Performance Analysis on Text Steganalysis Method Using A Computational Intelligence Approach

Roshidi Din, Shafiz Affendi Mohd Yusof, Angela Amphawan, School of Computing
University Utara Malaysia, UUM Sintok, Kedah, Malaysia
roshidi@uum.edu.my, shafiz@uum.edu.my, angela@uum.edu.my

Hanizan Shaker Hussain, Hanafizah Yaacob, Nazuha Jamaludin, Department of Computer Science
Kolej Poly-Tech MARA, KPTM Cheras, Kuala Lumpur, Malaysia
hanizan@gapps.kptm.edu.my, hanafi@gapps.kptm.edu.my, nazuha@gapps.kptm.edu.my

Azman Samsudin
School of Computer Sciences
University Sains Malaysia, USM Gelugor, Pulau Pinang, Malaysia
azman@cs.usm.my

Abstract— In this paper, a critical view of the utilization of computational intelligence approach from the text steganalysis perspective is presented. This paper proposes a formalization of genetic algorithm method in order to detect hidden message on an analyzed text. Five metric parameters such as running time, fitness value, average mean probability, variance probability, and standard deviation probability were used to measure the detection performance between statistical methods and genetic algorithm methods. Experiments conducted using both methods showed that genetic algorithm method performs much better than statistical method, especially in detecting short analyzed texts. Thus, the findings showed that the genetic algorithm method on analyzed stego text is very promising. For future work, several significant factors such as dataset environment, searching process and types of fitness values through other intelligent methods of computational intelligence should be investigated.

Keywords—genetic algorithm method; computational intelligence; steganography, text steganalysis; fitness function value; performance evaluation; statistical method

I. INTRODUCTION

The Information Hiding field have played a significant role in e-secret communication channel and e-business applications such as e-national security [1], e-military [2], multimedia property [3], and authentication application [4]. Most of the new attacks in information hiding called steganalysis are derived by analyzing hidden protocol techniques in order to detect and extract the hidden messages. Steganalysis is aimed at discovering the hidden message from useless covert messages on analyzed medium communication channels [5, 6]. Until recently, there are two types of methods in detecting the hidden message namely digital steganalysis and text steganalysis. Most of the digital steganalysis methods which are proposed by steganalist can be divided into three major domains such as image steganalysis [7, 8, 9], audio steganalysis [10, 11], and video steganalysis [12, 13]. Besides that, text steganalysis method is used to discover the existence of hidden message on the features of characteristics, statistical probabilities or linguistic structures of natural language domain [14]. Currently, there are several detection methods in text steganalysis methods on natural language structure which is divided into six categories such as feature-based, statistical method, rhetorical, syntactical-based, lexical, and semantic-based [15, 16, 17, 18, 19, 20] as shown in Fig. 1.

From the theoretical point of view in text steganalysis [21], one of the challenges for steganalist that need to be considered is to justify whether the analyzed text may or may not contain hidden messages that is embedded into them. Therefore, steganalysis does not know exactly that the analyzed text is a stego text. This means that the steganalist cannot decide whether the analyzed text contains hidden message or not. This is because most of stego text “looks or feels” just like an innocent text even though it contains a hidden message. However, there are boundaries of solutions which text steganography method does not know how competent the steganalysis detectors are in measuring the steganography methods intelligently [21]. Thus, it seems like a hard challenge from steganalist’s point of view. It also raises an interesting question on deciding which intelligent method should be utilized in steganalytic system. As far as our knowledge is concerned, there has been little effort on the usage of text steganalytic system to utilize the computational intelligence method.

In the meantime, several studies have been applying intelligent methods such as genetic algorithm on digital steganalysis domain in order to find the hidden messages. Most of the results have proven that the utilization of genetic algorithm based system has performed exceptionally [22, 23].

Despite the different genetic algorithm methods on digital steganalysis domain that has been proposed, the potential of exploring the usage of genetic algorithm method in steganalysis for text domain is still under-utilized.
Hence, the motivation of this study is to utilize a genetic algorithm method in order to produce a good performance analysis on text steganalysis domain. Our primary goal is to analyze the performance between the genetic algorithm method and the statistical method in order to detect the hidden message on analyzed text. This paper is organized as follows. The next section deals with the discussion of intelligent methods of computational intelligence approach involved on text domain and digital steganalysis domain. This is then followed by the experimental design and the discussion of the results. Last, the concluding remarks will be given.

II. COMPUTATIONAL INTELLIGENCE ON TEXT STEGANALYSIS DOMAIN

Nowadays, there are four main intelligent methods of computational intelligence approach that have been applied in text processing domain. These are bayesian network, neural network, genetic algorithm, and fuzzy logic. All these methods are widely used in order to understand the human intelligence. Therefore, many steganalyst have applied computational intelligence approach on digital steganalysis domain. The discussion of these four intelligent methods on text and text steganalysis domain are summarized in TABLE I and TABLE II.

TABLE I. SUMMARY OF COMPUTATIONAL INTELLIGENCE METHODS ON TEXT DOMAIN.

<table>
<thead>
<tr>
<th>Computational Intelligence Method</th>
<th>Text Domain</th>
<th>Digitally Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Network</td>
<td>[41]</td>
<td>Simplest Moments of Freq. Distribution [42]</td>
</tr>
<tr>
<td>Neural Network</td>
<td>Characteristic Functions [43]</td>
<td>Principle of Diminishing Marginal Distributions [44]</td>
</tr>
<tr>
<td>Genetic Algorithm</td>
<td>Immune System [47]</td>
<td>Classification Rules [49]</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Dynamic Evolving Neural Fuzzy Inference System [50]</td>
<td>JPEG Steganalysis through Neuro-Fuzzy Inference System [51]</td>
</tr>
</tbody>
</table>

In TABLE I and TABLE II, computational intelligence methods on text domain and the usage of computational intelligence methods in digital steganalysis are summarized. It has been identified that genetic algorithm and fuzzy logic are the most capable method to apply on natural language environment. However, genetic algorithm is a bestfit computational intelligence method in producing a systematic rule for feature selection of solution and very powerful for optimization in text domain. In addition, it has been identified as well to pass the detection of current steganalytic systems on digital domain and also worked effectively on audio steganalysis and image steganalysis. Thus, this study believes that the genetic algorithm method can be applied intelligently and will perform well on text steganalysis domain.

TABLE II. SUMMARY OF COMPUTATIONAL INTELLIGENCE METHODS ON DIGITAL STEGANALYSIS DOMAIN.

<table>
<thead>
<tr>
<th>Domain Types</th>
<th>Digital Steganalysis Methods</th>
<th>Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Network</td>
<td>Bayesian Framework [41]</td>
<td>Statistical Moments of Freq. Distribution [42]</td>
</tr>
<tr>
<td>Neural Network</td>
<td>Characteristic Functions [43]</td>
<td>Principle of Diminishing Marginal Distributions [44]</td>
</tr>
<tr>
<td>Genetic Algorithm</td>
<td>Immune System [47]</td>
<td>Classification Rules [49]</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Dynamic Evolving Neural Fuzzy Inference System [50]</td>
<td>JPEG Steganalysis through Neuro-Fuzzy Inference System [51]</td>
</tr>
</tbody>
</table>

III. EXPERIMENTAL DESIGN

There are four stages involved in the experimental design such as the preprocessing of dataset used, text genome initialization, genetic algorithm detection engine and
performance evaluation. The experimental design of this study is illustrated in Fig. 2.

![Experimental Design](image)

**Fig. 2.** An experimental design for text steganalysis domain.

The first stage can be assumed as a setting up step which is to justify the collection of datasets used in this study. It utilizes

a hidden dataset of analyzed text [52] containing 10 lines with 893 bytes. It also uses a cover text from an established database known as Reuters News 21578. Besides, a detection process uses an established Oxford dictionary as a database corpus in order to verify the words or sentences on analyzed text. Actually, the system does not know what words or sentences included in each line of the sentences of an analyzed text before the detection process is done. Once the system receives an analyzed text, it will pass on an analyzed text during detecting process.

Then, second stage is an initialization of stego text. The role of this stage is to initialize the genes features of text genome. It is based on the rules features of the cover text Tc, hidden text Tm and the stego key used, Kn. In this process, each of the hidden text, Tm is embedded into cover text Tc using stego key used Kn which is known as embedding process. The purpose of this stage is to produce a stego text in order to be used as analyzed text Tn with the fitness function values fn. This information can be used as an input to detection process which utilizes a genetic algorithm method.

Next, the third stage consists of genetic algorithm detector. A development system of genetic algorithm detector uses JAVA Genetic Algorithm Programming (JGAP) language with interface under NetBeans IDE 6.9.1 environment. In this stage, the patterns and the behaviours of the analyzed text Tn are studied as much as it could due to understanding the analyzed text Tn itself through a database corpus using genetic algorithm detector. The searching process will end whenever genetic algorithm detector gets optimum estimated values namely cost function values of analyzed text Tn.

As a final stage, performance evaluation is to measure the score values of several parameters between genetic algorithm method and statistical method. Finally, the results of these parameters will be discussed.

**IV. RESULT AND DISCUSSION**

The purpose of this section is to analyze the results of the performance of statistical method versus genetic algorithm method during the extraction of a hidden message from a cover text. There are five metric parameters used in this experiment such as running time, best fitness value, mean distributions, variance distributions and standard deviation distributions. These metric parameters were investigated on 100 cover texts files of dataset. The score values of these five metric parameters are shown from Fig. 3 until Fig. 7.

**A. Running Time**

Based on Fig. 3, the results showed that the running time for both methods fluctuates between 3 and close to 60 seconds. However, the running time of statistical method is almost higher than the running time of genetic algorithm method. It may be influenced by the searching process of the corpus used in good environmental datasets. Only several texts of the genetic algorithm method had sudden spikes which were at the running time of texts 60, 98, and 99. Thus, it is found that the running time performance of genetic algorithm method is more stable than the statistical method for good environmental datasets.

**B. Fitness Values**

The results showed that the best fitness performance for both methods fluctuates nearly between 14 and 40 of cost value which is plotted in Fig. 4. The range of the best fitness performance using statistical method is between 18 and 40 of cost value whereas the range of best fitness performance using genetic algorithm method is between 14 and 28 of cost value. Therefore, it is found that the best fitness performance for genetic algorithm method is better than the statistical method for the good dataset environment.
The results showed that the mean probability of statistical based method sustainably fluctuates from close to 0.65 to 0.9 which is demonstrated in Fig. 5. However, the mean probability of genetic algorithm method remains constant closely at 1. Therefore, the mean probability of statistical method is lower than the mean probability of genetic algorithm method. The results of mean distribution measurements for statistical method and genetic algorithm method showed that the mean distribution for genetic algorithm method has a smaller range and is more accurate compared to the statistical method for good dataset environment.

D. Probability of Variance

Based on Fig. 6, the results showed that the variance probability of statistical method fluctuates from close to 0.01 to 0.023. However, the variance probability of genetic algorithm method remains unchanged closely to 0. Therefore, the variance probability of statistical method is higher than the variance probability of genetic algorithm method. The results of variance distribution measurement for statistical method and genetic algorithm method show that the variance distribution for genetic algorithm method has a smaller range and accurate than the statistical method for good dataset environment.

E. Probability of Standard Deviation

The results showed that the standard deviation probability of statistical based method fluctuates from close to 0.01 to 0.015 which is presented in Fig. 7. However, the standard deviation probability of genetic algorithm method remains unchanged at 0. Therefore, the standard deviation probability of statistical based method is higher than the standard deviation probability of genetic algorithm method. The results of standard deviation distribution measurement for statistical method and genetic algorithm method showed that the standard deviation distribution for genetic algorithm method has a smaller range and is more accurate compared to the statistical method for good dataset environment.

V. CONCLUSION

The primary contribution of this paper is to present the usage of computational intelligence approach works which in return would contribute to text steganalysis domain. The work presented here is among the earliest effort on text steganalysis domain using computational intelligence.
approach. It justifies that the genetic algorithm method performs better compared to statistical methods based on used parameter metrics in order to identify hidden messages on an analyzed text. For future work, it is suggested that several factors such as dataset environment, searching process, and types of fitness values should be considered to increase the strength of steganalysis on text domain. Besides, the use of other computational intelligence methods such as neural network, fuzzy evaluation, swarm optimization, and ant colony optimization should be explored and investigated.

ACKNOWLEDGMENT

We would like to thank Assoc. Prof. Dr. Huda Ibrahim, Dean of School of Computing, Universiti Utara Malaysia (SoC CAS UUM) and Director of Research and Innovation Management Centre (RIMC) for their moral support for the realization of this work. This research was financially supported by the High Performance Individual Research Grant National (Grant NO. 37011) under RIMC Grant, Universiti Utara Malaysia (UUM).

Fig. 7. A probability of standard deviation for statistical based versus GA based.

REFERENCES
