ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

SITES

Smart And Innovative Technologies In Engineering And Sciences

Gyan Ganga College of Technology

Vol. 5, Special Issue 3, November 2016

Clustering Algorithm for Text Steganography

Mr. Sailesh .S. Iyer[1], Dr. Kamaljit Lakhtaria[2]

Sr. Lecturer, SKPIMCS-MCA

Associate Professor, RCC, Gujarat University, Ahmedabad, India

Abstract: Data Mining Techniques have been deployed in extracting solutions for Business problems. This paper is an attempt to explore the world of Data Mining techniques and how it can be effectively used in the field of Text Steganography for Information Exchange. This paper explores various DM Techniques like Clustering, Classification, Machine Learning, and Neural Networks to get an insight into the effectiveness of these techniques. This paper proposes an algorithm titled "K-Means Text Steganography" which uses k-Means clustering for embedding Text in Text. This algorithm has been found to be satisfying performance parameters like Similarity Measure and Embedding Capacity.

Keywords: Text Steganography, k-Means Clustering, Embedding Capacity, Similarity Measure, Machine Learning, Neural Networks, Classification, Data Mining Techniques, Information Exchange.

I. INTRODUCTION (DATA MINING & STEGANOGRAPHY)

Data Mining is an integration of many areas and has been practiced in various forms. Statistics, Machine Learning, Pattern Recognition, Visualization, Data Warehouse, High Performance Computing [6] etc. are some of the areas where Data Mining can be integrated to provide effective solutions.

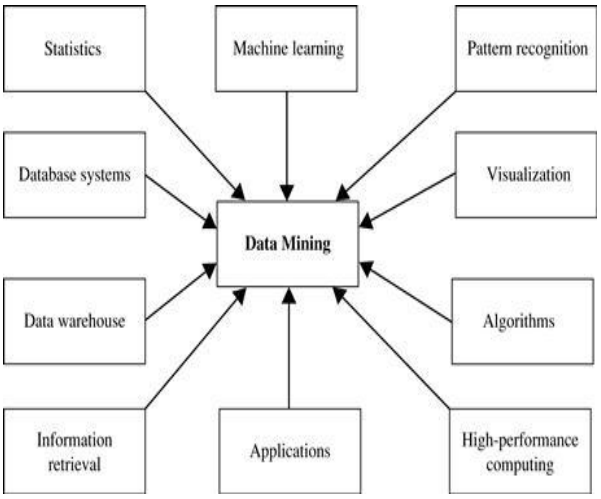

Fig.1 Data Mining confluences.

Data Mining Tasks can be primarily classified as Predictive and Descriptive. Classification technique which consists of Decision Tree Algorithms like ID3, J48 etc are used to predict the outcome for a given dataset. Also Naïve Bayes Classifier is a simple yet accurate method for prediction. Descriptive Data Mining consists of Clustering and Association Mining which are characterized by k Means, PAM and Apriori Algorithm.

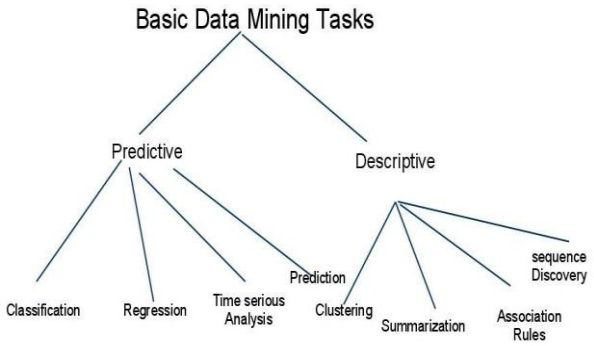

Fig.2. Data Mining Techniques at a glance

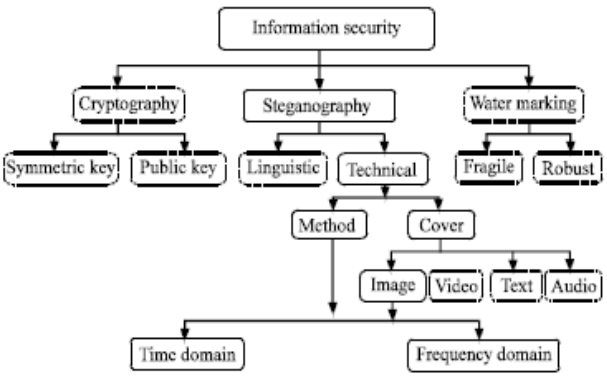

Fig.3. Information Security Classification

Information Security[5] can be classified into Cryptography, Steganography and Watermarking.
Steganography consists of components like secret message[7] to be hidden, hiding image or text (stego), method to hide text in stego and secret key to decrypt the

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

SITES

Smart And Innovative Technologies In Engineering And Sciences

Gyan Ganga College of Technology

Vol. 5, Special Issue 3, November 2016

hidden text. Steganography can be further divided into Image, Audio, Video and Text based on the method used. Our focus is going to be on various Steganography methods using Data Mining.

Given below in Fig. 4. is the blending of Data Mining and Steganalysis.

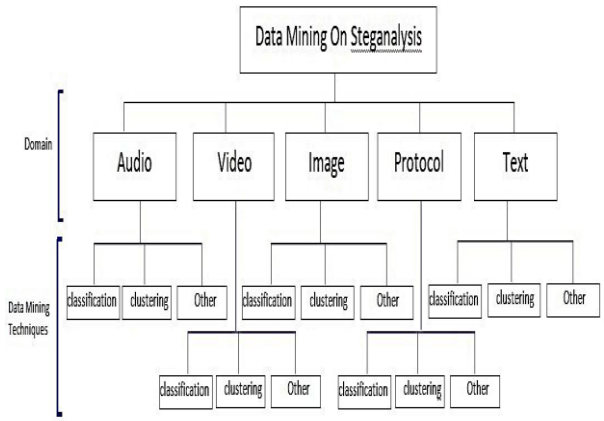

Fig. 4. Data Mining and Steganalysis

II. RELATED PAPER SUMMARY

The Literature survey conveys some very interesting observations. I have studied papers related to Classification, Clustering and Regression methods. The details of papers published from year 2010-2015 (6 years) according to Data Mining are listed in Table 1.

Table 1. Paper published year wise data

Year	Classification	Clustering	Regression
2010	21	12	15
2011	25	13	16
2012	28	17	21
2013	30	13	25
2014	33	13	26
2015	37	15	22

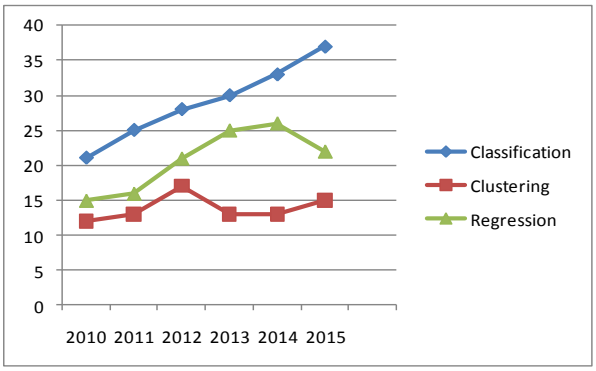

Fig. 5. Graphical Representation of Paper Publications

Papers published details show that more research work has been done on Classification method compared to Regression and Clustering. Figure 5 shows the Graphical Representation of above table. The line graph shows the blue line represented by Classification is far ahead of the other methods represented by green and red.

Text Steganography [2] methods have been suggested and some of the most popular methods have been mentioned below:

2.1 In Text Steganography by Hiding Information in Specific Character of Words approach, specific characters from some particular words are selected to hide the information. e.g. The first character of every alternative word hides the secret message.

2.2 Inter Word spacing method [2] was proposed to achieve the objective of reducing the size of objects created using Steganography. Experiments suggest that size reduction of objects is achieved but if the spaces are deleted then the message will be lost. One of the drawbacks of this method is that any alteration in spacing would leave this method ineffective.

2.3 Linguistic Steganography by context based substitution has been suggested [2]. Experiment results of the blind test demonstrate that the substitution can hardly generate syntax errors or unsuitable words, which implies the concealment of Steganography. This method has a relatively low embedding bit rate. Work can be extended to include imperfect filters, incomplete vocabulary, small scale dictionaries and limited training materials.

2.4 The proposed new Steganography method that uses a statistical compression technique called 'arithmetic coding'. This method has not been tested for long messages. This method has been tested for short messages which are proved by results of the experiments. Out of range messages are produced for large messages.

2.5 Text Rotation method is used in Excel to rotate the cell content by 1° if text and -1° if numeric. The cell text length is the key as if the cell length is 4 or less than 4, detection becomes very hard. If the text length is greater than 4 then it becomes easier to detect change in Stego text from Cover text.

2.6 Alphabet Pairing Algorithm [3] has been found to be very effective method for Embedding Text which has been found to be robust and qualifying on all parameters like embedding capacity, imperceptibility and time elapsed.

Table 2 consists of comparison of existing Text Steganography Algorithms which are compared in MATLAB environment.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

SITES

Smart And Innovative Technologies In Engineering And Sciences

Gyan Ganga College of Technology

Vol. 5, Special Issue 3, November 2016

Table-2. Comparative Analysis of Text Algorithms

STEGANOGRAPHY METHODS	ATTRIBUTES				
	DOMAIN	EMBEDDING CAPACITY	ROBUSTNESS	IMPERCEPTIBLY	INTEGRITY
Null Spaces	Text	Low	High	Low	Low
Synonym Substitution	Text	High	Low	High	Low
Context Based Equivalent Substitution	Text	Low	Low	High	Low
Frequency of Letters	Text	Low	Low	Low	Low
White Steg	Text	High	Low	High	Low
Reflection Symmetry	Text	High	High	Low	Low
Text Rotation Techniques	Text	Very High	High	High	High
Mixed Case Font	Text	Very High	High	Medium	High
Font Type	Text	Very High	High	Very High	High

III. REVIEW OF DATA MINING TECHNIQUES[1]

- **Regression-trees:**

This method uses decision trees for prediction of only numeric values. Each leaf has a numerical value, which is the average of all the training set values that the leaf, or rule, applies to. The relation between attributes is of prime importance here.

- **Model trees:**

Model trees offer regression trees as a mix with regression equations. The leaves of these trees contain regression equations rather than single predicted values. A model tree approximates continuous functions by several linear sub models.

- **Naïve Bayes classifier:**

This method proves to be simple yet provides accurate prediction. An adaptative classifier that can improve initial knowledge-based predictions for the class of a new instance by refining the model on the basis of the evidences provided by the whole history of processed cases.

- **Swarm Intelligence (SI):**

This method is used for numerical variables by training the system under the metaphor of the collective behavior of decentralized, self-organized systems, natural or artificial.

- **Clustering based on rules:** This provides grouping of homogeneous objects. Do not require number of classes as input. Can introduce prior expert knowledge as semantic bias. Guarantee interpretability of results and coherence with prior expert knowledge.

- **Simple linear regression:**

This method is used for predicting value of a quantitative variable for a new instance as a linear equation of a single numerical variable.

- **Multiple linear regression**:

This method is used to predict a new instance as a linear equation of several numerical variables.

- **Analysis of Variance:**

This method predicts a linear combination of one or two qualitative variables.

- **Time series:**

This methods predicts the value of a quantitative variable for a future instance as a linear combination of past values of the same variable.

- **Rule-based classifiers:**

This is a set of classification rules that can be used later to evaluate a new case and classify in a predefined set of classes.

- **Support Vector Machines (SVMs):**

They can provide discriminant functions to distinguish between two predefined classes that can be non-lineary separable.

- **Statistical clustering:**

This method is used to group of homogeneous objects. Might not require the number of classes. Can be efficient with big data sets. Sometimes difficult to understand the meaning of grouping provided.

IV. PROPOSED ALGORITHM

Proposed Algorithm has been suggested as a combination of Text Steganography and Data Mining technique. In this we are using a combination of k-Means Clustering

The proposed algorithm steps are as follows:

1. An ASCII table is prepared for all the alphabets and symbols.
2. The message to be embedded is decided.
3. k-Means clustering algorithm is used to form clusters of the same alphabets. Here we take k=6 i.e. number of clusters are 6. The embedding of the character takes place with an equivalent character from the same cluster.
4. The Euclidean distance between the first characters of the text to be embedded is found with the original text message.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

SITES

Smart And Innovative Technologies In Engineering And Sciences

Gyan Ganga College of Technology

Vol. 5, Special Issue 3, November 2016

5. The lowest distance obtained signifies that the character of the secret message would be embedded in this position from the cluster containing that character.
6. For decrypting the message, the same procedure is repeated in the reverse order.
7. The receiver of the message has the prior knowledge of the clusters. Since he/she receives the original message, the comparison of the first character with the cluster components of that cluster is done to which the character belongs.
8. The distance of the character with all the characters of the cluster is found out and the least distance is the embedded character.

This method has been tested and found to be satisfying two major criteria namely similarity measure i.e. Jaro Winkler distance which is the similarity measure if close to 1 and robustness.
Four experiments were performed where Text was embedded and measured for two parameters namely Average Jaro Winkler Score [19] i.e. Similarity measure and Average Embedding Capacity. The results of these experiments are tabulated and graphically explained.

Table 3. Experimental Results for Proposed Algorithm

Experiment	Av. % Capacity	Av. Jaro Score
Experiment-1	10.2	0.95
Experiment-2	10.6	0.98
Experiment-3	10.58	0.96
Experiment-4	10.75	0.99

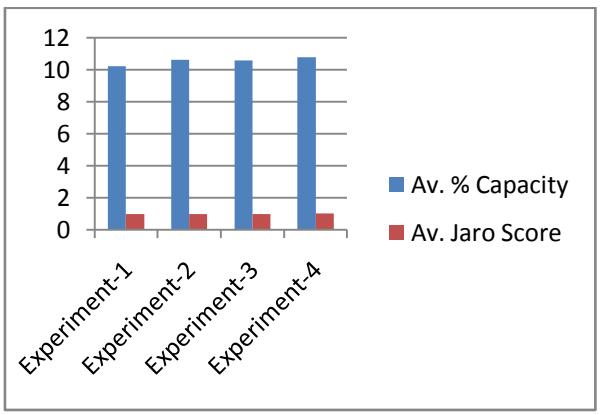

Fig 6 Graphical Representation of Experimental Values

These experimental results are within acceptable limits and perform better than some of the peer algorithms.

V. FUTURE WORK

This combination of Data Mining Clustering k-Means technique in Text Steganography has experimentally proved to be suitable. Future work may involve identifying still more Data Mining techniques which can be integrated with Steganography to give better results.

REFERENCES

[1] Rui Xi, Donald CW II 2005, Survey of Clustering Algorithms, IEEE Transaction on Neural Networks, 16: 645-678.
[2] Sailesh Iyer and Lakhtaria Kamaljit .I., "Practical Evaluation and Comparative Study of Text Steganography Algorithms", International Journal of Advance Engineering and Research Development e-ISSN 2348-4470 Volume 3 Issue 4, April 2016 pp 277-283.
[3] Sailesh Iyer and Lakhtaria Kamaljit .I., "New Robust and Secure Alphabet Pairing Text Steganography Algorithm", International Journal of Current Trends in Engineering and Research e-ISSN 2455-1392 Volume 2 Issue 7, July 2016 pp 15-21.
[4] Lakhtaria Kamaljit .I. "Protecting computer network with Encryption technique: A Study, "Ubiquitous Computing and Multimedia Applications. Springer Berlin Heidelberg 2011, 381-390..
[5] Lakhtaria, Kamaljit I. "Protecting computer network with encryption technique: A Study." International Conference on Ubiquitous Computing and Multimedia Applications. Springer Berlin Heidelberg, 2011.
[6] Kamaljit, I. Lakhtaria. "Prof. Bhaskar N. Patel,"Comparing Different Gateway Discovery Mechanism for Connectivity of Internet and MANET"."International Journal of Wireless Communication and Simulation 2.1 (2010): 209.
[7] Lakhtaria, Mr Kamaljit, et al. "Securing AODV for MANETs using Message Digest with Secret Key." Nettork Security & Its Applications (IJNSA) 1 (2009): 111-116.
[8] Abhishek Kolugiri, Sheikh Ghouse and Dr. P. Bhaskara Reddy, "Text Steganography Methods and its tools", International Journal of Advanced Scientific and Technical Research, March- April 2014.
[9] Reihane Saniei, Karim Faez, "The Capacity of Arithmetic Compression Based Text Steganography Method", 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP).
[10] Baharudin Osman, Roshidi Din, Tuan Zalizam Tuan Muda, Mohd. Nizam Omar, "A Performance of Embedding Process for Text Steganography Method", Recent Advances in Computer Science.
[11] Esra Satir, Hakan Isik, "A compression-based text steganography method", The Journal of Systems and Software 85 (2012) 2385– 2394.
[12] Prem Singh, Rajat Chaudhary and Ambika Agarwal, "A Novel Approach of Text Steganography based on null spaces", IOSR Journal of Computer Engineering (IOSRJCE) Volume 3, Issue 4 (July-Aug. 2012), PP 11-17.
[13] Indrajit Banerjee, Souvik Bhattacharya and Gautam Sanyal, "A Procedure of Text Steganography Using Indian Regional Language", I. J. Computer Network and Information Security, 2012, 8, 65-73.
[14] Sharon Rose Govada, Bonu Satish Kumar , Manjula Devarakonda and Meka James Stephen, "Text Steganography with Multi level Shielding, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.
[15] Anandaprova Majumder, Suvamoy Changder, "A Novel Approach for Text Steganography: Generating Text Summary using Reflection Symmetry", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.
[16] T. Moerland "Steganography and Steganalysis", May 15, 2003, www.liacs.nllhome/tmoerlaniprivtech.pdf.
[17] Fei Wang, Liusheng Huang, Zhili Chen, Wei Yang, Haibo Miao, "A Novel Text Steganography by Context-Based Equivalent Substitution", IEEE 2013.
[18] Sahil Kataria, Kavita Singh, Tarun Kumar, Mahendra Singh Nehra, "ECR(Encryption with Cover Text and Reordering) based Text Steganography", IEEE 2nd Intl. Conf. on IIP 2013.
[19] Reihane Saniei, Karim Fiez, "The Capacity of Arithmetic Compression Based Text Steganography Method", 8th Iranian Conference on Machine Vision and Image Processing (MVIP).
[20] Jaro-Winkler distance. 2015 Jan. Available from: http:// en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance