A Hybrid Approach of Fuzzy C-mean Clustering and Genetic Algorithm (GA) to Improve Intrusion Detection Rate

Kamaldeep Kaur¹, Navjot Kaur²

¹,²Computer Science and Engineering, Asracollege of Engineering & Technology, Bhawanigarh, India

Abstract: This paper describes a hybrid approach of Fuzzy C-means clustering and Genetic Algorithm (GA) is proposed that provides better accuracy & increases the intrusion detection rate. This approach provides better accuracy of detection as compared to K-means and FCM Clustering. With this proposed approach intrusion detection rate is improved considerably. A brief overview of a hybrid approach of genetic algorithm and fuzzy c-means clustering to improve anomaly or intrusion is presented. This paper proposes genetic algorithm and fuzzy c-means clustering to generate to detect intrusions. The goal of intrusion detection is to monitor network activities automatically, detect malicious attacks and to establish a proper architecture of the computer network security. We have been using fuzzy data mining techniques to extract patterns that represent normal behavior for intrusion detection. We describe a variety of modifications that we have made to the data mining algorithms in order to improve accuracy and efficiency.

Keywords: intrusion detection, clustering, fuzzy c-means clustering, genetic algorithm, Kddcup 99 Dataset

1. Introduction

An Intrusion Detection System is a type of security software that inspects all network activity and analyses it for any kind of malicious activities that violate computer security policy. With an increase in dependency on the internet, there is significant increase in the number of internet attacks. The challenges increases towards network security due to the introduction of new ways of attacks. An intrusion detection system (IDS) is a type of security software designed to automatically alert administrators when someone or something is trying to compromise information system through malicious activities or through security policy violations. The IDS device can be hardware, software or a combination of both that monitors the computer network against any unauthorized access.

2. Techniques of Intrusion Detection

2.1 Clustering

A cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “The process of organizing objects into groups whose members are similar in some way”.

2.1.1 Goal of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

2.2.2 Possible Applications

Clustering algorithms can be applied in many fields, for instance:

- Marketing: Finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- Biology: Classification of plants and animals given their features;
- Libraries: Book ordering;
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- City-planning: Identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: Clustering observed earthquake epicenters to identify dangerous zones;
- WWW: Document classification; clustering weblog data to discover groups of similar access patterns.

2.2.3 Fuzzy C-Means Clustering

The Algorithm Fuzzy C-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on minimization of the following objective function: Fuzzy-C means will tend to run slower than K-means, since it's actually doing more work. Each point is evaluated with each cluster, and more operations are involved in each evaluation. K-Means just needs to do a distance calculation, whereas fuzzy c means needs to do a full inverse-distance weighting.

Requirements

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;

International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

www.ijsr.net

Keywords: intrusion detection, clustering, fuzzy c-means clustering, genetic algorithm, Kddcup 99 Dataset

1. Introduction

An Intrusion Detection System is a type of security software that inspects all network activity and analyses it for any kind of malicious activities that violate computer security policy. With an increase in dependency on the internet, there is significant increase in the number of internet attacks. The challenges increases towards network security due to the introduction of new ways of attacks. An intrusion detection system (IDS) is a type of security software designed to automatically alert administrators when someone or something is trying to compromise information system through malicious activities or through security policy violations. The IDS device can be hardware, software or a combination of both that monitors the computer network against any unauthorized access.

2. Techniques of Intrusion Detection

2.1 Clustering

A cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “The process of organizing objects into groups whose members are similar in some way”.

2.1.1 Goal of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

2.2.2 Possible Applications

Clustering algorithms can be applied in many fields, for instance:

- Marketing: Finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- Biology: Classification of plants and animals given their features;
- Libraries: Book ordering;
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- City-planning: Identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: Clustering observed earthquake epicenters to identify dangerous zones;
- WWW: Document classification; clustering weblog data to discover groups of similar access patterns.

2.2.3 Fuzzy C-Means Clustering

The Algorithm Fuzzy C-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on minimization of the following objective function: Fuzzy-C means will tend to run slower than K-means, since it's actually doing more work. Each point is evaluated with each cluster, and more operations are involved in each evaluation. K-Means just needs to do a distance calculation, whereas fuzzy c means needs to do a full inverse-distance weighting.

Requirements

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
• discovering clusters with arbitrary shape;
• minimal requirements for domain knowledge to determine input parameters;
• ability to deal with noise and outliers;
• insensitivity to order of input records;
• high dimensionality;
• interpretability and usability.

2.2 Genetic Algorithm

Genetic algorithm use ideas based on the language of natural genetics and biological evolution. Genetic allows humans to contribute solution suggestions to the evolutionary process. Genetic algorithm finds application in computational science, engineering, economics, chemistry, manufacturing. Genetic algorithm requires two functions-
• Genetic representation- it can encode appearance, behaviour, physical qualities of individuals. Designing a good genetic representation is expensive and evolvable is a hard problem in evolutionary computation.
• Fitness function- it is a particular type of objective function that is used to summarize as a single figure of merit. It flow close a given design solution is to achieving the set aims.

2.2.1 Genetic Algorithm Advantages To Intrusion Detection Systems

The implementation of genetic algorithms offers many advantages to intrusion detection systems. The benefits of using genetic algorithms for intrusion detection can be summarized as:
• Genetic algorithms offer intrusion detection systems an intrinsic parallelism.
• Genetic algorithms are capable of working in multiple directions simultaneously. This makes them beneficial for analyzing the huge volumes of multi-dimensional data to be processed by an intrusion detection system.
• Genetic algorithms work with populations of solutions rather than a single solution. This makes them suitable for behaviour based intrusion detection, where the behaviour attributes may exhibit varying values;
• Genetic algorithms are highly re-trainable. Therefore, using genetic algorithms for intrusion detection will add to the adaptability of the system.
• Genetic algorithms evolve over time by using crossover and mutation. Property of evolving over time makes them a good choice for dynamic rule generation.

3. Fuzzy Logic

Fuzzy logic starts with and builds on a set of user-supplied human language rules. The fuzzy systems convert these rules to their mathematical equivalents. This simplifies the job of the system designer and the computer, and results in much more accurate representations of the way systems behave in the real world. Fuzzy logic include its simplicity and its flexibility. Fuzzy logic can handle problems with imprecise and incomplete data, and it can model nonlinear functions of arbitrary complexity. "If you don't have a good plant model, or if the system is changing, then fuzzy will produce a better solution than conventional control techniques," says Bob Varley.

A. Fuzzy Set- Fuzzy Set is any set that allows its members to have different grades of membership (membership function) in the interval [0,1]. A paradigm is a set of rules and regulations which defines boundaries and tells us what to do to be successful in solving problems within these boundaries.

B. Fuzzy Set Operations

1) Universe of Discourse - The Universe of Discourse is the range of all possible values for an input to a fuzzy system.
2) Fuzzy Set- A Fuzzy Set is any set that allows its members to have different grades of membership (membership function) in the interval [0,1].
3) Support- The Support of a fuzzy set F is the crisp set of all points in the Universe of Discourse U such that the membership function of F is non-zero.
4) Crossover point- The Crossover point of a fuzzy set is the element in U at which its membership function is 0.5.
5) Fuzzy Singleton- A Fuzzy singleton is a fuzzy set whose support is a single point in U with a membership function of one.

4. Methodology

Figure 1: Flow of Work

- Data
- Parametric Initialization
- Fitness calculation
- Maxima minima
- Mean
- Data Extraction
- Crossover
- Mutation
- Fit Value
- Best Value
- Plot
- Output
- Result

Data- Data is unprocessed form of information. All software is divided into two general categories: data and programs. Programs are collections of instructions for manipulating data.

Parametric Initialization- Assume the value of a parameter for the purpose of analysis.
• **Fitness Calculation** - After assuming the value of parameter for the purpose of analysis. Calculate the fitness of the parameters.

• **Maxima/Minima** - The maxima and minima are the largest and smallest value of the function, either within a given range or on the entire domain of a function.

• **Mean** - The mean is the average of the numbers: a calculated “central” value of a set of numbers.

• **Data Extraction** - Data extraction is the act or process of retrieving data out of data sources for further data processing or data storage.

• **Crossover** - Crossover is a process of taking more than one parent solutions and producing a child solution from them.

• **Mutation** - Mutation is a genetic operator used to maintain genetic diversity from one generation of a population of genetic algorithm chromosomes to the next.

• **Fit value** - The values for an output variable that have been predicted by a model fitted to a set of data.

• **Best value** - Tradeoff between price and performance that provides the greatest overall benefit under the specified selection criteria.

• **Plot** - Plot is a literary term that refers to how narrative points are arranged to make a story understandable to the reader or observer.

• **Output** - Output is defined as the act of producing something, the amount of something that is produced or the process in which something is delivered.

• **Results** - The definition of a result is how something ended or the outcome of some action.

### 5. Results

#### Table 1: Detection Rate of Data Pattern using FCM Clustering + Genetic Algorithm

<table>
<thead>
<tr>
<th>E-R (%)</th>
<th>FCM+GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73.54</td>
</tr>
<tr>
<td>2</td>
<td>93.79</td>
</tr>
<tr>
<td>3</td>
<td>68.10</td>
</tr>
<tr>
<td>4</td>
<td>78.54</td>
</tr>
<tr>
<td>5</td>
<td>91.01</td>
</tr>
<tr>
<td>6</td>
<td>97.30</td>
</tr>
<tr>
<td>7</td>
<td>87.73</td>
</tr>
<tr>
<td>8</td>
<td>70.07</td>
</tr>
<tr>
<td>9</td>
<td>70.20</td>
</tr>
<tr>
<td>10</td>
<td>91.02</td>
</tr>
<tr>
<td>Average</td>
<td>82.12</td>
</tr>
</tbody>
</table>

After the iteration got the 82.12% of elimination rate.

#### Table 2: Comparison of Detection Rate of Data Pattern using K-means, FCM clustering, FCM clustering + Genetic Algorithm

<table>
<thead>
<tr>
<th>E-R (%)</th>
<th>K-means</th>
<th>FCM</th>
<th>FCM+GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64.59</td>
<td>72.67</td>
<td>73.54</td>
</tr>
<tr>
<td>2</td>
<td>65.20</td>
<td>92.79</td>
<td>93.79</td>
</tr>
<tr>
<td>3</td>
<td>64.59</td>
<td>68.10</td>
<td>68.10</td>
</tr>
<tr>
<td>4</td>
<td>64.59</td>
<td>72.67</td>
<td>78.54</td>
</tr>
<tr>
<td>5</td>
<td>64.59</td>
<td>81.37</td>
<td>91.01</td>
</tr>
<tr>
<td>6</td>
<td>65.20</td>
<td>95.34</td>
<td>97.30</td>
</tr>
<tr>
<td>7</td>
<td>65.20</td>
<td>78.73</td>
<td>87.73</td>
</tr>
<tr>
<td>8</td>
<td>65.20</td>
<td>68.28</td>
<td>70.07</td>
</tr>
<tr>
<td>9</td>
<td>65.20</td>
<td>68.28</td>
<td>70.20</td>
</tr>
<tr>
<td>10</td>
<td>64.59</td>
<td>84.27</td>
<td>91.02</td>
</tr>
<tr>
<td>Average</td>
<td>64.90</td>
<td>80.77</td>
<td>82.12</td>
</tr>
</tbody>
</table>

Combination of FCM Clustering and Genetic Algorithm (GA) shows better results (82.12%) as compared to K-means (64.90 %) and FCM Clustering (80.77 %).

![Figure 2: Detection Rate of Data Pattern using FCM clustering + Genetic Algorithm](image)

Detection ratio means correctness in a model for detecting intrusion. Here also experimental result shows that the proposed algorithm performs better in term of correctness in detecting intrusion.

### 6. Conclusion

These days avoidance of security ruptures utilizing the current security innovations is unlikely. Therefore, interruption identification is an imperative segment in system security. Additionally, abuse identification strategy can’t recognize obscure assaults so the irregularity location system is utilized to distinguish these assaults. To enhance the precision rate of interruption discovery in irregularity based recognition information mining method is utilized. In this thesis a hybrid approach using FCM clustering and Genetic Algorithm are implemented. The detection rate of combination of FCM clustering and Genetic Algorithm is better than K-means, FCM clustering.
References


