Using k-mean Clustering Technique to Diagnosis Two Cryptography systems

Raja Salih
Institue of medical technology , Baghdad , Iraq

(Received: 24 / 3 / 2013---- Accepted: 15 / 5 / 2013)

Abstract

This paper presents a new using of k-means clustering techniques. It uses for diagnosis the cryptographic system of cipher messages depending on cipher text. System diagnosis is the first step in analyzing the encrypted messages by the cryptanalyst. The proposed method classify the messages that were encrypted using two systems: simple substitution cipher or transposition technique using k- means algorithm after modifying the Euclidean equation. This method proved to be successful diagnosis all the encrypting messages, that encrypted using one of the two systems, which are tested .

Key words : clustering techniques, clustering technique, cryptography, substitution cipher, transposition cipher, Clustering algorithm, k – means, algorithm .

Introduction to Clustering Technique

Clustering is the process of grouping a set of multidimensional patterns (objects or data points), such patterns in the same cluster have the most similar characteristics, and patterns within different clusters have the most dissimilar characteristics [1] . In most cases a cluster is represented by a cluster center or a "centroid".

Clustering has been applied to a wide range of applications, such as pattern recognition, image segmentation, spatial data analysis, machine learning, data mining, economic science, and internet portals. Classification is another data analysis method. The distinction between the two approaches is that classification is a supervised learning process which is trained on a set of pre-labeled patterns In contrast, clustering is either unsupervised or supervised, unsupervised clustering, has no predefined classes [2].

A basic outline of a general clustering process can be described as follows:

1) Select a proximity measure to be used. This is used to evaluate the similarity (or dissimilarity) of two data points. This could be the Euclidean distance, correlation coefficient or other measures.

2) Apply the clustering technique to classify the dataset. Many different clustering techniques have been used .

3) Validate the clusters. Cluster validation evaluates the clustering scheme obtained from step (2). Cluster validity indices are often used to assess the quality of the clusters [2] .

The objective of cluster analysis is the classification of objects according to similarities among them, and organizing of data into groups, it means clustering can be used as a reduction technique by storing the characteristics of the clusters instead of the individual data [3]. The main potential of clustering is to detect the underlying structure in data, not only for classification and pattern recognition, but for model optimization.

In general, clustering techniques can be divided into two main categories: hierarchical and partitional clustering. In each category, many subtypes and variants have been applied to diverse types of clustering problems. However, when using data from real world applications, the boundaries between clusters might be vague [1] .

We search the internet through google search , yahoo search ,answer search for such or related subject but found there is no one take the idea resemble this research, diagnosis of cryptographic system .

Clustering Analysis

Cluster analysis aims at dividing a data set into groups or clusters that consist of similar data. There is a large number of clustering techniques available with different underlying assumption about the data and the clusters to be discovered. A simple and common popular approach is the c-means clustering. For the c-means algorithm, it is assumed that number of clusters is known or at least fixed, i.e. the algorithm will partition a given dataset: $X = \{x_1, x_2, \ldots, x_n\}$ into clusters [4] .

The partitions found should in general have the following properties

Homogeneity: The data that are assigned to the same cluster should be similar.

Heterogeneity: The data that are assigned to different clusters should be different.

In most cases the data is in the form of real-valued vectors. The Euclidean distance is a suitable measure of similarity for these datasets. The partitions should then be such that the intra-cluster distance is minimized and the inter-cluster distance is maximized. The clustering techniques can be broadly classified as follows [5][6] :

a) Incomplete/Heuristic: Clusters are determined by heuristic methods based on visualization after dimensionality reduction of the data.

The dimensionality is reduced by using geometrical methods or projection techniques .

b) Deterministic crisp: Each datum is assigned to one and only one cluster.

c) Overlapping crisp: Each datum can be simultaneously assigned to several clusters .

d) Probabilistic: A probability distribution specifies the probability of assignment of the datum to a cluster

e) Possibilistic /Fuzzy: Degrees of membership indicate the extent to which the datum belongs to the

cluster. The sum of memberships of each datum across all the clusters may not be 1 for possibilistic clustering .

f) Hierarchical: These techniques are based on either dividing the data into more fine-grained classes or combining small classes to more coarse-grained classes.

g) Objective function based: These techniques are based on an objective or evaluation function that assigns a quality or error value to each possible partition or group of clusters. The partition that obtains the best evaluation is the chosen solution.

h) Cluster estimation: These techniques use heuristic functions to build partitions and estimate the cluster parameters.

Cryptography

Cryptography enables to secret the sensitive information or transmits it across insecure channel. So that it cannot be read by anyone except the intended recipient. A cryptographic algorithm, or cipher, is a mathematical function used in the encryption and decryption process. An encryption (also called enciphering), is a mapping of plaintext to cipher text, based on some chosen key. An encryption step is an encryption applied to a particular sequence of plaintext characters, using, in a way that depends on the encryption key, a particular encryption rule of an encryption algorithm .[7][8]

Decryption or deciphering is the inverse operation of an encryption; it maps cipher text to plaintext, based on the knowledge of the key. There are many types of algorithms designed to implement cryptosystems. The security of an algorithm depends on how hard to break [9] .

A cryptographic algorithm works in combination with a key to encrypt the plaintext. The same plaintext encrypts to different cipher text with different keys. The security of encrypted data is entirely dependent on two things: the strength of the cryptographic algorithm and the secrecy of the key.[10]

Cryptanalysis is the study of techniques for attempting to defeat cryptography techniques and information security services. Cryptanalysts are also called attackers. For each ciphering technique, there is a mechanism and tools to analysis it.

The first step of the analysis is the diagnosis of the cipher text to determine which techniques was used in the plain text.

- **Substitution cipher [9]**

A substitution cipher is one in which each character in the plaintext is substituted for another character in the cipher text. The receiver inverts the substitution on the cipher text to recover the plaintext. There are many types of substitution ciphers; one of them is simple substitution.

In a simple substitution cipher, each character of the plaintext is replaced with a corresponding character of cipher text .

- **Transposition ciphers [11]**

In a transposition cipher the plaintext remains the same, but the order of characters is shuffled around. For a symmetric-key block encryption scheme with block length t and K be the set of all permutations on the set (1, 2, …, t) . The encryption function for each $e \in K$ can be defined as $E_e(m) = (m_e(1)\, m_e(2) \ldots m_e(t))$, where $m = (m_1 m_2 \ldots m_t) \in M$ (the message space). The decryption key corresponding to e is the inverse permutation $d = e^{-1}$ .To decrypt $c = (c_1 c_2 \ldots c_t)$. to compute $D_d(c) = (c_{d(1)} c_{d(2)} \ldots c_{d(t)})$.

In cryptanalysis of transposition ciphers, since the letters of the cipher text are the same as those of the plaintext, a frequency analysis on the cipher text would reveal that each letter has approximately the same likelihood as in English. This gives a very good clue to a cryptanalyst, who can then use a variety of techniques to determine the right ordering of the letters to obtain then plaintext .

Clustering algorithm

There are several algorithms that are used in clustering techniques. Requirements that should be satisfied are [1]:

1. Scalability.
2. Dealing with different types of attributes .
3. Discovering clusters of arbitrary shape .
4. Ability to deal with noise and outliers .
5. High dimensionally.
6. Insensitivity to the order of attributes .
7. Interpretability and usability .

The problem encountered with clustering algorithms is :

1. Dealing with large number of dimensions and large number of objects can prohibitive due to time complexity.
2. The outcomes of the algorithm can be interpreted in different ways . The effectiveness of the algorithm depend on the definition of similarity distance .

means Algorithm

K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative[5].

The K-Means Algorithm Properties are:

a) There are always K clusters.

b) There is always at least one item in each cluster.

c) The clusters are non-hierarchical and they do not overlap.

d) Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

The proposed algorithm

The proposed algorithms depend on **K-Means Algorithm** in general, except modify the Euclidean equations as illustrated in step 4.

Input : messages frequencies
Output: Messages classification

Step1 : Determine the number of clusters K
Step2: Calculate the frequency of letters in each message.
Step3: Choose randomly two of message as initial member for each cluster.
Step4 : find the differences between letters frequency of each initial member and the frequency of all messages using the formula

$$D_{i,j} = \sqrt{\sum_{k=0}^{n26} (f_{ik} - f_{jk})^2} \quad \ldots\ldots\ldots (1)$$

Where N is the number of alphabet characters
i,k,: the frequency k of message i.
Step 5: Collect the messages that have small Euclidean distance.
Step6: Repeat the step 2 to 5 until satisfy the stop criterion.

Figure (1): The Proposed Algorithm

Result

The following example illustrates the use of the above algorithm to classify 5 messages that were encrypted using either simple substitution or transposition, so the value of k is 2.
The first step was to calculate the frequency of the letters for each message independently (table 1).

Table 1: Messages Letters Frequency

Ms	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	7	4	3	8	29	2	3	4	9	0	0	1	4	11	11	6	0	10	11	15	5	0	2	1	3	0
2	0	0	5	0	3	15	10	6	5	4	1	0	11	11	9	0	7	8	1	29	3	11	4	2	2	3
3	0	0	0	3	6	14	7	3	8	3	2	0	20	15	0	9	9	8	1	22	4	7	2	2	3	2
4	9	2	6	8	22	3	4	8	15	0	0	2	3	19	10	3	0	7	7	14	0	3	2	1	2	0
5	3	0	13	3	21	2	4	4	15	0	1	8	6	13	15	2	1	10	9	14	2	1	1	0	2	0

Note: MS means the message number.

The next step initializes the two clusters by choosing one of messages in each cluster. Message 1 was chosen in cluster 1, and message 3 in cluster 2. The choice was done randomly. And then calculate the differences between each cluster and the all messages using Equation (1). The results are shown in Table 2 .

Table 2 : Result of Round 1

Cluster	Message 1	Message 2	Message 3	Message 4	Message 5
$G\,1^{(1)}$	0	40.5	36.6	16.0	19.2
$G\,2^{(1)}$	38	20	0	35.8	37.3

The first row represents the differences between letters frequency initialization (message 1). While the second row represents the differences between letters frequency initialization (message 2). From table (2), The values are clustered in two different groups. The group that contains value >35 substitute by 1, while the other group contains values ≤ 2 0. So, the values in first group1 replaced by 1 and values in the second group by 0.
Depending on the result, the array MG^0 consists of the value of 0 and 1. 1 when the messages belong to the cluster, 0 on otherwise .

$$MG^0 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \begin{matrix} Group\,1 \\ Group\,2 \end{matrix}$$

The message s 1, 4, and 5 are in group 1, while the messages 2 and 3 in group 2.
After knowing this results, the frequency of messages letters in the same group are calculated together. The frequency of group 1 (messages 1, 4, and 5) and the frequency of group 2 (messages 2, and 3) are presented in table 3.

Table 3. Frequency of G1 and G2 letters

	A	B	C	D	E	F	G	H	I	J	K	L	M
G1	19	13	26	19	72	7	11	16	39	0	1	11	13
G2	0	0	5	3	9	29	17	9	11	7	3	0	31
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
43	36	11	1	27	27	43	7	4	5	2	7	0	43
26	9	9	16	16	2	51	7	18	6	4	5	5	26

Using equation (1), recalculation of the difference between the frequencies between each group (G1 and G2) and each message frequency is done The result are shown in Table 4.

Table 4: Results of Round 2

Group	Message 1	Message 2	Message 3	Message 4	Message 5
G1	84	103	104	84	85
G2	67	42	43	64	64

Depending on the result above, the a array MG^1 can be found as shown bellow .

$$MG^1 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad \begin{matrix} \text{group 1} \\ \text{group 2} \end{matrix}$$

Because of the similarity of the results of MG0 and MG1, the algorithm will halt and will give the classification results .

Cluster 1 (G1) contains the messages 1, 4, and 5.

Cluster 2 (G2) contains the messages 2 and 3.

Note:G1 represent group 1and G2 represent group 2.

Conclusions and Recommendations

This paper provides a make use of k- mean algorithm for classify the messages that was encrypted using transposition system and simple substitution system after modifying Euclidean equation. The results of the classification tests were successful for all messages that encrypted in either substitution or transposition systems depending on cipher text only.

We recommend this technique to classify other systems, and use other techniques in diagnosis other cryptographic systems such as neural network and fuzzy system .

References

[1] Jiawei Han, "Data Mining: Concepts and Techniques", Second Edition , by Elsevier Inc ISBN 13: 978-1-55860-901-3,2006.

[2] Mehmed Kantardzic , "Data Mining: Concepts, Models, Methods, and Algorithms" , John Wiley & Sons, ISBN:0471228524, 2003 .

[3] Hore P., Hall L., Goldgof D., 2009, "A scalable framework for cluster ensembles", Pattern Recognition, Journal homepage: www.elsevier.com/ locate/pr .

[4] Klawonn F., Hoppner F., 2011, "Kernel Function from Fuzzy Clustering and Their Application, thesis, University of Tsukuba .

[5] Xu R. , "Survey of Clustering Algorithm", IEE , Transaction neural network , vol 16, no.3 .

[6] Kanade P., 2004, "Fuzzy Ants as a Clustering Concept", MSC. thesis, University of South Florida .

[7] William Stallings ,Cryptography and Network Security principles and Practices", printice Hall, 2011.

[8] Chebrolu S., Abraham A., Thomas J., 2004, "Feature deduction and ensemble design of intrusion detection system", Computer & Security , (www.elsevier.com/locate/cose) .

[9] Konheim, "Computer Security and Cryptography", John Wiley & sons, inc, publication, 2007.

[10] Willy Bruce, " Applied Cryptography, Second Edition: Protocols,Algorthms, and Source Code in C" , Wiley Computer Publishing , John Wiley & Sons, Inc.ISBN: 0471128457 Pub Date: 01/01/96 .

[11] Shneier Brouce, " Applied Cryptography", Wiley computer publishing, 1998 .

استخدام خوارزمية K-mean للعنقدة في تشخيص نظامي تشفير

رجاء صالح

المعهد الطبي التقني ، بغداد ، العراق

(تاريخ الاستلام: 24 / 3 /2013 ---- تاريخ القبول: 15 / 5 /2013)

الملخص

هذا البحث يقدم استخدام جديد لتقنية k- mean للعنقدة. استخدمت لتشخيص النظام الشفري للرسائل المشفرة اعتمادا على النص المشفر. ويعتبر تصنيف الرسالة المشفرة هو الخطوة الاولى التي يعملها محلل الشفرة عند عمله على تحليل الرسائل المشفرة. الطريقة المقترحة تصنفت الرسائل المشفرة بنظامين هما: التعويض البسيط والانتقالي باستخدام خوارزمية k-mean بعد تحوير معادلة اقليدس. وقد نجحت الطريقة في تصنيف كل الرسائل المشفرة باستخدام احد هذين النظامين والتي اجريت التجارب عليها .