Exploration of Association Rules by applying Automated Cogency

Kiruthika B, Hema M.S.

Abstract — A new EARAC (Exploration of Association Rules by applying Automated Cogency) algorithm is presented to explore association rules. Cogency is estimated using pairwise item conditional probability. EARAC algorithm explores association rules by only one pass through the dataset. Interesting rules are discovered from the set of all possible rules by using automated adjustment of Cogency.

Index Terms — Association rules, Cogency, Support, L-matrix, Transactions, Itemsets, Count.

1 INTRODUCTION

Data Mining is delineated as the form of inspecting large pre-existing databases in order to bring out new information. It can also be outlined as mining knowledge from data.Diverse approaches involved in data mining process are Classification, Clustering, Association Rules, Regression, Genetic Algorithm, Nearest Neighbor method, Artificial Intelligence, Neural Networks, Decision Trees.

1.1. Association Rule Mining

Association rule mining detects frequent patterns, correlations ,associations among sets of items in transaction databases. It is used to forecast customer buying habits by finding associations and correlations between different items that customers purchase.

Rule form: Antecedent → Consequent [support , confidence]

Support and Confidence are user defined threshold values of interestingness measure. Support indicates the percentage of items in antecedent and consequent part that are purchased together. Confidence indicates percentage of transactions that contains antecedent also contains consequent.

Example : Cricket Bat → Cricket Ball [30%,60%]

1.2 Market Basket Analysis

Market Basket Analysis predicts that if a customer buys a certain collection of items, then the customer is more or less likely to buy another collection of items. It guides to discover if there are combinations of products that frequently co-occur in transactions.

Example : Customers who buy flour and sugar may also buy eggs. A retailer can use information from market basket analysis for :

1. Store layout - Align products that co-occur together close to one another for better customer shopping experience.

2. Marketing - Target customers who buy flour with offers on eggs to boost them to spend more on their shopping basket.

Online retailers use market basket analysis to:

a. Inform the placement of items on their online sites or products in their catalogue

b. Run recommendation engines (Flipkart: customers who bought a particular product also bought few other products)

c. Carry out targeted marketing (e.g. emailing customers who bought specific items with other items and offers on those items that are likely to be purchased by them.)

1.3 Terminology

Items:

Objects in transactional databases for which patterns and associations are identified. For an online retailer each item in the shop is a product.

Itemset:

A group of items is an item set.

Transactions:

Instances of groups of items co-occuring together.

Rules:

Rules are statements of the form {i1, i2,...} ⇒ {ik} i.e., if there are items on the left hand side (LHS) of the rule then it is likely that a visitor will be interested in the item on the right hand side(RHS). For example:

{flour, sugar} ⇒ {eggs}

2. LITERATURE SURVEY

2.1 Apriori Algorithm

In Apriori algorithm [1] , Level wise search and Candidate itemset generation are involved .In first step, set of frequent 1-itemsets L1 is found. L1 is used to find L2, the set of frequent 2-itemsets.L2 is used to find L3 and so on until no frequent k-itemsets can be found. In second step, strong association rules are generated from the frequent itemsets.

2.2 FP Growth Algorithm

FP Growth algorithm [2] does not involve candidate itemset generation. It builds a compact data structure called FP Tree. Frequent itemsets are directly extracted from the FP tree. Frequent itemsets are the itemsets which satisfy the minimum threshold. From the frequent itemsets extracted from the tree,association rules are mined. FP growth algorithm is more efficient compared to apriori algorithm.

2.3 Pareto Based Genetic Algorithm

Association rule discovery is considered as a multiobjective problem [3]. It uses three measures namely Comprehensability , Interestingness and Predictive accuracy.Fitness values are used to find association rules. Fitness values are calculated based upon the ranks which are calculated from non-dominance property. Rules which has higher fitness value are added to association rules list. Rules with low fitness value are pruned.

2.4 The Fiti Algorithm

FITI algorithm [4] involves mining of inter transaction association rules. It uses three phases to extract association rules.

First phase involves mining of intra transaction itemsets and storing the intra transaction itemsets. Second phase involves database transformation.Original database is transformed into another database which stores the ids of frequent intratransaction itemsets to avoid regeneration of frequent intratransaction itemsets. Third phase involes mining of frequent inter transaction itemsets. Association rules are mined from the generated inter transaction itemsets.

2.5 Ranwar Algorithm

Rank based weighted mining [5] is presented in this paper.Two interestingness measures namely weighted condensed support and weighted condensed confidence are used.

WCS and WCC depends on the rank of items.Rank is computed for all the items present in the dataset.Using the rank computed,weight is assigned to each item.Based on weight,association rules are mined.

2.6 Row Enumeration Algorithm

High confidence association rules are mined from micro array datasets. A Row enumeration rule mining method [6] is proposed to explore rules. An interesting measure called MAX-CONF is used to mine rules. MAX-CONF employs two confidence pruning methods level 1 and level 2 to effectively prune the search space.Gene relationships are discovered from micro array data.

2.7 Multi Objective Evolutionary Algorithm

This paper focuses on negative dependencies for mining quantitative association rules. MOPNAR [7] is a multi objective evolutionary algorithm to mine a reduced set of positive and negative quantitative association rules.Negative association rules offer information that can be used to support decisions for applications.MOPNAR focues on comprehensibility, interestingness and performance.Rules which satisfy all the three support measures are added to the association rule list..

2.8. Generalized Association Pattern Mining Algorithm

This algorithm [8] focuses on mining generalized associations of semantic relations from the textual content of web documents. Two steps are involved in mining associations. In first step meta data representing semantic relations are extracted from raw text. It is carried out by using natural language processing techniques. The second step involves using the GP-close algorithm to discover the association patterns from meta data.

2.9 Arm Algorithm With Statistical Measures

This algorithm [9] uses three main steps for discovering binding cores in protein-DNA binding.In the first step,it discovers protein-DNA associated patterns.The discovered pattern must be within a specified width limit and must satisfy the criteria of support and confidence.

The second step computes the threshold value for each associated pattern.Patterns which does not satisfy the threshold value are filtered out..

The third step involves meta processing on the associated patterns. It involves four phases namely ranking , verification , evaluation, redundancy removal. Highly ranked associated patterns are mined.

3. THE PROPOSED EARAC ALGORITHM

The Proposed EARAC algorithm mines effective association rules by automated adjustment of cogency. It involves four main phases :

1. L-Matrix Construction
2. Cogency Calculation
3. Frequent Itemset Generation
4. Association Rules Exploration

3.1 L-Matrix Construction

L-Matrix gives the total number of occurrences of items in the dataset. It specifies the number of occurrences of each item as well as each item's occurrence in pairwise order with all other items in the given dataset.

L-Matrix :

	ITEM 1	ITEM 2	..	ITEM N
ITEM 1	L_{11}	L_{12}		L_{1N}
ITEM 2	L_{21}	L_{22}		L_{2N}
..				
ITEM N	L_{N1}	L_{N2}		L_{NN}

L denotes the link between items in the dataset. L_{11} specifies total number of times item 1 occurred in the dataset. Similarly L_{22}, L_{33}, L_{44}, …L_{NN} gives the total number of occurrences of each item in the dataset. Values of main diagonal gives the total number of occurrences of each item in the given dataset. L_{12} gives the total number of times item 1 and item2 occurred together in the given dataset.Similarly L_{13}, L_{14},... L_{NN} gives the pairwise occurrence of items.

3.2 Example

TRANSACTION IDS	ITEMSETS
1	Pencil, Pen
2	Pencil, Pen, Marker
3	Pencil, Marker
4	Pen, Marker
5	Marker

L-Matrix

	Pencil	Pen	Marker
Pencil	3	2	2
Pen	2	3	2
Marker	2	2	4

1-Itemset :

Main diagonal values give the total number of occurrences of Pencil, Pen and Marker.

ITEMSET	COUNT
Pencil	3
Pen	3
Marker	4

2-Itemset :

ITEMSET	COUNT
{ Pencil,Pen }	2
{ Pencil,Marker }	2
{ Pen,Marker }	2

3.3 L-Matrix Steps :

Step 1 : Initially set values of L-matrix to be 0.
Step 2 : Load the dataset.
Step 3 : Scan the dataset and read each transaction.
Step 4 : While reading each transaction, the count of items present in that particular transaction is incremented by '1' in the L-Matrix.
Step 5: If certain items are not present in the transaction, their values in L-Matrix remains unmodified.
Step 6: Repeat all steps until the end of transactions.

One itemset and two itemsets can be derived directly from L-Matrix. Hence it requires only one scan of the dataset.

3.4 Cogency Calculation

Cogency is a threshold value [13] similar to support and confidence. It is calculated based on pairwise item conditional probability [11]. It is used to mine effective association rules from set of possible association rules.

3.4.1 Cogency Formula :

Cogency = Cogency * (L(ay) / L(yy))

- Initially Cogency value is set to be 1.
- 'a' denotes each item in frequent 2-itemset, frequent 3-itemset,..,frequent n-itemset.
- 'y' denotes each item in frequent 1-itemset.
- L_{ay} gives the total number of times item a and item y occurred together in the dataset.
- L_{yy} gives the total number of times item y occurred in the dataset.

3.5 Automated setting of minimum cogency 1 and minimum cogency 2

Minimum cogency 1 and minimum cogency 2 are threshold values to find frequent 1-itemsets and effective association rules respectively. In existing [10] system minimum cogency 1 and minimum cogency 2 are predefined by the user. If minimum cogency is considered high, only few frequent itemsets and rules are generated. If it is considered low, many rules are generated among which some may be uninteresting.

In proposed EARC algorithm, minimum cogency 1 and minimum cogency 2 are automatically generated using dataset statistics [12]. Due to automated adjustment of minimum cogency, effective rules can be mined without missing any rules.

3.5.1 Steps to calculate automated cogency :

Step 1 : Find initial cogency by analyzing the itemsets and their frequency.
Step 2: Cumulative Cogency for the subsequent levels are found based on the previous level cogency and the items considered in the current level.

3.6 Frequent Itemset Generation

Step 1 : Generate Frequent 1-itemset and frequent 2-itemset by applying minimum support threshold to L- Matrix .
Step 2 : Let each item in frequent 2-itemset be X and each item in frequent 1-itemset be Y. For each item X in frequent 2-itemset, find Y in frequent 1-itemset (X → Y) such that X ∩ Y is empty.
Step 3 : Compute cogency for X and Y.
Step 4 : If computed Cogency(X → Y) > minimum cogency 1, generate frequent 3-itemsets by adding each Y to its corresponding X. Else prune the itemset.
Step 5 : Repeat until all frequent itemsets becomes empty.

3.7 Association Rules Exploration

Step 1 : For each frequent itemset generated , check Cogency (X → Y) > minimum cogency 2.
Step 2 : If Cogency (X→Y) > minimum cogency 2, add rule X → Y to association rules list.
Step 3 : If Cogency (X→Y) < minimum cogency 2, prune the rule.
Step 4 : Repeat until frequent itemsets become empty.

3.8 Example

- Minimum support = 30 % =0.3
- Frequent 1-itemset= {{pencil},{pen},{marker}}
- Frequent 2-itemset={{pencil,pen},{pencil,marker}, {pen,marker}}
- Predefined minimum cogency 1=0.4
- Predefined minimum cogency 2=0.4

1. {pencil,pen},{marker}

{pencil,marker}

Step 1: initial cogency = 1
Step 2: if L(ay) < S * n return -1
2 is not less than 0.3 * 5
Step 3: cogency = cogency * L(ay)/L(yy)
= 1 * 2/4 = 0.5

{pen,marker}

Step 4: cogency=0.5*2/4=0.25
cogency < min cog 1. So, Prune the itemset.

2. {pencil,marker},{pen}

{pencil,pen}

Step 1: initial cogency = 1
Step 2: if L(ay) < S * n return -1
2 is not less than 0.3 * 5
Step 3: cogency = cogency * L(ay)/L(yy)
= 1 * 2/3 = 0.6667

{marker,pen}
Step 4: cogency=0.6667*2/3=0.4445
cogency > min cog 1,min cog 2. So add {pencil,marker} → {pen} to association rules list.

3. {pen,marker},{pencil}

{pen,pencil}

Step 1: initial cogency = 1
Step 2: if L(ay) < S * n return -1
2 is not less than 0.3 * 5
Step 3: cogency = cogency * L(ay)/L(yy)
= 1 * 2/3 = 0.6667

{marker,pencil}
Step 4: cogency=0.6667*2/3=0.4445
cogency > min cog 1,min cog 2. So add {pen,marker} → {pencil} to association rules list.

4 CONCLUSION AND FUTURE WORK

The suggested EARAC algorithm explores association rules by only one pass through the dataset. Hence it is faster and consumes less memory due to its one time file access. Au-

tomated adjustment of cogency according to dataset statistics helps to mine more efficient association rules.

Incremental and streaming dataset using EARAC algorithm can be considered in the future work and the algorithm's efficiency can be enhanced further from software and hardware aspect in future.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in Proc. 20th Int. Conf. Very Large Data Bases, San Francisco, CA, USA, 1994, pp. 487–499.

[2] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation:A frequent-pattern tree approach," Data Mining Knowl. Discovery, vol. 8, no. 1, pp. 53–87, 2004.

[3] A. Ghosh and B. Nath, "Multi-objective rule mining using genetic algorithms," Inf. Sci. , vol. 163, nos. 1–3, pp. 123–133, Jun. 2004.

[4] Anthony K.H. Tung, Member, IEEE, Hongjun Lu, Member, IEEE, Jiawei Han, Member, IEEE, and Ling Feng, Member, IEEE,"Efficient Mining of Intertransaction Association Rules", ieee transactions on knowledge and data engineering, vol.15, no. 1, january/february 2003.

[5] Saurav Mallik, Anirban Mukhopadhyay, "RANWAR: Rank-Based Weighted Association Rule Mining From Gene Expression and Methylation Data," ieee transactions on nanobioscience, vol. 14, no. 1, january 2015.

[6] Tara McIntosh and Sanjay Chawla, "High-Confidence Rule Mining for Microarray Analysis," Ieee/Acm Transactions On Computational Biology And Bioinformatics, Vol. 4, No. 4, October-December 2007.

[7] Diana Mart´ın, Alejandro Rosete, Jes´us Alcal´a-Fdez, "A New Multiobjectiv Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules," Ieee Transactions On Evolutionary Computation, Vol. 18,No.1, February 2014.

[8] Tao Jiang, Ah-Hwee Tan, "Mining Generalized Associations of Semantic Relations from Textual Web Content," Ieee Transactions On Knowledge And Data Engineering, Vol. 19, No. 2, February 2007.

[9] Man-Hon Wong, Ho-Yin Sze-To, Leung-Yau Lo, Tak-Ming Chan, and Kwong-Sak Leung,"Discovering Binding Cores in Protein-DNA Binding Using Association Rule Mining with Statistical Measures," Ieee/Acm Transactions On Computational Biology Vol. 12,Jan2015.

[10] Azadeh Soltani and M.-R. Akbarzadeh-T., Senior Member, IEEE, " Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets" ieee transactions on neural networks and learning systems, vol. 25, no. 11, november 2014.

[11] A R. Hecht-Nielsen, "Cogent confabulation," Neural Netw., vol. 18, no. 2, pp.111–115, Mar. 2005.

[12] C.S.Kanimozhi Selvi, and A.Tamilarasi, "An Automated Association Rule Mining Technique With Cumulative Support Thresholds," Int. J. Open Problems in Compt. Math, Vol. 2, No. 3, September 2009 ISSN 1998-6262; Copyright © ICSRS Publication, 2009.

[13] M. J. Heravi, "A study on interestingness measures for associative classifiers," M.S. thesis, Dept. Comput. Sci., Alberta Univ., Edmonton, AB, Canada, 2009.

[14] Y. H. Hu and Y. L. Chen, "Mining association rules with multiple minimum supports: A new mining algorithm and a support tuning mechanism," Decision Support Syst., vol. 42, no. 1, pp. 1–24, 2006.

[15] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 32–41.