, linear and nonlinear classifiers[7] have been used for identification and their quantified results given.

The block cipher AES 128[9] has been used for encryption of plain text, LPC-coded plain speech and CVSD-coded plain speech.

For classification, experimentation was carried out on 1000 messages of encrypted text and LPC coded encrypted speech. Classification was carried out on 150 messages of encrypted text and CVSD-coded encrypted speech. The message length for both the cases was 4000 bits.

2. FEATURE EXTRACTION TECHNIQUES

The feature extraction technique converts binary encrypted data into measurement vectors. The binary bit stream is divided into 2^n (n = 3, 4 and 5) bit blocks. For each block the length of longest run of one's is computed. These lengths of longest run of ones are categorized into n classes C_1, C_2, C_3, …, C_n.

For k =1, 2,…, n–1, if the length of longest runs was = n–2+k, then the author assigned C_k. If the length of longest runs was >= 2n–2, then the author assigned C_n.

The frequencies of these n classes constitute the n components of the measurement vector called longest run frequency (LRF) – 2^n vector.

Received 28 September 2009, Revised 1 March 2010

3. PROJECTION PURSUIT TECHNIQUES

The objective of projection pursuit technique[8] is to reduce the dimensionality of the pattern vector from h-space to d-space ($h > d$ and $d = 2$) so that some inherent structure of the set of measurement vectors in high dimensional h-space is preserved in d-space. In particular, one is primarily interested in those relationships that reveal clusters in the measurement vectors in both the spaces. In this case, $h = 3$ for LRF-2[3], $h = 4$ for LRF-2[4] and $h = 5$ for LRF-2[5].

For a given data set of N samples, where each of these samples is described by an h-dimensional ($h > 2$) vector $X = \{\underline{X}_i\}$ $i = 1, 2, \ldots, N$

A nonlinear mapping (NLM) maps the vector X of high dimension to a vector Y of two-dimension; NLM: $X \rightarrow Y$

Thus Y is a collection of N two-dimensional vectors in the d-space, where $d = 2$,

So one gets $Y = \{\underline{Y}_i\}$ $i = 1, 2, \ldots, N$

Let $d_{ij}^* = \text{dis}(X_i, X_j)$ be distance between pattern X_i and pattern X_j in the h-space, and $d_{ij} = \text{dis}(Y_i, Y_j)$ be distance between transformed pattern Y_i and Y_j in the d-space. The structure is preserved if $d_{ij}^* \approx d_{ij}$ for all i and j.

The consequence of this kind of approximate preservation is the introduction of an error e_{ij} where $e_{ij} = d_{ij}^* - d_{ij}$ or an error function

$$E = f(e_{ij}) = f(d_{ij}^* - d_{ij}) \tag{1}$$

The problem is to minimise the value of this error function. The two algorithms for such nonlinear mapping are described as under.

3.1 Sammon's Nonlinear Algorithm[8]

Let the distance or dissimilarity measure between the vectors X_i and X_j in the h-space be the Euclidean metric

$$d_{ij}^* = \left\| X_i - X_j \right\|_2 = \left[\sum_{k=1}^{h} \left(x_{ik} - x_{jk} \right)^2 \right]^{1/2} \tag{2}$$

Similarly, let the distance or dissimilarity measure between the vectors Y_i and Y_j in the d-space (d=2) be the Euclidean metric

$$d_{ij} = \left\| Y_i - Y_j \right\|_2 = \left[\sum_{k=1}^{d} \left(y_{ik} - y_{jk} \right)^2 \right]^{1/2} \tag{3}$$

The error function E, which represents how well the present configuration of the N vectors in the d-space fits the N vectors in the h-space, is defined as

$$E = f(d_{ij}^* - d_{ij}) = \frac{1}{\sum_{i<j}^{N} d_{ij}^*} \sum_{i<j}^{N} \frac{\left[d_{ij}^* - d_{ij} \right]^2}{d_{ij}^*} \tag{4}$$

Initially the authors choose at random a set of Y_i's in the d-space. This set of vectors was the initial configuration of the d-space. The next step in the algorithm was to adjust the N vectors in the d-space so as to decrease the value of the error function. This was achieved by carrying out a steepest-descent procedure to search for the minimum of the error function by successively computing the new vectors Y_i's governed by the following recursive relation

$$Y_p' = Y_p - \alpha \left[\frac{\partial E}{\partial Y_p} \Big/ \left| \frac{\partial^2 E}{\partial Y_p^2} \right| \right] \quad p = 1, 2, \ldots, N$$

$$\alpha \text{ is a correction factor} \quad \text{and } \alpha = 0.01 \tag{5}$$

where

$$\frac{\partial^2 E}{\partial Y_p^2} = -\frac{2}{c} \sum_{j=1}^{N} \frac{1}{d_{pj}^* d_{pj}} \left[\left(d_{pj}^* - d_{pj} \right) - \left\| Y_p - Y_j \right\|_2^2 \frac{d_{pj}^*}{d_{pj}} \right]$$

$$\text{where } c = \sum_{i<j}^{N} d_{ij}^* \tag{6}$$

This adjustment of d-space configuration continues until the minimum value of E or the pre-specified number of iterations is achieved.

3.2 Chien's Nonlinear Algorithm[8]

The basic procedure of this method is similar to that of Sammon's method. It also consists of three steps. First, an initial configuration of the Y_i's in the d-space is chosen. Second, the Y_i's are adjusted according to an error function, using an optimisation process. Third, the adjustment process continues until the minimum value of E is achieved or the algorithm has undergone fixed number of iterations.

Let Y_p and Y_q be the pair of Y_i's selected for adjustment. Let the new vectors corresponding to Y_p and Y_q be Y_p' and Y_q', respectively. Then one gets

$$Y_p' = Y_p - \frac{1}{2} \left(\frac{1 - d_{pq}^* / d_{pq}}{1 + d_{pq}^*} \right) \left(Y_p - Y_q \right) \tag{7}$$

$$Y_q' = Y_q + \frac{1}{2} \left(\frac{1 - d_{pq}^* / d_{pq}}{1 + d_{pq}^*} \right) \left(Y_p - Y_q \right) \tag{8}$$

These two algorithms convert three-, four- or five-dimensional data into two-dimensional data.

4. CLASSIFICATION MODELS

The classifiers minimum distance classifier (MDC) and maximum likelihood classifier (MLC)[7] have been used for identification between encrypted text and LPC-coded encrypted speech and between encrypted text and CVSD-coded encrypted speech.

5. EXPERIMENTAL RESULTS

Encrypted data were converted into two-dimensional vectors using feature extraction techniques as mentioned in Section 2 and projection pursuit techniques as described in Section 3. The average learning classification success rate (L) and average testing classification success rate (T) using classifiers MDC and MLC are shown for encrypted text versus LPC-coded encrypted speech and encrypted text versus CVSD-coded encrypted speech.

5.1 Encrypted Text versus LPC coded Encrypted speech

*Sammon's NLM projection pursuit technique:*Sammon's NLM Projection Pursuit technique shows overlapping results as shown in Table 1 for encrypted text and LPC coded encrypted speech.

Table 1. Sammon's NLM projection pursuit technique

	LRF 8		LRF 16		LRF 32	
	MDC	MLC	MDC	MLC	MDC	MLC
L	52.39	52.78	50.39	51.23	51.78	52.95
T	50.00	52.50	50.56	50.39	50.50	50.84

Chien's NLM projection pursuit technique: The results of Chien's NLM projection pursuit technique gives better result as can be seen by quantified results in Table 2.

Table 2. Chien's NLM projection pursuit technique

	LRF 8		LRF 16		LRF 32	
	MDC	MLC	MDC	MLC	MDC	MLC
L	66.57	68.44	76.38	76.56	82.75	88.25
T	53.0	54.75	67.75	68	80.25	86.75

5.2 Encrypted Text versus CVSD-coded Encrypted Speech

*Sammon's NLM projection pursuit technique:*This technique does not give good result as evident by quantified results as shown in Table 3.

Table 3. Sammon's NLM projection pursuit technique

	LRF 8		LRF 16		LRF 32	
	MDC	MLC	MDC	MLC	MDC	MLC
L	52.7	53	52.4	53.5	55	53.5
T	51.4	51.2	51.2	51	52.2	52.9

Chien's NLM Projection Pursuit technique: The quantified result as given in Table 4 shows that identification using this technique is giving very good result.

Table 4. Chien's NLM projection pursuit technique

	LRF 8		LRF 16		LRF 32	
	MDC	MLC	MDC	MLC	MDC	MLC
L	90	92	92	94	90	94
T	84	86	86	88	84	86

6. CONCLUSIONS

In this paper, identification of encrypted text and encrypted speech has been done. Three feature extraction techniques and two projection pursuit techniques have been explored using two classifiers MDC and MLC. Following conclusions have been drawn from the above experimentation

(i) For both LPC- and CVSD-coded speeches, Chien's NLM projection pursuit technique is giving better result in comparison to Sammon's NLM projection pursuit technique.

(ii) The classification of encrypted text and CVSD-coded encrypted speech is giving encouraging result in comparison to that of encrypted text and LPC-coded encrypted speech.

ACKNOWLEDGEMENTS

The authors are thankful to Dr PK Saxena, Outstanding Scientist and Director SAG, for his continuous support and encouragement for this work. We are also thankful to Dr SS Bedi, Associate Director and Head, Evaluation Division, for his valuable guidance for the experimental work.

REFERENCES

1. Meisel, W.S. Computer-oriented approaches to pattern recognition. Academic Press, New York, 1972.
2. Duda, R.O.; Hart, P.E. & Stork, D.G. Pattern classification, Ed. 2, Wiley-Interscience Publication, 2001.
3. Schalkoff, R. Pattern recognition: statistical, structural and neural approaches. John Wiley & Sons, Inc. 1992.
4. Theodoridis, S. & Kourtroumbas, K. Pattern recognition. Academic Press, 1999.
5. Mizutani, E.; Sun, C.T. & Jang, J.S.R. Neuro-fuzzy and soft computing. Eastern Economy Edition, 1997.
6. Menzes, A.; Oorschot, P.V. & Vanstone, S. Handbook of applied cryptography. CRC Press, 1996.
7. Tou, J.T. & Gonzalez, R.C. Pattern recognition principles. Addison-Wesley Publishing Co. 1974.
8. Chien, Y.T. Interactive pattern recognition. Marcel Dekker Inc, 1978.
9. Stallings, W. Cryptography and network security, Ed. 3. Pearson Education Inc, 2005.

Contributors

Mr Rajesh Asthana has done post-graduation in Mathematics from the University of Gorakhpur. He joined Scientific Analysis Group (SAG) in 2003 and is presently working as Scientist C. He is working in the area of supervised and unsupervised pattern recognition techniques, support vector machines, genetic algorithms, cryptology and evaluation of crypto algorithms.

Mrs Neelam Verma has done her post graduation in Mathematics from IIT Delhi. She joined SAG, DRDO in 1986 and at present she is working as Scientist F. She is working in the area of supervised and unsupervised pattern recognition techniques, multivariate statistical analysis, randomness testing, cryptology and evaluation of crypto algorithms.