Stream Data Mining Classification for an efficient Anomaly Intrusion Detection

Mr. Ravi Jethva[1], Mr. Kaushal Madhu[2]

[1]PG Student, Computer Engineering, LJIET, Ahmedabad, Gujarat, India
[2]Assistant Professor, Department of Computer Engineering, LJIET, Ahmedabad, Gujarat, India

ABSTRACT

Intrusion Detection System using Data Mining algorithms is a wide scope of Research. Wherein, various classification techniques can be used for a better classification of Known and Unknown type of attacks. An IDS (Intrusion Detection System) monitors the network traffic and then sends the suspicious activity reports to the System Administrator. In order to improve the efficiency of classification, various different techniques such as GNP, Fuzzy class Association, Hoeffding Tree Algorithm and Neural Network algorithm are used, but they fall short on some or other factors. So, in our work we've proposed and implemented a combination of Fuzzy GNP Association Rule Mining along with Probability Density Function which overcome the problems of sub-attribute utilization problem and is efficient in terms of time taken in classification as well as reduces False Alarms and improves Detection Ratio.

Keyword : *- Data Mining, IDS, Anomaly Detection, GNP, Fuzzy Rule Mining, Probability Density Function ;*

1. Intrusion Detection System

Computer Systems and Networks now a day are highly susceptive to Confidential and Sensitive data which are being received from and sent out to different networks. An IDS has been developed for such kind of network attacks which can access or alter confidential data or information. An IDS is a software application that monitors the underlying network for malicious activities. IDS is mainly focused on identifying possible incidents and to report them or to store them in the directory in order to prevent the same kind of incidents in future. Prevention techniques alone are not sufficient since it is impossible to have an absolute secure system, hence IDS is designed for identifying the attacks and to classify them for Historical Analysis purpose. New intrusions continually emerge and new techniques are needed to defend against those intrusions. IDS is the second line of defense, since it comes into the picture after an occurrence of an intrusion [1].

The main function of Data Mining techniques in Intrusion Detection is to classify the attacks as Normal or Malicious. Data Mining also provides Data summarization and Visualization which provides Historical Analysis. Data mining generally refers to the process of extracting useful rules from large stores of data. The recent rapid development in data mining contributes to developing wide variety of algorithms suitable for network-intrusion-detection problems. Intrusion detection can be thought of as a classification problem: we wish to classify each audit record into one of discrete sets of possible categories, normal or a particular kind of intrusion [12] [13].

2. Existing Work

As one of the most popular data mining methods for wide range of applications, association-rule mining is used to discover association rules or correlations among a set of attributes in a dataset. The relationship between datasets can be represented as association rules. An association rule is expressed by $X \Rightarrow Y$, where X and Y contain a set of attributes. This means that if a tuple satisfies X, it is also likely to satisfy Y. The most popular model for mining association rules from databases is the a priori algorithm [14]. This algorithm measures the importance of association rules with two factors: support and confidence. However, this algorithm may suffer from large computational complexity for rule extraction from a dense database.

In order to discover interesting rules from a dense database, genetic algorithm (GA) [15], [16] and genetic programming (GP) [17], [18] have been applied to association-rule mining. In the GA, the method evolves the rules during generations and individuals or population themselves represent the association relationships [19]. However, it is not easy for GA to extract enough number of interesting rules, because a rule is represented as an individual of GA. GP improves the interpretability of GA by replacing the gene structures with the tree structures, which enables higher representation ability of association rules [20], [21].

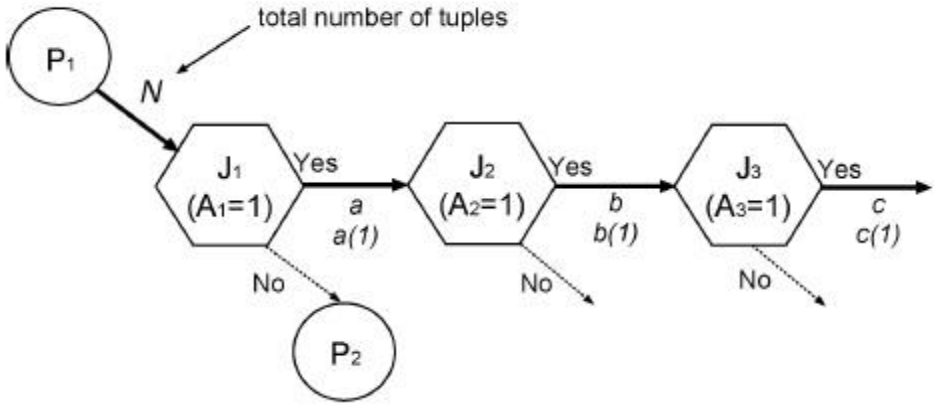

Fig.1 Node Transitions to find Class Association Rules [1]

As an extended evolutionary algorithm of GA and GP, genetic network programming (GNP) that represents its solutions using directed graph structures has been proposed [22]-[25].Originally, GNP is applied to dynamic problems based on inherent features of the graph structure such as reusability of nodes like automatically defined functions (ADFs) in GP, a compact structure without bloat and applicability to partially observable Markov decision process. However, to extend the applicable fields of GNP, an association-rule mining technique using GNP has been developed [26]. The advantage of association-rule mining methods is to extract sufficient number of important rules for user's purpose rather than to extract all the rules meeting the criteria. Like most of the existing association-rule mining algorithms, conventional association-rule mining based on GNP is able to extract rules with attributes of binary values. However, in real-world applications, databases are more likely to be composed of both binary and continuous values.

3. Proposed Work

In this work, concept of GNP based Fuzzy class association Rule Mining is introduced in detail. Where one more algorithm which is Probability Density function is applied to the output of GNP-Fyzzy rules. Using Fuzzy Rule Mining, the Information Loss problem has been overcome. GNP Structure is built up for the purpose of Rule extraction. Once the rules are extracted they are used for the classification.

For Misuse detection of IDS, normal and intrusion pattern rules are extracted directly from Training Data. Whereas, for Anomaly detection we're extracting as many normal-intrusion pattern rules as possible since this kind of detection will require high probability of data as an input to be predicted.

The use of Probability Density Function provides combination of discrete and continuous attributes in a rule and efficiently extracts many rules for classification which is practically useful for Real Network related databases. The Fitness function used in proposed method provides more rules with higher accuracy. In the proposed method, no experienced knowledge is required for deriving rules for classification. High Detection Rates are obtained for both Misuse and Anomaly detection.

GNP examines the attributes of tuples at judgment nodes and calculates the measurements of Association rules at processing nodes. The extracted fuzzy class-association rules are stored in a rule pool through generations. When an important rule is extracted by GNP, it is stored in the pool with its support, confidence, $\chi 2$ value, and the parameters of the fuzzy membership function. Occasionally, a fuzzy rule already stored in the pool would be extracted again. In that case, the membership function and $\chi 2$ value might be changed. If the fuzzy rule has higher $\chi 2$ value, it will replace the same old fuzzy rule in the pool along with its fuzzy parameters. Therefore, the pool is updated every generation and only important fuzzy rules with higher $\chi 2$ values and better-adapted fuzzy parameters are stored. Finally classification methods are proposed using FUZZY GNP.

3.1 Data Preprocessing

Extract data from DARBA dataset. It includes "list file," which identifies each network connection's time stamps, service type, source IP address, source port, destination IP address, destination port and the type of each attack. A preprocessor is a program that processes its input data to produce output that is used as input to another program. The output is said to be a preprocessed form of the input data. Preprocessing contains eliminating missing values from the Dataset.

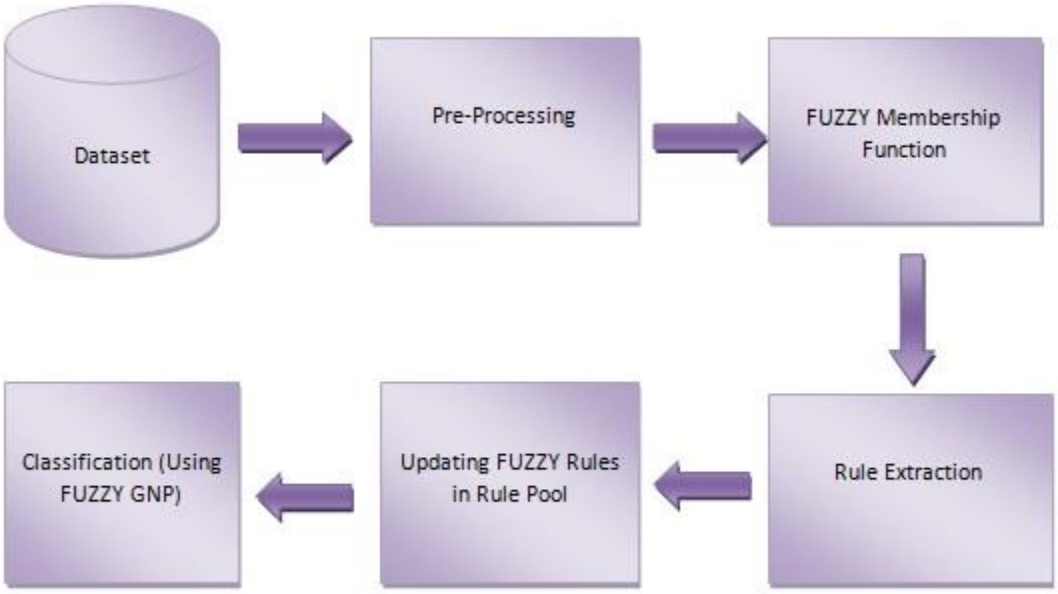

Fig.2 General Architecture for Rule Extraction using Fuzzy GNP

3.2 Fuzzy Membership Function

Network connection data have their own characteristics, such as discrete and continuous attributes, and these attribute values are important information that cannot be lost. We introduce a sub-attribute-utilization mechanism concerning binary, symbolic, and continuous attributes to keep the completeness of data information. Binary attributes are divided into two sub-attributes corresponding to judgment functions. The symbolic attribute was divided into several sub-attributes, while the continuous attribute was also divided into three sub-attributes concerning the values represented by linguistic terms (low, middle, and high) of fuzzy membership functions predefined for each continuous attribute. Each value of continuous attributes in the database is transformed into three linguistic terms (low, middle, and high). In other words, each continuous attribute is divided into three sub attributes with linguistic terms. A predefined membership function is assigned to each continuous attribute and the linguistic terms can be expressed by the membership function. The parameters α, β, and γ in a fuzzy membership function for attribute Ai are set as follows:

β = average value of attribute Ai in the database;
γ = the largest value of attribute Ai in the database;
$\alpha + \gamma = 2\beta$.

3.3 GNP Rule Extraction

GNP examines the attributes of tuples at judgment nodes and calculates the measurements of association rules at processing nodes. Judgment nodes judge the values of the assigned sub-attributes, e.g., Land= 1, Protocol=tcp, etc. The GNP-based fuzzy class-association rule mining with sub-attribute utilization successfully combines discrete and continuous values in a single rule. P1 is a processing node that serves as a starting point of class association rules and connects to a judgment node. The Yes-side of the judgment node is connected to another judgment node, while the No-side is connected to the next processing node. Judgment nodes shown here have the functions that examine the sub-attributes including both discrete and continuous attributes. Confidence and Support is used to extract important normal class-association rules. Rules with support values greater than threshold are extracted as important class-association rules of normal behaviors

3.4 Fitness and Genetic Function

Before defining the fitness of an individual, the fitness of extracted rule r is defined as follows:

$$fitness_r = \frac{Nt_c}{Nt} - \frac{Nn_i}{Nn}$$

Where,

Ntc - the number of connections correctly detected by rule r;

Nt - the number of connections in the training data;

Nni - the number of normal connections incorrectly detected by rule r;

Nn - the number of normal connections in the training data.

Each obtained rule is checked by the training data to get the fitness value. The scale of the fitness value is [–1, 1]. Higher fitness of a rule results in high DR and low positive false rate (PFR), which means the rate of incorrectly assigning normal connections to a intrusion class. On the other hand, lower fitness results in low DR and high PFR. When a rule is extracted by GNP, the overlap of the attributes between the rule and the already stored rules is checked to confirm whether the rule is newly extracted or not. The fitness of a GNP individual for network intrusion problems is defined by

$$F = \sum_{r \in R} \{w1 * fitness_r + w2 * \propto_{new} (r)\}$$

3.5 Classification using Probability Density Function

As this GNP-based fuzzy class-association approach is designed for databases containing both discrete and continuous attributes, specific classification methods are proposed. The matching degree between the continuous attribute *Ai* with linguistic term *Qi* in rule *r* in class *k* and the value *ai* of attribute *Ai* of a testing data is defined as

$$MatchDegree_K\ (Q_t, a_t) = FQ_t(a_t)$$

Where *FQi* represents the membership function for linguistic term *Qi* . Then, the matching between rule *r* in class *k* (including *p* continuous attributes and *q* discrete attributes) and new unlabeled connection *d* is defined as

$$MATCH_K\ (d, r) = \frac{1}{p+q} \left(\sum_{t \in CA} MatchDegree_k\ (Qt, at) + t\right)$$

Where

i - index of continuous attributes in rule r;

CA - set of suffixes of continuous attributes in rule r;

P - the number of continuous attributes in rule r;

q - the number of discrete attributes in rule r;

t - the number of matched discrete attributes with new unlabeled connection d in rule r.

Matchk (d, r) ranges from 0 to 1. If Matchk (d, r) equals 1.0, rule r matches connection data d completely. Else if Matchk (d, r) equals 0, rule r does not match connection data d at all. Then, the average matching between connection data d and all the rules in class k in a certain rule pool is defined as

$$MATCH_K\ (d) = \frac{1}{|R_k|} \sum_{r \in R_k} Match_k\ (d, r)$$

where R_k is a set of suffixes of the extracted rules in class k in the rule pool.

The classifier calculates the average matching between new connection data d_{new} and all the rules in the normal rule pool, i.e., $MATCH_n\ (d_{new})$ and the average matching between new connection data dnew and all the rules in the intrusion rule pool, i.e., $MATCH_i\ (d_{new})$. If $MATCH_n\ (d_{new}) \geq MATCH_i\ (d_{new})$, new connection data d_{new} is labeled as normal. If $MATCH_n\ (d_{new}) < MATCH_i\ (d_{new})$, new data d_{new} is labeled as intrusion.

4. Implementation Result

Fig.3 Extracted Rules by Min Support, Confidence

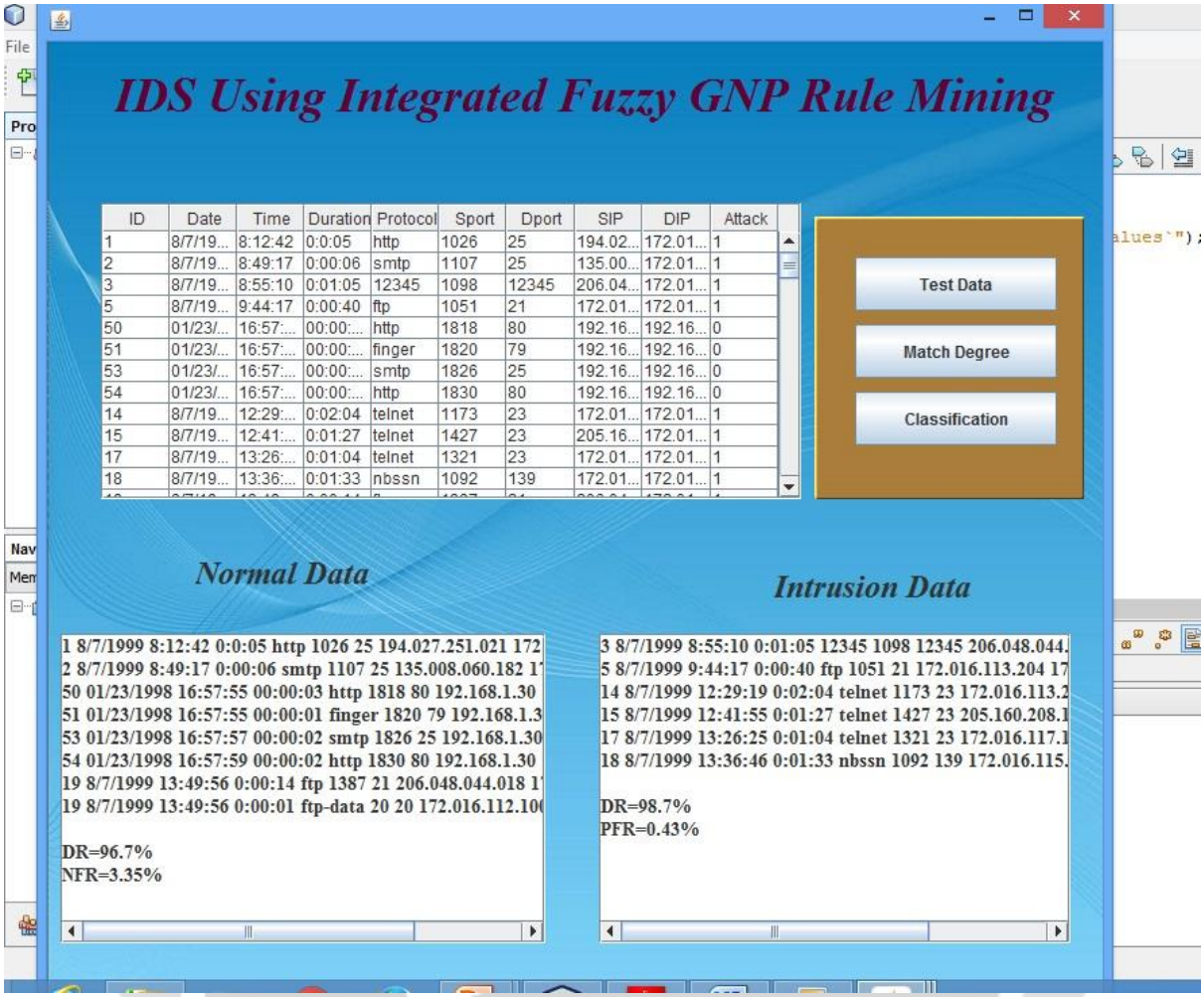

Fig.4 Data classified as Normal or Intrusion with DR and PFR

Figure shows extracted rules based on given minimum support and confidence values, and as an output, when the test data is entered from the system, the designed system classifies the data whether Normal or Intrusion. The classification along with the Detection rate, Positive False Ratio and Negative False Ratio is shown to prove that algorithm efficiently emerges so as to increase DR and decrease PFR.

Technique	Detection Ratio (DR) (%)	Positive False Rate (PFR) (%)
Genetic Programming	91.0	0.36
Fuzzy Rule Mining	95.8	0.48
Fuzzy GNP with PDF	98.7	0.43

Table 1 Comparison of techniques with DR and PFR measurements

The comparison in above table shows that Genetic Programming and Fuzzy Rule Mining when were implemented as a standalone technique, fallen short in DR and PFR respectively, when combined these 2 techniques along with Probability Density function shows that the DR has also increased along with less PFR.

5. Conclusions

This paper proposes an improved Fuzzy GNP using Probability Density Function which improves the overall Detection Ratio of an Intrusion Detection System, along with that it reduces False alarms and PFR. PDF helps to improve the efficiency when Anomaly Detection is to be taken into consideration and it emerges as an efficient algorithm to improve the performance of IDS.

6. References

[1] Shingo Mabu, Nanan Lu, Kaoru Shimanda and Kotaro Hirasawa "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming", IEEE Transactions On Systems, Man, And Cybernetics —Part C: Applications and Reviews, Vol. 41, No. 1, January 2011

[2] Shilpreet Sigh, Meenakshi Bansal, "A Survey on Intrusion Detection System in Data Mining", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume No. 2, Issue No. 6, June 2013

[3] Ketan Sanjay, Desale Chandrakant, Namdev Kumathekar,"Efficient Detection System using Stream Data Mining Classification Technique", 2015 International Conference on Computing Communication Control and Automation, 978-1-4799-68923/15 © 2015 IEEE

[4] Kailas Elekar, M.M. Waghmare, Amrit Priyadarshi "Use of rule base data mining algorithm for Intrusion Detection ", International Conference on Pervasive Computing (ICPC) , -1-4799-6272-3/15 (c)2015 IEEE

[5] Manish Kumar , Dr. M. Hanumanthappa "Intrusion Detection System using Stream Data Mining and Drift Detection Method ", IEEE – 31661 4th ICCCNT 2013

[6] Warusia Yassin, Azizol Abdullah " Signature-Based Anomaly Intrusion Detection using Integrated Data Mining Classifiers", 2014 International Symposium on Biometrics and Security Technologies (ISBAST), 978-1-4799-6444-4/14 ©2014 IEEE

[7] Hai Jin, Jianhua Sun, Hao Chen, Zongfen Han, " Efficient Intrusion Detection System using Stream Data Mining Classification Technique", Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems (FTDCS'04) Technologies (ISBAST), 0-7695-2118-5/04 © 2004 IEEE

[8] C. So, N. Mongkonchai, P. Aimtongkham, K. Wijitsopon and K. Rujirakul " An Evaluation of Data Mining Classification Models for Network Intrusion Detection", DICTAP 2014, Page No: 90 – 94

[9] J. Sanejunthichai, "Real Time Network Communication Data Analysis System in Order to Detect Internet Worm by Using Decision Tree Technique," In Proceedings of National Symposium on Applied Computing Technology and Information System, pp.118–124, 2011

[10] Damon Sotoudeh, Aijun An,CIKM'10 Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Pages 769-778, 2010

[11] Parekh S P, Madan B S, Tugnayat R M, "Approach for Intrusion Detection System Using Data Mining", Journal of Data Mining and Knowledge Discovery, ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 2, 2012, pp.-83-87.

[12] S. Manganaris, M. Christensen, D. Serkle, and K. Hermix, "A data mining analysis of rtid alarms," presented at the 2nd Int. Workshop Recent Adv. Intrusion Detect., West Lafayette, IN, 1999.

[13] D. E. Denning, "An intrusion detection model," IEEE Trans. Softw. Eng., vol. SE-13, no. 2, pp. 222–232, Feb. 1987.

[14] R. Agrawal and R. Srikant, "Fast algorithms formining association rules," in Proc. 20th VLDB Conf., Santiago, Chile, 1994, pp. 487–499.

[15] J. H. Holland, Adaptation in Natural and Artificial Systems. Ann Arbor, MI: Univ. Michigan Press, 1975.

[16] D. E. Goldberg, Genetic Algorithm in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[17] J. R. Koza, Genetic Programming, on the Programming of Computers by Means of Natural Selection. Cambridge, MA: MIT Press, 1992.

[18] J. R. Koza, Genetic Programming II, Automatic Discovery of Reusable Programs. Cambridge, MA: MIT Press, 1994.

[19] A. A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms. New York: Springer-Verlag, 2002.

[20] M. Crosbie and G. Spafford, "Applying genetic programming to intrusion detection," presented at the AAAI Fall Symp. Series, AAAI Press, Menlo Park, CA, Tech. Rep. FS-95-01, 1995..

[21] W. Lu and I. Traore, "Detecting new forms of network intrusion using genetic programming," Comput. Intell., vol. 20, no. 3, pp. 474–494, 2004.

[22] S. Mabu, K. Hirasawa, and J. Hu, "A graph-based evolutionary algorithm: Genetic network programming (GNP) and its extension using reinforcement learning," Evol. Comput., vol. 15, no. 3, pp. 369–398, 2007.

[23] T. Eguchi, K. Hirasawa, J. Hu, and N. Ota, "A study of evolutionary multiagent models based on symbiosis," IEEE Trans. Syst.,Man, Cybern. B, Cybern., vol. 36, no. 1, pp. 179–193, Feb. 2006.

[24] K. Hirasawa, T. Eguchi, J. Zhou, L. Yu, and S. Markon, "A doubledeck elevator group supervisory control system using genetic network programming," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 38, no. 4, pp. 535–550, Jul. 2008.

[25] K. Hirasawa, M. Okubo, H. Katagiri, J. Hu, and J. Murata, "Comparison between genetic network programming (GNP) and genetic programming (GP)," in Proc. Congr. Evol. Comput., 2001, pp. 1276–1282.

[26] K. Shimada, K. Hirasawa, and J. Hu, "Genetic network programming with acquisition mechanisms of association rules," J. Adv. Comput. Intell. Intell. Inf., vol. 10, no. 1, pp. 102–111, 2006.