International Journal of Computer Science and Informatics

Volume 3 | Issue 2 Article 13

October 2013

ASSOCIATION RULE MINING FOR KDD INTRUSION DETECTION DATA SET

ASIM DAS
Department of Computer Science, Pondicherry University, Pondicherry, India, asimdas407@gmail.com

S. SIVA SATHYA
Department of Computer Science, Pondicherry University, Pondicherry, India, ssivasathya@gmail.com

Recommended Citation
DAS, ASIM and SATHYA, S. SIVA (2013) "ASSOCIATION RULE MINING FOR KDD INTRUSION DETECTION DATA SET," *International Journal of Computer Science and Informatics*: Vol. 3 : Iss. 2 , Article 13.
DOI: 10.47893/IJCSI.2013.1133
Available at: https://www.interscience.in/ijcsi/vol3/iss2/13

ASSOCIATION RULE MINING FOR KDD INTRUSION DETECTION DATA SET

ASIM DAS[1] & S.SIVA SATHYA[2]

Department of Computer Science, Pondicherry University, Pondicherry, India
E-mail : { asimdas407,ssivasathya}@gmail.com

Abstract - Network intrusion detection includes a set of malicious actions that compromise the integrity, confidentiality and availability of information resources. Several techniques for mining rules from KDD intrusion detection dataset [10] enables to identify attacks in the network. But little research has been done to determine the association patterns that exist between the attributes in the dataset. This paper focuses on the association rule mining in KDD intrusion dataset. Since the dataset constitutes different kinds of data like binary, discrete & continuous data, same technique cannot be applied to determine the association patterns. Hence, this paper uses varying techniques for each type of data. The proposed method is used to generate attack rules that will detect the attacks in network audit data using anomaly detection. Rules are formed depending upon various attack types. For binary data, Apriori approach is used to eliminate the non-frequent item set from the rules and for discrete and continuous value the proposed techniques are used. The paper concludes with experimental results.

Keywords - *KDD CUP 99, intrusion detection system, Apriori approach, association rules mining.*

I. INTRODUCTION

KDD [10] dataset covers four major categories of attacks: Denial of Service (DOS), user-to-root (U2R), remote-to-local (R2L) and probing attack. KDD dataset is divided into labeled and unlabeled records. Each labeled record consisted of 41 attributes and one target value. Association rule mining is generally used to find the interesting rules from a large database depending upon the user defined support and confidence. In market basket analysis it finds relationship among the items present in the transactional database. A frequent item set is defined as one that occurs more frequently in the given data set than the user given support value. One more threshold confidence is used to restrict the association rules to a limited number. Confidence also includes the item sets having low support but from which high confidence rules may be generated. However the market basket analysis always works with binary values which means if item is present then the value is 1, otherwise value is 0(zero). But in reality or in KDD intrusion dataset, all the values may not be in binary. The dataset includes binary data, discrete and continuous data. Hence a generic technique will not work on all these data. So Association rule mining has to consider the type of data also. In general Association rule mining could be explained as follows:

Let I= {i_1, i_2, i_3, … i_n} be a set of items and T be a set of transactions. Each transactions is a set of items such that I_i subset of $_I$. An item set X is a set of items {i_1,i_2,……i_k}(1≤k≤n) such that X subset of I. An item set containing k number of items is called k-item set. An association rules is an implication of the form, A=>B, where A subset of I, B subset of I & A∩B=Ø. The rules A=>B holds in T with support s if s% of the transactions in T contain both A and B. Similarly the rule A=>B holds in T with confidence c if c% of the transactions in T support A also support B. To discover association rules from T having support and confidence greater than min_support and min_confidence.

Support, S (A=>B) = $\frac{\varphi(AUB)}{\varphi(N)}$

Confidence C (A=>B) =$\frac{\varphi(AUB)}{\varphi(A)}$

It was felt that identifying association patterns in KDD intrusion data set will help to design better Intrusion Detection System (IDS). Since the dataset is very large comprising of variety of data ranging from binary, discrete and continuous, these different association rule mining have been proposed in this paper. The rest of the paper is organized as follows: Section -2 describes the background and related work, Section-3 describes proposed association rule mining algorithm, Section-4 describes experimental result, Section-5 contains the conclusion followed by the reference section.

II. BACKGROUND AND RELATED WORK

M.Sulaiman khan, Maybin Muyeba and Frans Coenen[1] described weighted association rule mining from fuzzy data in their paper. Then some other proposed association rule mining for weighted value not necessarily binary value. The value should be continuous or discrete value to be presented in the database. In 2009 Flora S. Tsai [9] described network intrusion detection system using association rule mining in his paper. This helped to generated interesting rules from the KDD data set. The intrusion detection system contains various types of attacks. The KDD dataset contains variety of data starting from binary, discrete and continuous data. So, it is very difficult to generate rules for a particular attack using same approach. To find the association rules from the KDD dataset, different approaches should be applied for different kind of dataset as follows.

2.1. Association Rule Mining for Binary Value

For binary weighted value it is easy to find out the frequent item set. Apriori algorithm generates the frequent item set from a large number of data set. In association rules mining weights are considered as the highest priority. Apriori algorithm can be imagined as two steps. Firstly it generates candidate sets. Secondly, it prunes the entire non-frequent item set after each step using the minimum support and the weight of the item from the data base. Pruning process can eliminate many item sets which are not frequent.

2.2. Association Rule Mining for Continuous Value

If a particular attribute takes a value in the range [0…1] it is considered to be a continuous attribute in the Tanagra tool. This could be taken as Fuzzy data and hence fuzzy weighted Association rule mining as described in [5] could be used here. The weight of fuzzy data can be defined as Fuzzy Item Weight (FIW). Now Fuzzy Item set Transaction Weight (FITW) is the aggregate weights of all the fuzzy sets associated with the items in the item set present in a single transaction. From this FITW support and Confidence value can be calculated as per [5].

2.3. Association Rule Mining for Discrete value

If the range of values that an attribute in the data set can take is very large then normalization of the data becomes very difficult. The traditional approach to deal with this type of data in association analysis is to convert each value into a set of binary values. The discrete attributes are normalized i.e. we find a set of thresholds that can be used to convert the attributes into a categorical variable. This kind of normalization affects the accuracy of the rule generation technique which may lead to higher misclassification rate.

III. PROPOSED ASSOCIATION RULE MINING ALGORITHM FOR KDD IDS

3.1. Binary value

To find the association rules from binary data set already many algorithm have been proposed. For KDD data set Apriori algorithm is used to calculate the support and confidence value for various attributes. Apriori algorithm find the frequent item set from the database depending upon user define support and confidence. Here the rules can be generated for strong support and confidence value of the attributes. The item set having less support and confidence are automatically pruned. By this approach the number of generated rules can be restricted.

3.2. Continuous value

A continuous dataset D consists of transaction T={t_1,t_2,…t_n} associated with each item in I={i_1,i_2,i_3,….i_n}, which can contains the attributes weighted fuzzy value as L={l_1,l_2,…l_n}. Here we can assign weight for each attributes which are associated with i. Each attribute $t_i[i_j]$ is associated with several fuzzy set. The value of the attributes are between [0, 1] as fuzzy set. The "k th " weighted set for the "jth " item in the "ith" fuzzy transaction is given by $t_i[i_j[l_k[w]]]$ as mentioned in [1] .

Weighted Support(X) =

$$\frac{\sum_{i=1}^{n}\prod_{k=1}^{|L|}\left(\forall\left[i[l[w]]\right]\in X\right)ti\left[ij[lk[w]]\right]}{n}$$

Fuzzy Weighted Confidence is the ratio of sum both satisfying XUY to the sum of X. Here Z can be defined as Z=XUY.

Weighted Confidence(X=>Y)=

$$\sum_{i=1}^{n}\frac{\prod_{k=1}^{|Z|}\left(\forall[z[w]]\in Z\right)ti[zk[w]]}{\prod_{k=1}^{|X|}\left(\forall[i[w]]\in X\right)ti[xk[w]]}$$

3.3. Discrete value

As the value of the attributes falls in a wider range, it is very difficult to find the support value of such data without any fixed range. Unlike the traditional normalization method that is applied to the data set, in this approach, first the sum of the values for a particular attribute is calculated. Then the individual value is divided by the aggregate value. The newly calculated value is always between the range [0, 1].For one item set the methodology is same as the traditional one, but when we consider multiple item sets at a time, the minimum of the calculated value is taken as the support value. Now we can generalize it for any number of discrete values, where in the attributes should be paired to get the support and confidence value. This is explained as follows:

Let I= {i_1,i_2,….,i_n} be set of item sets in a transaction data base with transaction sets T = {t_1,t_2,…..t_n}. Each transaction contains the weighted value as W= {w_1,w_2,w_3,…..w_n} . Now support can be calculated as, Support(X)=$\frac{\mathrm{T[wj]}}{\sum_{j=1}^{|N|} T[wj]}$

For calculating support of more than two attributes the minimum value from the two item sets can be considered. The average should be taken from the minimum value to get the desired support. In table 1 it describes the item sets with the discrete values. In table 2 the calculated support value for each item set is shown. The support value for item A and item B, taken at a time, can be calculated by selecting minimum value from each row of the table. The set will contain the value as

T_{min}={Min(0.08,0.12),Min(0.08,0.16),Min(0.12,0.12), Min(0.16,0.16),Min(0.14,0.12),Min(0.08,0.06),Min(0 .14,0.16),Min(0.12,0.07)}

={0.08,0.08,0.12,0.16,0.12,0.06,0.14,0.07}

Support= 0.83/8=0.10

Table1: discrete data set to find the support.

TID	A	B	C	D	E
t1	12	33	34	18	78
t2	12	45	54	32	99
t3	18	34	32	23	32
t4	25	45	55	11	22
t5	21	33	56	27	90
t6	12	18	54	29	99
t7	21	45	32	34	97
t8	18	21	11	23	22

Table2: calculated support for each item set.

TID	A	B	C	D	E
t1	0.08	0.12	0.10	0.09	0.14
t2	0.08	0.16	0.16	0.16	0.18
t3	0.12	0.12	0.09	0.11	0.05
t4	0.16	0.16	0.16	0.05	0.04
t5	0.14	0.12	0.17	0.13	0.17
t6	0.08	0.06	0.16	0.14	0.18
t7	0.14	0.16	0.09	0.17	0.17
t8	0.12	0.07	0.03	0.11	0.04

3.4. Rule Generation Algorithm

After finding the frequent item set, support and confidence vale, rules have to be generated for each attack types. The following algorithm is general for any kind of data set. Here F contains the largest frequent item set. Min_supp defines the user define support and Min_conf defines the user defines confidence. RULE contains the desired rules generated from data set. The algorithm is as follows:

Algorithm:

```
1. Take the largest frequent item set F with Min_Supp and Min_Conf value.
2. Generate all possible subsets of F and store it in SUB.
3. Count SUPP and CONF value for each elements of SUB.
4. If (SUPP>=Min_Supp && CONF>=Min_Conf) then
    a. Choose the particular elements of SUB and store in RULE
    b. Generate various rules and store in RULE.
5. Else reject the particular element of SUB and go to step 3.
6. Return RULE.
7. End.
```

IV. EXPERIMENTAL RESULT

The rules are generated using the proposed approach and are showed in the table3. The result shows sample rules for a particular attack .The attributes with high confidence and support values are considered to generate the rules. The attributes having very low support and confidence value are rejected automatically.

Table3: Generated rules with support and confidence

Attack Type	**Attributes Selected**	**Rules**

1. Neptune Attack	29,30,34,35, 38,39	29=>30
		support=0.004
		confidence=0.065
		29,30=>34
		support=0.0002
		confidence=0.054
		29=>30,34
		support=0.0002
		confidence=0.0030
		34=>29,30
		support=0.0002
		confidence=0.0047
		30=>29,34
		support=0.0002
		confidence=0.05
		29,34=>30
		support=0.0002
		confidence=0.038
		29,34=>30,35
		support=0.000016
		confidence=0.0030
		30=>34
		support=0.003
		confidence=0.0714
		30,34=>29
		support=0.0002
		confidence=0.66

2. Apache2 Attack	25,26,27,28, 40,41	25=>26 support=0.28
		confidence=0.903
		25=>27 support=0.03
		confidence=0.096
		25=>28 support=0.03
		confidence=0.096
		26=>28 support=0.03
		confidence=0.044
		26=>40
		support=0.157
		confidence=0.234
		25,26,27
		support=0.01
		confidence=0.032
		25,26=>27,28
		support=0.00464
		confidence=0.0165

Attack Type	Attributes Selected	Rules

3. Back Attack	34,35,38,39,
	40,41
		34=>35
		support=0.005
		confidence=0.005
		34=>38
		support=0.00156
		confidence=0.00157
		34=>39
		support=0.00156
		confidence=0.00157
		35=>39
		support=0.0015
		confidence=0.2727
		35=>34
		support=0.005
		confidence=0.9090
		38=>34
		support=0.00156
		confidence=0.99
		39=>34
		support=0.00156
		confidence=0.99

4. Mail bomb Attack	34,35,40	34=>35
		support=0.0399confidence=0.042
		34=>40
		support=0.038
		confidence=0.040
		35=>40
		support=0.0048
		confidence=0.106
		35=>34
		support=0.039
		confidence=0.88
		34,35=>40
		support=0.00236
		confidence=0.059
		35=>34,40
		support=0.00236
		confidence=0.052
		40=>34,35
		support=0.00236
		confidence=0.042
		34,40=>35
		support=0.00236
		confidence=0.062

5. Smurf Attack	34,35,36	34=>36
		support=0.999
		confidence=0.999
		36=>34
		support=0.999
		confidence=0.999
		34=>35
		support=4.66E^{-5}
		confidence=0.00046
		35=>36
		support=4.66E^{-5}
		confidence=0.848
		35=>34,36
		support=4.25E^{-5}
		confidence=0.77

6. Ware master Attack	34,35,37	34=>35
		support=0.0238
		confidence=0.033
		35,37=>34
		support=0.00178
		confidence=0.0679
		34,35=>37
		support=0.000178
		confidence=0.074
		35=>37
		support=0.00262
		confidence=0.029
		37=>34
		support=0.0624
		confidence=0.088

7. Process table Attack	34,35,38,39,
	40,41
		34=>35
		support=0.062
		confidence=0.11
		34=>38
		support=0.2090
		confidence=0.37
		34=>39
		support=0.259
		confidence=0.46
		35=>39
		support=0.084
		confidence=0.49
		35=>40
		support=0.0802
		confidence=0.42

V. CONCLUSION

In this paper a generalized approach for mining the weighted association rules from KDD intrusion detection dataset with binary and fuzzy attributes has been proposed. Different techniques to count the support and confidence value from the dataset have been used. A number of association rules have been derived for each type of attack. The approach used here is effective to analyze the database containing discrete and continuous attributes with weighted settings. Here the poor rules having less support and confidence value have also been removed. The association rules thus generated will guide the IDS in evolving better rules to identify various attacks.

VI. ACKNOWLEDGEMENT

This work is a part of the AICTE funded project titled „Bio-inspired Intrusion Response System through feature relevance Analysis on Attack Classification, Under the Research Promotion Scheme (RPS), Ref. No: 8023/BOR/RID/RPS-59/2009-10.

REFERENCES

[1] M. Sulaiman Khan, Maybin Muyeba, Frans Coenen, ”Weighted Association Rule Mining from Binary and Fuzzy Data”.

[2] Tao, F., Murtagh, F., Farid, M, “Weighted Association Rule Mining Using Weighted Support and Significance Framework”. In: Proceedings of 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 661- 666, Washington DC (2003).

[3] Lu, S., Hu, H., Li, F, “Mining Weighted Association Rules”, Intelligent data Analysis Journal, 5(3), 211-255 (2001).

[4]. M. Sulaiman Khan, Maybin Muyeba, Frans Coenen, David Reid,”Mining Fuzzy Association Rules from Composite Items”.

[5] Michael Steinbach, PangNing Tan, Hui Xiong, Vipin Kumar,”Generalizing the Notion of Support”.

[6] R.Aggarwal, T. Imielinski, A. Swami, “Mining association rules between sets of items in very large database,” Proceedings of ACM SIGMOD conference, 1993.

[7] Y. Wang, Inyoung Kim, G. Mbateng, S.Y Ho, “A latest class modeling approach to detect network intrusion”, Computer Communications, 30, 93-100,2006.

[8] Agarwal, R. Srikant, “Fast Algorithms for Mining Association Rules”. In: “20 th VLDB Conference,pp.487-499(1994).

[9] Flora S. Tsai, “Network Intrusion Detection Using Association Rules”. In International Journal of Recent Trends in Engeerin, vol 2, No 2, November 2009.

[10] D. Newman,”KDD cup 1999 Data”, The UCI KDD Archive, Information and Computer Science, University of California, Iravin.