International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

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

Kamaldeep Kaur[1], **Navjot Kaur**[2]

[1, 2]Computer 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 the 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 couldbe “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.

2.2.1Genetic 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

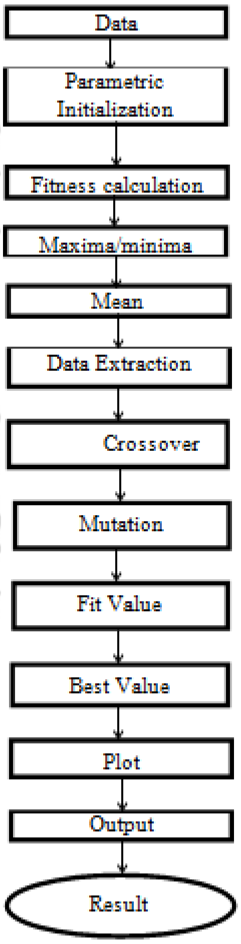

Figure 1: Flow of Work

- **Data**- 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 n**ext**
- **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

E-R (%)	FCM+GA
1	73.54
2	93.79
3	68.10
4	78.54
5	91.01
6	97.30
7	87.73
8	70.07
9	70.20
10	91.02
Average	82.12

After the iteration got the 82.12% of elimination rate

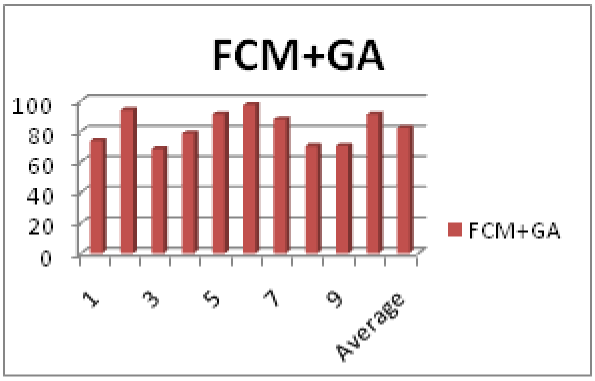

Figure 2: Detection Rate of Data Pattern using FCM clustering + Genetic Algorithm

We improve the intrusion detection rate with kddcup99 Dataset and technique FCM clustering and genetic algorithm

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

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

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 %)

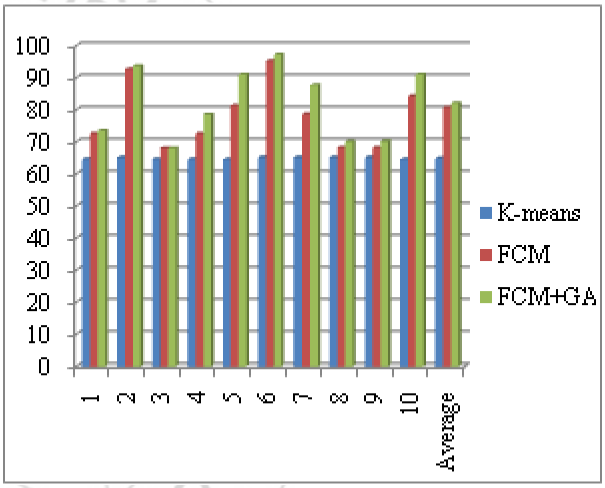

Figure 3: Comparison of Detection Rate of Data Pattern using K-means, FCM clustering, FCM clustering + Genetic Algorithm

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

[1] Hu, Liang, et al. (2015) "False positive elimination in intrusion detection based on clustering." Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on.IEEE, 2015.

[2] Lin, Wei-Chao, Shih-Wen Ke, and Chih-Fong Tsai. (2015) "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors." Knowledge-based systems 78 (2015): 13-21.

[3] Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita. (2014) "Network anomaly detection: methods, systems and tools." Communications Surveys & Tutorials, IEEE 16.1 (2014): 303-336.

[4] Kim, Gisung, Seungmin Lee, and Sehun Kim. (2014) "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection."Expert Systems with Applications 41.4 (2014): 1690-1700.

[5] Majeed, Parry Gowher, and Santosh Kumar. (2014) "Genetic algorithms in intrusion detection systems: A survey." International Journal of Innovation and Applied Studies 5.3 (2014): 233.

[6] Hassan, Mostaque Md. (2013) "Current studies on intrusion detection system, genetic algorithm and fuzzy logic." arXiv preprint arXiv:1304.3535 (2013).

[7] Katkar, Vijay D., and Siddhant Vijay Kulkarni. (2013) "Experiments on detection of Denial of Service attacks using ensemble of classifiers."Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on. IEEE, 2013.

[8] Lin, Shih-Wei, et al. (2012) "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection." Applied Soft Computing 12.10 (2012): 3285-3290.

[9] Galar, Mikel, et al. (2012) "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42.4 (2012): 463-484.

[10] Muniyandi, AmuthanPrabakar, R. Rajeswari, and R. Rajaram. (2012) "Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm." Procedia Engineering 30 (2012): 174-182.

[11] Shetty, Monali, and N. Shekokar. (2012) "Data Mining Techniques for Real Time Intrusion Detection Systems."International Journal of Scientific & Engineering Research 3.4 (2012).

[12] Mabu, Shingo, et al. (2011) "An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 41.1 (2011): 130-139.

[13] Smith, Michael R., and Tony Martinez. (2011) "Improving classification accuracy by identifying and removing instances that should be misclassified."Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011.

[14] Abdullah, B., et al. (2009) "Performance evaluation of a genetic algorithm based approach to network intrusion detection system." Proceedings of the International Conference on Aerospace Sciences and Aviation Technology.Military Technical College, 2009.

[15] Alserhani, Faeiz, et al. (2009) "Snort performance evaluation." Proceedings of Twenty Fifth UK Performance Engineering Workshop (UKPEW 2009), Leeds, UK. 2009.

[16] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. (2009) "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 15.

[17] Ghorbani, Ali A., Wei Lu, and MahbodTavallaee. (2009) “Network intrusion detection and prevention: concepts and techniques.” Vol. 47.Springer Science & Business Media, 2009.

[18] Dhanalakshmi, Y., and I. Ramesh Babu. (2008) "Intrusion detection using data mining along fuzzy logic and genetic algorithms."International Journal of Computer Science and Network Security 8.2 (2008): 27-32.

[19] Guntur, Gudlavalleru Guntur Rajamandry. (2008) "Modeling an intrusion detection system using data mining and genetic algorithms based on fuzzy logic." IJCSNS 8.7 (2008): 319.

[20] Alcalá, Rafael, JesØsAlcalá-Fdez, and Francisco Herrera. (2007) "A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection."Fuzzy Systems, IEEE Transactions on 15.4 (2007): 616-635.

[21] Banković, Zorana, et al. (2007) "Improving network security using genetic algorithm approach." Computers & Electrical Engineering 33.5 (2007): 438-451.

[22] Patcha, Animesh, and Jung-Min Park. (2007) "An overview of anomaly detection techniques: Existing solutions and latest technological trends." Computer networks 51.12 (2007): 3448-3470.

[23] Depren, Ozgur, et al. (2005) "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks." Expert systems with Applications 29.4 (2005): 713-722.

[24] Diaz-Gomez, Pedro A., and Dean F. Hougen. (2005) "Improved Off-Line Intrusion Detection Using a Genetic Algorithm." ICEIS (2). 2005.

[25] Gong, RenHui, Mohammad Zulkernine, and PurangAbolmaesumi. (2005) "A software implementation of a genetic algorithm based approach to network intrusion detection." Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing,

[26] Han, Sang-Jun, and Sung-Bae Cho. (2005) "Evolutionary neural networks for anomaly detection based on the behavior of a program."Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 36.3 (2005): 559-570.

[27] Abbes, Tarek, Adel Bouhoula, and MichaëlRusinowitch. (2004) "Protocol analysis in intrusion detection using decision tree."Information Technology: Coding and Computing, 2004. Proceedings.ITCC 2004.International Conference on.Vol. 1.IEEE, 2004.

[28] Li, Wei. (2004) "Using genetic algorithm for network intrusion detection."Proceedings of the United States Department of Energy Cyber Security Group (2004): 1-8.

[29] Qin, Min, and Kai Hwang. (2004) "Frequent episode rules for internet anomaly detection."Network Computing and Applications, 2004.(NCA 2004).

Proceedings.Third IEEE International Symposium on.IEEE, 2004.

[30] Ning, Peng, and SushilJajodia. (2003) "Intrusion detection techniques."The Internet Encyclopedia (2003).

[31] Dasgupta, Dipankar, and Fabio González. (2002) "An immunity-based technique to characterize intrusions in computer networks."Evolutionary Computation, IEEE Transactions on 6.3 (2002): 281-291.

[32] Dokas, Paul, et al. (2002) "Data mining for network intrusion detection." Proc. NSF Workshop on Next Generation Data Mining. 2002.

[33] AdhityaChittur., (2001) " ModelGeneratisson for an IntrusionDetection System Using GeneticAlgorithms"November 27, 2001Ossining High School Ossining

[34] Portnoy, Leonid. (2000) "Intrusion detection with unlabeled data using clustering." (2000).

[35] Lee, Wenke, and Salvatore J. Stolfo. (1998) "Data mining approaches for intrusion detection."Usenix security. 1998.

[36] Sundaram, Aurobindo. (1996) "An introduction to intrusion detection."Crossroads 2.4 (1996): 3-7.

[37] Heady, Richard, et al. (1990) "The architecture of a network level intrusion detection system." Department of Computer Science, College of Engineering, University of New Mexico, 1990.

[38] Tomek, Ivan. (1976) "An experiment with the edited nearest-neighbor rule."IEEE Transactions on Systems, Man, and Cybernetics 6 (1976): 448-452. 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks.SNPD/SAWN 2005.Sixth International Conference on.IEEE, 2005.

[39] F.Wikimedia,"Intrusiondetectionsystem,"http://en.wikipedia.org/wiki/Intrusion-detection system, Feb 2009.