A Context-free Grammar based Association Rule Mining Technique for Network Dataset

S Saranyadevi¹, R Murugeswari¹ and S Bathrinath²
¹Department of Computer Science and Engineering,
²Department of Mechanical Engineering, Kalasalingam Academy of Research and Education, Krishnankoil-626126, Tamilnadu, India.
saranyadevism@gmail.com

Abstract. Among various data mining concepts like prediction, clustering, classification, association and outlier discovery, association is a useful technique to extract the interesting relations among data items effectively. Association technique is applied in a number of applications like marketing, education, chemical, bioinformatics, computational linguistics and etc. The important purpose of association is to provide useful information of buying preferences of customers in supermarket in order to increase the sales opportunity, which is called as market-basket analysis. Till now there are many algorithms were developed, but the usage of formal grammars in association rule mining (ARM) is a latest technique to mine required data by means of grammars. In this paper ARM is performed using Context-free Grammar (CFG) – (ARM – Grammar) and the experiments are conducted on MATLAB 2017 software using network dataset, KDDCUP’99. Experimental outcomes prove that the proposed ARM – Grammar is effective than the traditional ARM approach.

1. Introduction
At present, a huge amount of data is generated and processed in various fields such as medical, banking, chemical, educational, market-basket analysis and etc, day by day. And so there is a necessity of managing such bulky and composite data using some effective techniques. Data mining is one of the ideal methods of extracting huge amount of existing and hidden data from databases. Data mining process consists of the following steps:
  • Identifying data from the database.
  • Preprocessing data to filter out the unwanted and missing values.
  • Choosing the data limits which are to be evaluated.
  • Selecting the required data from the large data, which is considered as the target data.
  • Inferring and displaying the final results, which are considered as knowledge.
The above steps are the important processes to extract the knowledge and also called as Knowledge Discovery from Databases (KDD). There are several data mining techniques such as prediction, clustering, association, outlier analysis and classification [1].
Association – Association [2] identifies the interesting relations among data patterns in the database. This method uses association rule to create relationship between different items of the dataset. The association rule is represented as,

\[ \text{Antecedent} \rightarrow \text{Consequent} \]

means that when the antecedent part occur, the consequent part will also occur with some relation.

a. Classification – Classification rules [3] are used to place the data objects into various predefined classes like safe and unsafe.

b. Clustering – Clustering is used to group the objects into number of clusters or groups [4].

c. Prediction – Prediction identifies the latest information from the collection of existing data [5].

d. Outlier Analysis – This technique spots out the exclusions in the data object which occur abnormally [6].

The purpose of ARM is to find relations among items in the given database. Apriori algorithm was the first algorithm developed for finding associations between items from large size of transaction data. The whole process was divided into two main steps: first step generates all combination of items whose support value was greater than the predefined minimum threshold (minsup) value, can be named as the large itemsets. Second step produced association rules from the large itemsets found from the first step whose calculated confidence value was greater than the minimum threshold value (minconf).

A Frequent-Pattern-tree structure (FP-tree) method of association rule generation was proposed in [7] without candidate itemset generation process. First the given database was scanned to find the items and their frequency count whose count was greater than the minsupport threshold, were considered as large 1-itemsets. Then the large 1-itemset is arranged in a descending order. Based on this order, the database was again scanned to construct the complete FP-Tree for all transactions in the database. The approach involved in the mining process is called as FP-Growth.

A weighted ARM approach was proposed in [9], which uses weight as a factor for mining rules. Weighted support and weighted confidence were described to extract the important association rules. The formal grammar named as CFG (Context-free Grammar) based ARM approach was proposed to mine rare association rules. In [8], a way of ARM with quantitative attributes from the transaction data was discussed. For each quantitative attributes, number of partitions was found out, then a set of consecutive integers were assigned for each of the identified attributes. The approach found the large itemsets whose support value was greater than the minsupport threshold and finally the association rules were generated from these identified large itemsets. There are some other works exist in the literature using the extensions of CFG with combination of fuzzy logic and optimization algorithms like GA and ACO [18 - 21].

The mining process using association rules is performed using various methods and techniques as discussed in literature, but this paper uses formal grammar named Context – free Grammar (CFG) to mine association rules. This way of mining is a different and new technique for mining interesting data items. The usage of CFG in ARM process is a recently known concept which is an enhancement of traditional ARM process. The main intention of this research is to mine association rules by means of grammar.

The rest of the paper is discussed as follows. Section 2 represents the proposed methodology including block diagram and the steps of the proposed method. Section 3 presents the experimental analysis part and the relevant result discussion with various performance metrics. Section 4 provides the conclusion part of the proposed work.
2. Proposed Methodology

In this paper, association rule mining is performed with the help of formal grammar, called, Context-free Grammar (CFG).

2.1 Definition of CFG

A Context-free Grammar is a quadruple, \( G = (\Sigma, N, P, S) \) where, \( \Sigma \) and \( N \) are the finite set of terminal and non-terminal symbols; \( P \) is the finite set of production rules of the form \( A \rightarrow \lambda \), where \( A \rightarrow \lambda, A \in N, \lambda \in (\Sigma \cup N)^* \); \( S \in N \) is the start symbol [10].

In this research work, Non–terminal symbols are the dataset attributes. Terminal symbols are the attribute values. Production rules are the grammar conditions or rules, based on the rules final categorization is made. Starting symbol is the first record of the identified dataset.

The principle of CFG is to perform all processes by itself using the defined rules or conditions by the specified grammar whereas in traditional ARM approach, all steps are performed by users using only Boolean values (0/1) in the input data.

2.2 Proposed Association Rule Mining Grammar (The ARM - Grammar)

Figure 1 represents the process flow of the proposed ARM – Grammar.

![Flow chart for CFG based ARM – The ARM – Grammar.](image-url)
The proposed concept of ARM – Grammar involves the following steps;
1. First the data is identified for mining purpose and the missing and irrelevant values are eliminated in the preprocessing step. Also the initial step defines the grammar conditions like finding the maximum and minimum values from dataset attributes as reference values for performing the mining process.
2. Then the grammar has set of defined rules to mine the interesting data. Based on the grammar rules the frequently occurring values are extracted.
3. And next the extracted values are placed in the available categories.
4. The most frequent value is placed in one category. And the following frequent values are placed in the subsequent categories.
5. The placement of instances in the categories is done based on the difference between the attribute values of data instances.

3. Experimental Analysis
The experiments are conducted using the processor, Intel Core-i5, along with 8 GB RAM and the Operating System is 32-bit and the experiments are executed in MATLAB 2017 software.

3.1 Dataset Description
The identified dataset for our experiment is the KDD CUP’99 dataset [11 – 14], from the UCI machine learning library. It is a well-known benchmark dataset which is used for intrusion detection. This dataset consists of an exemplar set of data to be inspected, with ‘good’ as normal connections and ‘bad’ connections called intrusions or attacks in a network environment. The whole dataset is divided as training set and test set. Training set has the size of 494021 instances and the test set has the size of 311029 instances. The dataset contains 41 attributes either continuous values or discrete values and output selector classes which are segregated into 5 major categories such as DOS (Denial of Service), Probe, R2L (Root 2 Local), U2R (User 2 Root) and Normal. The KDDCUP99 dataset is described in Table 1.

| Dataset   | Training Set Size | Test Set Size | Attributes | Number of Classes |
|-----------|-------------------|---------------|------------|-------------------|
| KDDCUP99  | 494021            | 311029        | 41         | 5                 |

3.2 Experimental Results and Discussion
Experimental results of KDDCUP’99 dataset provide an accurate categorization of instances in the relevant category. The main aim of utilizing this dataset is to recognize that the particular instance is belonging to either the normal or attack (ed) connection. The proposed algorithm outperforms than the existing method in terms of Accuracy, Error Rate, Precision, Recall, F-measure, False Detection [14-17]. The size of the network dataset considered in this research consists of 1500 data instances with 41 attributes.
3.3 Performance Metrics
Till now categorization of the dataset in Machine Learning (ML) field is performed into binary, multi-labeled, multi-class, hierarchical means. In this paper, multi-class categorization concept is applied where the input data is divided into ‘n’ number of non-overlapping classes or categories. The performance of our proposed method is evaluated by means of the evaluation metrics such as Accuracy, Error Rate, Precision, Recall, F-measure and False Detection is mentioned in Table 2.

Table 2. Performance Evaluation Metrics

| Performance Measure | Formula |
|---------------------|---------|
| Accuracy            | \[ \text{Accuracy} = \frac{\text{Correctly Categorized Instances}}{\text{Total Instances Categorized}} \] |
| Error Rate          | \[ \text{ErrorRate} = 100 - \text{Accuracy} \] |
| Precision           | \[ \text{Precision} = \frac{\text{Number of Appropriate Instances}}{\text{Total Number of Retrieved Instances}} \] |
| Recall              | \[ \text{Recall} = \frac{\text{Number of Appropriate Instances Retrieved}}{\text{Total Number of Appropriate Instances}} \] |
| F-measure           | \[ F = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \] |

3.3.1 Accuracy Comparison
Ratio of the total number of predictions that were correct is called as accuracy. Figure 2 represents the comparison of both traditional ARM and ARM – Grammar methods in terms of accuracy. From Figure 2 it is clear that the accuracy of proposed ARM – Grammar method is improved than existing traditional ARM method.

Figure 2. Accuracy – Comparison.
3.3.2 Error Rate

Error rate is the degree of errors occurred during the whole process. Figure 3 represents the error rate graph comparison of both proposed (ARM – Grammar) and existing (traditional ARM) methods. The error rate is reduced in the proposed method than the existing method.

![Error Rate Comparison](image)

**Figure 3.** Error Rate – Comparison.

3.3.3 Precision Comparison

Precision is the fraction of appropriate instances retrieved to the total number of appropriate and inappropriate instances retrieved. Figure 4 represents the comparison of both traditional ARM and ARM – Grammar methods in terms of Precision. From Figure 4 it is clear that the Precision of proposed ARM – Grammar method is improved than existing traditional ARM method.

![Precision Comparison](image)

**Figure 4.** Precision – Comparison.

3.3.4 Recall Comparison

Recall is the fraction of appropriate instances retrieved to the whole number of appropriate instances within the database. Figure 5 represents the comparison of both traditional ARM and ARM –
Grammar methods in terms of Recall. From Figure 5 it is clear that the Recall of proposed ARM – Grammar method got improved than existing traditional ARM method.

![Figure 5. Recall – Comparison.](image)

3.3.5 F-Measure Comparison
The harmonic mean of precision and recall measures is called F-Measure. Figure 6 represents the comparison of both traditional ARM and ARM – Grammar methods in terms of F-Measure. From Figure 6 it is clear that the F-Measure of proposed ARM – Grammar method got improved than existing traditional ARM method.

![Figure 6. F-Measure – Comparison.](image)

3.3.6 Accuracy Comparison with Existing Algorithms
The proposed algorithm is compared with existing algorithms such as Decision Tree CART, Decision Tree, Regression, Bayesian Classifier and Improved KNN. From the results it is proven that the ARM Grammar algorithm works more accurate than the existing methods and is displayed in Figure 7.
Figure 7. Comparison of Accuracy with Existing Algorithms.

The categorization concept is explained using the pictorial representation of each category which has been divided by both proposed (Figure 8) and existing (Figure 9) methods. From the two pie charts it is clear that the proposed method has eight classes as same as the dataset whereas the existing method has only seven classes. And the number of instances placed in each class using both the methods differ each other. The genuineness of our proposed algorithm has been explained using the performance evaluation metrics in the section 2.1.

3.3.7 Comparison of Dataset Categorization

Figure 8. Categorization of Instances using ARM – Grammar method (Proposed).
3.3.8 Observations

Although there are five classes of network connections in the dataset descriptions, we got eight unique selector class values in the KDDCUP’99 dataset namely Ipsweep, Multihop, Named, Normal, Smurf, Snmpgetattack, Xlock, Xsnoop. The attack Ipsweep comes under the main class Probing. Multihop comes under R2L class, further we have noticed that the attack Named, Xlock and Xsnoop have a similar behavior as Multihop. Hence all the four attacks were grouped together and placed in the same class, R2L. In the same way, Smurf comes under DOS class, and the Snmpgetattack has a similar behavior as Smurf. Hence both the attacks were combined together and placed in the same class DOS. The next one is the Normal connection which is free of attacks. As we have considered a part (15K instances) of the whole KDDCUP’99 dataset in order to run the experiments in which we notice that there are no attacks of class U2R were found. So the identified eight unique classes from our algorithm can be matched with the five defined classes as given in the dataset description and is mentioned in Table 3. Table A1 (in Appendix section) shows the number of different types of attacks obtained from the proposed and existing algorithms.

**Table 3. Attacks Categorization**

| S.No | Risk Factors  | Notation |
|------|---------------|----------|
| 1    | Ipsweep       | Probing  |
| 2    | Multihop      | R2L      |
| 3    | Named         | R2L      |
| 4    | Normal        | Normal   |
| 5    | Smurf         | DOS      |
| 6    | Snmpgetattack | DOS      |
| 7    | Xlock         | R2L      |
| 8    | Xsnoop        | R2L      |
4. Conclusion
This paper provides a methodology based on the combination of CFG based ARM for the detection of Probing, Remote to Local, Denial of Service and User to Root attacks. With the minimum false detection rate, the proposed algorithm proved to achieve maximum intrusion detections. A commendable intrusions percentage (approximately 94%) was detected by our approach. This makes the proposed algorithm more powerful to detect attacks efficiently in today’s innovative networking environment for information systems. Any hybrid evolutionary algorithms can be used to strengthen the proposed algorithm in order to detect the new attacks created by the intelligent attackers in the field of privacy and security.

5. Appendix

Table A1. Attacks Types

| Normal    | Probing   | DOS        | R2L        | U2R        |
|-----------|-----------|------------|------------|------------|
| Traditional ARM | ARM-Grammar | Traditional ARM | ARM-Grammar | Traditional ARM | ARM-Grammar | Traditional ARM | ARM-Grammar |
| Normal (523) | Normal (6) | Ipsweep (472) | Ipsweep (1209) | Smurf (13) | Smurf (13526) | Multihop (178) | Multihop (175) |
| Snmpgetattack (67) | Snmpgetattack (70) | Named (218) | Named (10) | NIL        |
| Xlock (13529) | Xlock (2)   | Named        | NIL        |
| Xsnoop (2)   |           |             |            |

| 523 | 6 | 472 | 1209 | 80 | 13506 | 13925 | 189 | 0 |

References
[1] Kaur M and Kang S 2016 Market Basket Analysis: Identify the changing trends of market data using association rule mining. Procedia computer science, 85, 78-85.
[2] Wu X, Kumar V, Quinlan J R, Ghosh J, Yang Q, Motod H and Zhou Z H 2008 Top 10 algorithms in data mining. Knowledge and information systems, 14 (1), 1-37.
[3] Saurkar A V, Bhujade V, Bhagat P and Khaparde A 2014 Review Paper on Various Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 4 (4), 98-101.
[4] Verma Sheenu and Bhatnagar Sakshi 2014 An Effective Dynamic Unsupervised Clustering Algorithmic Approach for Market Basket Analysis, International Journal of Enterprise Computing and Business Systems 4 (2).
[5] Bharati M and Ramageri M 2010 Data mining techniques and applications, Indian Journal of Computer Science and Engineering:1 (4):301-305.
[6] Hawkins D M 1980 Identification of outliers 11 London: Chapman and Hall.
[7] Lin K C, Liao I E, and Chen Z S 2011 An improved frequent pattern growth method for mining association rules. *Expert Systems with Applications*, 38 (5), 5154-5161.

[8] Sun K and Bai F 2008 Mining weighted association rules without preassigned weights. *IEEE transactions on knowledge and data engineering*, 20 (4), 489-495.

[9] Luna J M, Romero J R, Romero C and Ventura S 2014 Reducing gaps in quantitative association rules: A genetic programming free-parameter algorithm. *Integrated Computer-Aided Engineering*, 21 (4), 321-337.

[10] Hopcroft J E, Motwani R, and Ullman J D 2001 Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32 (1), 60-65.

http://archive.ics.uci.edu/ml/datasets/kdd+cup+1999+data

[11] Siddiqui M K and Naahid S 2013 Analysis of KDD CUP 99 dataset using clustering based data mining. *International Journal of Database Theory and Application*, 6 (5), 23-34.

[12] Ozgur A and Erdem H 2016 A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015. *PeerJ PrePrints*.

[13] Aggarwal P and Dahiya D 2016 Contribution of four class labeled attributes of KDD dataset on detection and false alarm rate for Intrusion Detection System. *Indian Journal of Science and Technology*, 9 (5).

[14] Lin W, Alvarez S A and Ruiz C 2002 Efficient adaptive-support association rule mining for recommender systems. *Data mining and knowledge discovery*, 6 (1), 83-105.

[15] Sokolova M and Lapalme G 2009 A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45 (4), 427-437.

[16] Choudhary T and Manikandan M S 2016 Robust Photoplethysmographic (PPG) based biometric authentication for wireless body area networks and m-health applications. In *Communication (NCC), 2016 Twenty Second National Conference on IEEE*. 1-6.

[17] Saranyadevi S, Murugeswari R and Bathrinath S 2019 Road risk assessment using fuzzy Context-free Grammar based Association Rule Miner. *Sādhanā*, 44 (6), 1-14.

[18] Saranyadevi S, Murugeswari R and Bathrinath S 2019 Association Rule Mining for Rainfall Prediction Using Fuzzy Context-free Grammar. *Journal of Advanced Research in Dynamical Control System*, 11(08) (2019), 850-858.

[19] Saranyadevi S, Murugeswari R, Bathrinath S and Sabitha M S 2019 Hybrid Association Rule Miner Using Probabilistic Context-Free Grammar and Ant Colony Optimization for Rainfall Prediction. In *International Conference on Intelligent Systems Design and Applications*, Springer, Cham. 683-695.

[20] Saranyadevi S, Murugeswari R and Bathrinath S 2019 Fuzzy Logic based Recommendation System for Road Accident Data Analysis. In *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing*, IEEE. 1-6.