International Journal of Computer Applications Technology and Research
Volume 2– Issue 3, 378 - 381, 2013

A survey of Anomaly Detection using Frequent Item Sets

Gaurav Shelke
RITS, Bhopal
Bhopal, India

Anurag Jain
RITS, Bhopal
Bhopal, India

Shubha Dubey
RITS, Bhopal
Bhopal, India

Abstract: Knowledge extraction is a process of filtering some informative knowledge from the database so that it can be used wide variety of applications and analysis. Due to this highly efficient algorithm is required for data mining and for accessing data from large datasets. In frequent item sets are produced from very big or huge data sets by applying some rules or association rule mining algorithms like Apriori technique, Partition method, Pincer-Search, Incremental, Border algorithm and many more, which take larger computing time to calculate all the frequent itemsets. As the network traffic increases we need an efficient system to monitor packet analysis of network flow data. Due to this frequent itemsets mining is basic problem in field of data mining and knowledge discovery. Here in this paper a brief survey of all the techniques related to frequent item sets generation has been given.

1. INTRODUCTION

Data mining can be achieved by Association, Classification, Clustering, Predictions, Sequential Patterns, and Similar Time Sequences. In Association, the relationship of a particular item in a data transaction on other items in the same transaction is used to predict patterns. In Classification, the methods are intended for learning different functions that map each item of the selected data into one of a predefined set of classes. Given the set of predefined classes, a number of attributes, and a ―learning (or training) set,‖ the classification methods can automatically predict the class of other unclassified data of the learning set. Cluster analysis takes ungrouped data and uses automatic techniques to put this data into groups. Clustering is unsupervised, and does not require a learning set. It shares a common methodological ground with Classification [1]. Prediction analysis is related to regression techniques. The key idea of prediction analysis is to discover the relationship between the dependent and independent variables, the relationship between the independent variables. Sequential Pattern analysis seeks to find similar patterns in data transaction over a business period. Existing algorithms for mining association rules are mainly worked on a binary database, termed as market basket database. On preparing the market basket database, every record of the original database is represented as a binary record where the fields are defined by a unique value of each attribute in the original database. The fields of this binary database are often termed as an item. For a database having a huge number of attributes and each attribute containing a lot of distinct values, the total number of items will be huge. Storing of this binary database, to be used by the rule mining algorithms, is one of the limitations of the existing algorithms. Another aspect of these algorithms is that they work in two phases. The first phase is for frequent item-set generation. Frequent item-sets are detected from all-possible item-sets by using a measure called support count (SUP) and a user defined parameter called minimum support. Support count of an item set is defined by the number of records in the database that contains all the items of that set. If the value of minimum support is too high, number of frequent item sets generated will be less, and thereby resulting in generation of few rules. Again, if the value is too small, then almost all possible item sets will become frequent and thus a huge number of rules may be generated. Selecting better rules from them may be another problem. After detecting the frequent item-sets in the first phase, the second phase generates the rules using another user defined parameter called minimum confidence [2] and [3-5].

2. RELATED WORK

Association Rules

Association rules are if and then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). Association rule is expressed as X=>Y, where X is the antecedent and Y is the consequent. Each association rule has two quality measurements, support and confidence. Support implies frequency of occurring patterns, and confidence means the strength of implication [1-3] and [9]. Associations: itemsets and sets of rule the result of mining transaction data in rules are associations. Conceptually, associations are sets of objects describing the relationship between some items (e.g., as an itemset or a rule) which have assigned values for different measures of quality. Such measures can be measures of significance (e.g., support), or measures of interestingness (e.g., confidence, lift), or other measures (e.g., revenue covered by the association). All types of association have a common functionality in rules comprising the following methods:

• Summary to give a short overview of the set and inspect to display individual associations,

• Length () for getting the number of elements in the set,

• Items for getting for each association a set of items involved in the association (e.g., the union of the items in the LHS and the RHS for each rule),

• sorting the set using the values of different quality measures (sort),

• Subset extraction,

• Set operations (union, intersect and set equal), and

• Matching elements from two sets (match),

• Write for writing associations to disk in human readable form. To save and load associations in compact form, use save and load from the base package.

The associations currently implemented in package a rules are sets of itemsets (e.g., used for frequent itemsets of their closed or maximal subset) and sets of rules (e.g., association rules). Both classes, itemsets and rules, directly extend the virtual class associations and provide the functionality described. Class itemsets contains one item Matrix object to store the items as a binary matrix where each row in the matrix represents an itemset. In addition, it may contain transaction ID lists as an object of class tidLists. Note that when representing transactions, tidLists store for each item a transaction list, but here store for each itemset a list of transaction IDs in which the itemsets appears. Such lists are currently only returned by éclat [4-6]. Class rules consists of two itemMatrix objects representing the left-hand-side (LHS) and the right-hand- side (RHS) of the rules, respectively. The items in the associations and the quality measures can be accessed and manipulated in a safe way using access or replace methods for items, lhs, rhs, and quality. In addition the association classes have built-in validity checking which ensures that all elements have compatible dimensions. It is simple to add new quality measures to existing associations. Since the quality slot holds a data frame, additional columns with new quality measures can be added. These new measures can then be used to sort or select associations using sort () or subset (). Adding a new type of associations to a rules is straightforward as well [6-8].

Rules interestingness measures

The aim of the association rules is to reveal interesting relations between data. For that reason certain are used which evaluate the level of importance of each rule. These are:

Confidence: The confidence of an association rule is the proportion of the isolates that are covered by the LHS of the rule that are also covered by the RHS. Values of confidence near value 1 are expected for an important association rule.

Support: The support of an association rule is the proportion of the isolates covered by LHS and RHS among the total number of isolates. Support can be considered as an indication of how often a rule occurs in a data set and as a consequence how significant a rule is.

Coverage: The coverage of an association rule is the proportion of isolates in the data that have the attribute values or items specified on the LHS of the rule. Values of coverage near value 1 are expected for an important association rule.

Leverage: The leverage of an association rule is the proportion of additional isolates covered by both the LHS and RHS above those expected if the LHS and RHS were independent of each other. Leverage takes values inside [-1, 1]. Values equal or under value 0, indicate a strong independence between LHS and RHS. On the other hand values near 1 are expected for an important association rule.

Lift: The lift of an association rule is the confidence divided by the proportion of all isolates that are covered by the RHS. This is a measure of the importance of the association that is independent of coverage [7] and [9-10].

Ignasi Paredes-Oliva, proposed an efficient technique of classifying frequent patterns on the basis of traffic patterns [16], here in this paper based on elegant combination of frequent item-set mining with decision tree learning.

Farah Hanna AL-Zawaidah and Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona et. al. presented a novel association rule mining approach that can efficiently discover the association rules in large databases. The proposed approach is derived from the conventional Apriori approach with features added to improve data mining performance. They had performed extensive experiments and compared the performance of the algorithm with existing algorithms found in the literature. They developed a visualization module to provide users the useful information regarding the database to be mined and to help the user manage and understand the association rules. Future work includes:

1) Applying the proposed algorithm to more extensive empirical evaluation; 2) Applying the developed approach to real data like retail sales transaction and medical transactions to confirm the experimental results in the real life domain; 3) Mining multidimensional association rules from relational databases and data warehouses (these rules involve more than one dimension or predicate, e.g. rules relating what a customer shopper buy as well as shopper's occupation); 4) Mining multilevel association rules from transaction databases [1].

Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K et. al. is to find all the possible optimized rules from given data set using genetic algorithm. The rule generated by association rule mining algorithms like priori, partition, pincer search, incremental, border algorithm etc, does not consider negation occurrence of the attribute in them and also these rules have only one attribute in the consequent part. By using Genetic Algorithm (GAs) the system can

predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. The major advantage of using GAs in the discovery of prediction rules is that they perform global search and its complexity is less compared to other algorithms as the genetic algorithm is based on the greedy approach. They have dealt with a challenging association rule mining problem of finding optimized association rules. The frequent itemsets are generated using the Apriori association rule mining algorithm. The genetic algorithm has been applied on the generated frequent itemsets to generate the rules containing positive attributes, the negation of the attributes with the consequent part consists of single attribute and more than one attribute. The results reported in this paper are very promising since the discovered rules are of optimized rules [3].

Peter P. Wakabi Waiswa and Dr. Venansius Baryamureeba et. al. present a Pareto based multi objective evolutionary algorithm rule mining method based on genetic algorithms. They used confidence, comprehensibility, interestingness, and surprise as objectives of the association rule mining problem. Specific mechanisms for mutations and crossover operators together with elitism have been designed to extract interesting rules from a transaction database. Empirical results of experiments carried out indicate high predictive accuracy of the rules generated.

In this paper deal with the ARM problem as a multi objective problem rather than as a single one and try to solve it using multi-objective evolutionary algorithms with emphasis on genetic algorithms (GA). The main motivation for using GAs is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining tasks. Multi-objective optimization with evolutionary algorithms is well discussed. The proposed algorithm was tested on a dataset drawn from the UCI repository of machine learning databases. For brevity, the data used is of a categorical nature. In this paper they had dealt with a challenging NP-Hard association rule mining problem of finding interesting association rules [5].

Xin Li et al [11] proposed Frequent Itemsets Mining in Network Traffic Data. They think about the problem of frequent itemset mining problem in network traffic data, and propose an algorithm for mining frequent itemsets. They try to minimize the size of results and only maximal frequent itemsets are considered. To protect the privacy, intermediate mining results are encrypted using hashing method by different servers. The proposed algorithm is evaluated from the perspectives of accuracy and efficiency.

Mining of frequent itemsets using Genetic algorithm was proposed in [12]. This work carried out with logic of GA to improve the scenario of frequents itemsets data mining using association rule mining. The main benefit of using GA in frequent itemsets mining is to perform global search with less time complexity. This scheme gives better results in huge or larger data set. It is also simple and efficient.

Another frequent itemsets mining approach based on genetic algorithm for non binary dataset was proposed by G. Vijay Bhasker et al [13]. They present an efficient algorithm for generating significant association rules among database items. GA is used to improve the scenario and system can predict about negative attributes in generated rules. As per results obtained this scheme is simple and efficient one. The Time complexity of the algorithm is also less and suitable for non binary data sets.

In continuation with this R. Vijaya Prakash et al [14] proposed similar method mining frequent itemsets for large data set using Genetic Algorithm. They implement frequent itemsets mining for numeric attributes also. Association rule mining is used to find relationship among attributes of database. This process was much time consuming and applied on discrete attributes. GA gives the facility of global search and minimum complexity. This algorithm avoids the necessity of discretizing apriori in attribute domain. They used an evolving algorithm to find the most appropriate amplitude of the intervals that be conventional a k-itemset, so that they have an elevated support value without being the intervals too extensive.

Sanat Jain and Swati Kabra [15] proposed Mining & Optimization of Association Rules Using Effective Algorithm. In this they work on association rules organization and frequent itemsets generation using positive and negative association rule mining. They proposed an apriori based algorithm to find valid positive and negative association rule in confidence framework or structure. As per result this algorithm is efficiently works for mining of positive and negative association rules in database and also optimize positive and negative association rule using genetic algorithm. This approach also reduces the search space and improved usability of mining rules that uses correlated coefficient to judge which association rule is used to mine.

3. CONCLUSION

Here in this paper a brief introduction and survey of different algorithm that is used for the knowledge discovery. An association rule based mining approach is also given and different item sets mining approach for the knowledge extraction.

4. REFERENCES

[1] Farah Hanna AL-Zawaidah, Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona, ―An Improved Algorithm for Mining Association Rules in Large Databases‖, World of

Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 2011, pp. 311 316.

[2] Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad, ―Optimization of Association Rule Mining using Improved Genetic Algorithms‖ IEEE 2004.

[3] Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., ―Optimized association rule mining using genetic algorithm‖, Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp-01-04.

[4] Rupali Haldulakar and Prof. Jitendra Agrawal, ―Optimization of Association Rule Mining through Genetic Algorithm‖, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Mar 2011, pp. 1252-1259.

[5] Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba, ―Extraction of Interesting Association Rules Using Genetic Algorithms‖, Advances in Systems Modelling and ICT Applications, pp. 101- 110.

[6] 1S. Dehuri,A. K. Jagadev, A. Ghosh and R. Mall, ―Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations‖, American Journal of Applied Sciences 3 (11), 2006, pp. 2086-2095.

[7] Ansaf Salleb-Aouissi, Christel Vrain andCyril Nortet, ―QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules‖, IJCAI- 2007, pp. 1035 1040.

[8] M. Ramesh Kumar and Dr. K. Iyakutti, ―Genetic algorithms for the prioritization of Association Rules‖, IJCA Special Issue on ―Artificial Intelligence Techniques - Novel Approaches & Practical Applications‖ AIT, 2011, pp. 35-38.

[9] Duke Hyun Choi, Byeong Seok Ahn, Soung Hie Kim, Prioritization of association rules in data mining: Multiple criteria decision approach, Expert Systems with Applications: An International Journal, v.29 n.4, p.867- 878, November, 2005.

[10] Choi et al., (2005).Prioritization of association rules in data mining: Multiple criteria decision approach. Expert Systems with Applications. v29. 867-878.

[11] Xin Li, Xuefeng Zheng, Jingchun Li, Shaojie Wang "Frequent Itemsets Mining in Network Traffic Data", 2012 Fifth International Conference on Intelligent Computation Technology and Automation, pp. 394- 397, 2012.

[12] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, pp. 133 – 143, October 2010.

[13] G. Vijay Bhasker, K. Chandra Shekar, V. Lakshmi Chaitanya "Mining Frequent Itemsets for Non Binary Data Set Using Genetic Algorithm", International Journal Of Advanced Engineering Sciences And Technologies (IJAEST), ISSN: 2230-7818, Vol. 11, Issue No. 1, pp. 143 – 152, 2011.

[14] R. Vijaya Prakash, Dr. Govardhan, Dr. S.S.V.N. Sarma "Mining Frequent Itemsets from Large Data Sets using Genetic Algorithms", IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" (AIT-2011), ISSN: 0975 – 8887, Special issue No. 4, Article -7, pp. 38- 43, 2011.

[15] Sanat Jain, Swati Kabra "Mining & Optimization of Association Rules Using Effective Algorithm", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 4, pp. 281- 285, April 2012.

[16] Ignasi Paredes-Oliva, Ismael Castell-Uroz, Pere Barlet-Ros, Xenofontas Dimitropoulos and Josep Sol´e-Pareta, "Practical Anomaly Detection based on Classifying Frequent Traffic Patterns",IEEE 2012.