2009 Third International Conference on Network and System Security

A Comparison Between Rule Based and Association Rule Mining Algorithms

Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle
School of Computing Science, Faculty of Business and Informatics
Central Queensland University, Rockhampton,
QLD-4702, Australia.
E-mail: {m.mazid, s.ali, k.tickle@cqu.edu.au}

***Abstract*—Recently association rule mining algorithms are using to solve data mining problem in a popular manner. Rule based mining can be performed through either supervised learning or unsupervised learning techniques. Among the wide range of available approaches, it is always challenging to select the optimum algorithm for rule based mining task. The aim of this research is to compare the performance between the rule based classification and association rule mining algorithm based on their rule based classification performance and computational complexity. We consider PART (Partial Decision Tree) of classification algorithm and Apriori of association rule mining to compare their performance. DARPA (Defense Advanced Research Projects Agency) data is a well-known intrusion detection problem is also used to measure the performance of these two algorithms. In this comparison the training rules are compared with the predefined test sets. In terms of accuracy and computational complexity we observe Apriori is a better choice for rule based mining task.**

Keywords-Association Rule Mining, Classification, Apriori, Partial Decision Tree (PART), DARPA (Defense Advanced Research Projects Agency).

I. INTRODUCTION

Generating concise and accurate classifier by using Association Rule Mining is one of the central interests of data mining and machine learning researchers. So there is a great concern in finding ways to fine tune this technique to make it work more effectively. There are number of Association Rule Mining algorithms that are available to researchers such as Apriori [1], Predictive Apriori [2], Tertius [3], CLOSET [4], MAFIA [5], ELACT [6], CHARM [7], and many more. Various searching methods and different types of techniques have been used to polish up association algorithms such as best-first search [1-3], FP (Frequent Pattern)-Tree [4], depth-first search [5],etc. On the other hand, a number of techniques have been proposed to perform classification tasks such as Decision Trees [8], Rule Based Learning [9-10], Naïve-Bayes Classification [11], C4.5 [8], CN2 [9], RIPPER [10], etc. The common characteristic of these techniques is that they use a heuristic/greedy search. In greedy search, a single path is followed through tree nodes that take closest to the goal. But this searching method does not guarantee to find optimal solution each time. Thus the key objective of this research is to compare these two types of mining technique and to explore their similarities and dissimilarities between them.

Classification is one of the most significant areas in data mining. It is also known as pattern recognition, discrimination or prediction. Classification algorithms extract patterns by using data files with a set of labeled training examples. Classification algorithms are in the supervised learning group because they build a classifier/model based on supplied classes. It uses classifiers to predict classes. A classifier is a global model which generates a concise and eloquent description for each class by using attributes of data files [17]. A classifier is computed with decision functions $f(\mathbf{x}_i, \alpha_i) = y_i, \alpha_i \in \Lambda, \forall \langle \mathbf{x}_i, y_i \rangle \in D$ where D is a dataset with I independently identically distributed samples: $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)$; samples are set of feature vectors with length m; binary class $y_i \in \{+1, -1\}$ is the target value and Λ is a set of abstract parameters[19]. Classification algorithms have made significant inroads in the fields of bioinformatics, medical diagnosis, weather prediction, fraud detection, loan risk prediction, customer segmentation, target marketing, text classification and engineering fault detection. Because classification covers such a wide range of data mining, researchers have discovered many approaches such as rule-based classification, associative and instance-centric approaches, genetic algorithm based approaches, probability theory, etc. Each approach has at least one or more popular algorithms, for instance PART (Partial Decision Trees), NN (Neural Networks), SVM (Support Vector Machine), Naive Bayes, etc. In this research we will investigate the performance of PART, which is a rule based classification algorithm.

Association Rule Mining (ARM) is another important and substantial technique in machine learning. It is particularly important for extracting information from large databases and does so by discovering frequent itemsets and associating item relationships between or among items in a data file. Association Rule Mining is an unsupervised learning because it extracts rules without any prior class information. The credit for the development of Association Rule Mining is mostly attributed to Agrawal [1]. The Association rule is an expression of X→Y (read as 'if X then Y'), where X and Y are itemsets in a database D. The expression can be illustrated as 'if a customer buys item X then the customer is also likely to buy item Y' or 'if a patient is infected by disease X then the patient is likely to be infected by disease Y', and so on. The itemset of the left hand side of the arrow is called the antecedent and the itemset of the right hand side of the arrow is called the

978-0-7695-3838-9/09 $26.00 © 2009 IEEE
DOI 10.1109/NSS.2009.81

consequent. Each expression is called a rule. A rule can contain from two to an unrestricted number of items, with or without AND or OR operands. An item of a rule is selected from the frequent itemsets of the data file. Frequent itemsets are the items that occur more frequently. Basically, ARM follows two steps to produce rules from a data file: first, find all the frequent itemsets; second, generate strong association rules from the frequent itemsets. The best rules are selected on the basis of different types of interestingness measurement rules. ARM is a powerful exploratory technique with a wide range of applications including marketing policies, medical diagnosis, financial forecast, credit fraud detection and many other areas.

Our research compares two popular algorithms that represent two different data mining techniques. One is Apriori of Association Rule Mining and the other is PART algorithm of Classification. We have use DARPA [12] data sets for this experiment and Weka [13] data mining tools to generate rules. The rest of the paper is organized as follows. Section 2 briefly describes two algorithms which are Apriori and PART. Section 3 details of data that we have used in this experiment. Section 4 describes how the experiment performed. Section 5 is about experiment result discussion and finally Section 6 is the conclusion of this experiment.

II. Algorithm Descriptions

Apriori and PART are popular algorithms in the data mining community. They use different learning philosophy to produce rules. Apriori generates rules from unsupervised problem and PART generates rules from supervised problem. Brief descriptions of these two algorithms are summarized in below.

A. Apriori

The basic idea of the Apriori algorithm is to generate frequent itemsets for a given dataset and then scan those frequent itemset to distinguish most frequent items in this dataset. The process is iterative. Because generated frequent itemsets from a step can construct another itemsets by joining with previous frequent itemsets. Apriori is a confidence-based Association Rule Mining algorithm. The Confidence [18] is simply accuracy to evaluate the rules, produced by the algorithm. The rules are ranked according to the confidence value. If two or more rules share the same confidence then they are initially ordered using their support and secondly the time of discovery. Support is the percentage of a particular record in a data file. Basic steps for rule generation by Apriori are:

- Produce frequent itemsets of length 1
- Repeat until count of new frequent itemsets are zero(0)
  - From length *n* frequent itemsets, produce *n*+1 candidate itemsets.
  - Prune infrequent candidate of length *n*.
  - Count the support of each candidate by scanning the database.
  - Retaining the frequent candidate, eliminate the infrequent one.
- Produce Apriori rules based on support and confidence.

B. PART

The PART algorithm was developed by Frank and Witten [14]. PART is acronym of Partial Decision Tree. This name was chosen because this algorithm generates rules by repeatedly producing partial decision trees. This algorithm is derived from C4.5 and RIPPER algorithms. Both C4.5 and RIPPER use decision trees to generate the rule set. Unlike those rules, PART does not need to perform global optimization. In Global Optimization decision tree is been generated, then transformed it into a rule set and finally it simplifies the rules. For huge data sets, Global Optimization needs excessive time to generate rules. PART uses "separate-and-conquer" strategy [15]. In this strategy, one rule is generated at a time. Then it removes the instances covered by that rule and iteratively induces further rules for the remaining instances until none is left. In a multi-class setting this automatically leads to an ordered list of rules. An ordered list of rules is a type of classifier that is termed as 'decision list'. It differs from the standard approach in the way that each rule is created. To generate a single rule, a pruned decision tree is built for the current set of instances. Tree nodes with the largest coverage are made into a rule and the tree is discarded. This avoids hasty global generalization. PART is an ordinary decision tree that contains branches to undefined sub-trees [14].

III. Data Description

We have used the DARPA intrusion detection datasets from MIT Lincoln Lab [12]. These datasets provide researchers to evaluate performance of different IDS methodologies. DARPA99 [12] dataset represents data as rows. Each row comprises various information regarding pre and post login activities of users. The DARPA99 dataset contains 494,019 rows . Individual row consists of 41 attributes that describe about features of the network connection. These attributes can be grouped into 4 categories:

- **Basic attributes:** these attributes are about the packet header of a connection.
- **Content attributes:** these are about domain knowledge and some other information such as failed login attempts.
- **Time-based traffic attributes:** these attributes are relating to the time window, such as attempts to connect with same host within 2 second interval.
- **Host-based traffic attributes:** these attributes are about individual host history within a timeframe.

IV. Experimental Design

We have used Discriminant Analysis [16] to pre-process DARPA99 dataset. Performing Descriminant analysis we

have choose the first 8 attributes and the last attributes for our experiment. Out of 42 attributes, the extracted attributes for this experiment are bellow.

- Protocol type – different kind of protocol that have been used to establish the network connection.
- Service – destination network service such as http, ftp, smtp, etc.
- Land – whether connection source different or same port/host.
- Wrong fragment – number fragmentation that was incorrect
- Num_failed_logins – Number of failed logins.
- Logged_in – indication for successful user logged in.
- Root_shell – whether the root shell is obtained by the user.
- Is_guest_login – whether the logged in person is guest.

The last attribute "class" has 38 types of attacks which can be categorized into four types [16]. Those are Denial of service(dos), Remote to local(r2l), User to root(u2r) and Probe. We have modified the database to acquire better classification by eliminating less frequent data. We have sorted the data according to their class type and omitted those data that appeared less than 100 times out of 494019. Then we segment data into training set (first 90% of data file) and test set (10% of last portion of data file). After data modification, we have applied PART and Apriori algorithm, both are available in popular data mining tool WEKA 3.4 [13]. For Apriori, we choose first best fifty rules from the training phase. In PART analysis we have considered all the generated rules to fix up the model. . Finally we compared the testing data that are generated by the two algorithms. The computer configuration was Intel Core2 Duo CPU 2.33GHz and 4GB RAM.

V. Experimental Results

The main objective of this experiment is to perform the classification task with one of the ARM algorithms such as Apriori and compare the result with another classification algorithm such as PART. Table 1 and Table 2 shown the rules generated by the both algorithms. According to the generated rules, Apriori has picked up 4 classes out of 11 in rules. Classification accuracy between training data and test data is 87.5% for Apriori. In contrast, PART has detected more classes (7 classes) compared to Apriori. However accuracy of classification between Training and Test data by PART is inferior to Apriori i.e. 46.67%. In terms of computational time, Apriori shows the supremacy which is shown in Table 3.

TABLE 1: CLASSIFICATION WITH APRIORI

Class Name	Rules	Training Data	Test Data
smurf	logged_in=0	√	√
	root_shell=0	√	√
	is_guest_login=0	√	√
	logged_in=0 root_shell=0	√	√
	logged_in=0 is_guest_login=0	√	√
	root_shell=0 is_guest_login=0	√	√
	logged_in=0 root_shell=0 is_guest_login=0	√	√
neptune	logged_in=0	√	√
	root_shell=0	√	√
	is_guest_login=0	√	√
	logged_in=0 root_shell=0	√	√
	logged_in=0 is_guest_login=0	√	√
	root_shell=0 is_guest_login=0	√	√
	logged_in=0 root_shell=0. is_guest_login=0	√	√
normal	root_shell=0	√	√
	is_guest_login=0	√	√
	logged_in=1	√	√
	root_shell=0 is_guest_login=0	√	√
	root_shell=0 logged_in=1	√	√
	is_guest_login=0 logged_in=1	√	√
	root_shell=0 is_guest_login=0 logged_in=1	√	√
	is_guest_login=0 logged_in=0	×	√
	root_shell=0 is_guest_login=0 logged_in=0	×	√
	root_shell=0 logged_in=0	×	√

TABLE 2: CLASSIFICATION WITH PART

Class Name	Rules	Training Data	Test Data
ipsweep.	service = eco_i	√	√
neptune.	service = private AND protocol_type = tcp	√	√
	service = telnet AND num_failed_logins = 0	√	×
	protocol_type = tcp	√	×
normal.	logged_in = 1	√	√
	protocol_type = udp AND wrong_fragment = 0	√	√
	service = http	√	√
	service = urp_i	√	√
	service = urh_i	√	×
	protocol_type = tcp AND service = ftp_data	√	×
	protocol_type = tcp AND service = finger	√	×
pod.	service = ecr_i	√	×
satan.	service = other	√	×
smurf.	protocol_type = icmp AND service = ecr_i AND wrong_fragment = 0	√	√
teardrop.	protocol_type = udp	√	×

TABLE 3: COMPUTATIONAL TIME (MEASURED BY SEC.)

Algorithms	Training Data (444458instances)	Test Data (49384 instances)
Apriori	18	2
PART	129	17

VI. CONCLUSION

This is a comparative study between two rules based algorithms. We found that PART produces rules which are based on Class attribute. PART has picked up more classes than Apriori. However Apriori provides greater accuracy in terms of Training and Test data comparison. It requires substantially less computational time as well. But Apriori does not produce rules relating to class attribute each time. If such feature could be included with this algorithm, it will perform well for rule-based classification. We aim to continue this research with more rule-based data mining algorithm and few number of different types of data file.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining associations between sets of items in large databases," In Proc. of the ACM SIGMOD Int'l Conference on Management of Data. - Washington D.C. , 1993, pp. 207-216,

[2] T. Scheffer, "Finding association rules that trade support optimally against confidence," In proceedings of the 5th European Conference on Principles and Practice of Knowlege Discovery in Databases(PKDD'01). - Freiburg, Germany : Springer-Verlag, 2001, pp. 424-435.

[3] P.A. Flach and N. Lachiche, "Confirmation-guided discovery of first-order rules with Tertius," Kluwer Academic Publishers. - The Netherlands, 2001, Vol. 42. - pp. 61-95.

[4] J. Pei, J. Han, and R. Mao, "Closet: An efficient algorithm for mining frequent closed itemsets", In Proc. SIGMOD Int'l Workshop Data Mining and Knowledge Discovery. 2000.

[5] D. Burdick, Calimlim M. and J. Gehrke, "MAFIA: a maximal frequent itemset algorithm for transactional databases," In Intlernational conference on Data Engineering, 2001.

[6] M. J. Zaki, "Scalable algorithms for association mining," IEEE Transaction on Knowledge and Data Engineering, 2000. vol. 12.

[7] M. J. Zaki, and C.J. Hsiao, "CHARM: An efficient algorithm for closed association rule mining," Computer Science Department, Rensselaer Polytechnic Institute, New York, 1999.

[8] J. R. Quinlan, "C4.5: Program for machine learning," Morgan Kaufmann. 1992.

[9] P. Clark, and T. Niblitt, "The CN2 induction algorithm," Machine Learning, 1989, vol. 3, issue 1.

[10] W. Cohen, "Fast effective rule induction," In Proceedings of ICML-95, San Francisco, CA. 1995.

[11] R. Duda and P. Hart, "Pattern classification and scene analysis," Wiley, New York. 1973.

[12] Lincoln Labrotary. DARPA-99 Intrusion Detection Evaluation Data Set. Massachusetts Institute of technology, Retrieved January 12, 2009 from http://www.ll.mit.edu/ mission/communications/ist/ corpora/ideval/data/ 1999data.html

[13] I.H. Witten, and E. Frank, "Data mining: practical machine learning tool and technique with java implementation," Morgan Kaufmann, San Francisco, 2000.

[14] E. Frank and I.H. Witten, "Generating accurate rule sets without global optimisation," in Proceedings of the Fifteenth International Conference, Morgan Kaufmann, San Francisco, CA, 1998.

[15] G. Pagallo and D. Haussler, "Boolean feature discovery in empirical learning," Machine Learning, 1990, vol. 5(1), pp. 71-99.

[16] F.J emili, M. Zaghdoud and M. B. Ahmed, "A framework for an adaptive intrusion detection system using bayesian network," IEEEXplore, 2007. pp 66.

[17] A.B.M. S. Ali, and K. A. Smith, "On learning algorithm for classification," Applied Soft Computing, 2004, pp. 119-138.

[18] A.B.M. S. Ali, and S. A. Wasimi, "Data mining: methods and techniques," Thomson Publishers, Victoria, Australia, 2007.

[19] A.B.M. S. Ali, "Automated support vector learning algorithms," PhD Thesis, Monash University, Victoria, Australia. 2005.