ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Enhanced GenMax Algorithm for Data Mining

Saifullahi Aminu Bello[1], Abubakar Ado[2], Tong Yujun[3], Abubakar Sulaiman Gezawa[4], Abduarra'uf Garba[5]

Lecturer, Department of Computer science, Kano University of Science and Technolgy, Kano, Nigeria [1]

Lecturer, Department of Computer Science, Northwest University, Kano, Nigeria [2]

Professor, Electrical/Electronics, Liaoning University of Technology, Jinzhou, Liaoning Province, China[3]

System Analyst, MIS Department, Bayero University Kano, Kano, Nigeria[4]

Lecturer, Department of Computer Science, Northwest University, Kano, Nigeria [5]

Abstract: Association rules mining is an important branch of data mining. Most of association rules mining algorithms make use of only one minimum support to mine items, which may be of different nature. Due to the difference in nature of items, some items may appear less frequent and yet they are very important, and setting a high minimum support may neglect those items. And setting a lower minimum support may result in combinatorial explosion. This result in what is termed as "rare item problem". To address that, many algorithm where developed based on multiple minimum item support, where each item will have its minimum support. In this research paper, a faster algorithm is designed and analyzed and compared to the widely known enhanced Apriori algorithm. An experiment has been conducted and the results showed that the new algorithm can mine out not only the association rules to meet the demands of multiple minimum supports and but also mine out the rare but potentially profitable items' association rules, and is also proved to be faster than the conventional enhanced Apriori.

Keywords: Data Mining, Association Rules, Multiple Minimum Item supports, Apriori, GenMax Algorithm.

I. INTRODUCTION

Due to the ease at which data is saved and availability of memory spaces, data is growing exponentially. Data mining can extract the unknown but potentially useful information [1] from a mass, incomplete and noisy raw data.
The Barcode technology use in the supermarkets allows easy recording of sale transactions. Data mining technologies in the sales centers/e-commerce have entered a practical stage and achieved good effects.

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.

Following the original definition by Agrawal et al[1]. The problem of association rule mining is defined as: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called *items*. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A *rule* is defined as an implication of the form $X \Longrightarrow Y$ where $, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short *itemsets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

2.4.1 SUPPORT: The *support* Supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

$$\text{support } (X \rightarrow Y) = \frac{(X \cup Y).count}{n}$$

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Usually one user specified minimum support is used when mining itemsets, any itemset whose proportion is greater than or equals the minimum item support is considered to be frequent.

Due to the different nature of the items in a data set, some items are more frequent than the others, if a single minimum support is applied, the less frequent items cannot be mined except if the support is very low, and in that case the more frequent items will associates with all items in any possible way.

Multiple Minimum Item Support (MMIS): the concept of multiple minimum item support is introduced to address the rare items problem.

2.4.2 CONFIDENCE: Confidence of a rule is the conditional probability of Y given X. Using probability notation: confidence (X implies Y) = P (Y given X).

$$\textbf{confidence (X}\rightarrow\textbf{Y)} = \frac{support(X \cup Y)}{support(X)} = \frac{(X \cup Y).count}{X.count}$$

This paper analyses the practical application of data mining on an e-commerce data to find the association rules between items.

An algorithm use in [2] called GenMax algorithm which was used to mine out association rules between items based on single Minimum Item Supports is enhanced, thus allowing it to mine items based on multiple minimum item support as devised in [3], which is an enhancement of the conventional apriori algorithm.

II. RELATED WORK

2.1 ENHANCED APRIORI

Tong Yujun [3] analyzed an algorithm based on the original apriori algorithm according to a multiple user specified Minimum Item supports and support difference constraint.

```
Input: transaction T, MIS, φ
Output: frequent k-itemsets
Line 1   M=sort (I, MIS);
Line 2   L=init-traverse(M, T);
Line 3   F1={{l}| l∈ L, l.count/n≥MIS(l)};
Line 4   for (k=2; Fk-1 ≠Φ; k++) {
Line 5     if k=2 then  Ck=candidate-generate-level2(L, φ) [2];
Line 6     else  Ck=candidate-generate (Fk-1, φ) [2];
Line 7     for each transaction t∈ T {
Line 8       for each candidate c∈ Ck {
Line 9         if c is contained in t then c.count++;
Line 10        if c-{c [1]} is in t then (c-{c[1]}).count++;
Line 11      }
Line 12    }
Line 13    Fk = {c∈ Ck| c.count/n≥MIS(c[1])};
Line 14  }
Line 15  return F= ∪k Fk;
```

In the improved algorithm, the key operation is to sort the items of I in ascending order basing on their minimum item support values as is used in following operations (line 1), all the 1-itemset can be obtained in F_1(line 2,3), function candidate-generate-level2(L, φ) can let the Down Closure Property principle be valid as usual(line 5), function candidate-generate (F_{k-1}, φ) (line 6) is similar to that in the Apriori algorithm but different in pruning step, and counting c-{c[1]} is to calculate the confidence of association rules (line 10). At last, the algorithm returns the union of frequent itemsets from 1-itemsets to k-itemsets (line 15).

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

2.2 GENMAX ALGORITHM

GenMax [2] is a backtrack search based algorithm for mining maximal frequent itemsets. GenMax uses a number of optimizations to quickly prune away a large portion of the subset search space. It uses a novel *progressive focusing* technique to eliminate non-maximal itemsets, and uses *diffset propagation* for fast frequency checking.

//Invocation: MFI-bactrack(Ø, F_1, 0)
MFI-backtrack(I_l, C_l, l)

1. **for each** $x \in C_l$
2. $I_{l+1} = I_l \cup \{x\}$
3. $P_{l+1} = \{y: y \in C_l \text{ and } y > x\}$
4. ***if*** $I_{l+1} \cup P_{l+1}$ *has a superset in* ***MFI***
5. ***Return*** // all subsequent branches pruned!
6. $C_{l+1} = \text{FI} - \text{combine}(I_{l+1}, P_{l+1})$
7. **if** C_{l+1} is empty
8. **if** I_{l+1} has no superset in **MFI**
9. **MFI** = **MFI** $\cup I_{l+1}$
10. else $\text{MFI} - \text{backtrack}(I_{l+1}, C_{l+1}, l+1)$

III. FAST ALGORITHM WITH MMIS

The algorithm is an application of the idea developed in [3], the enhanced apriori with the one in [2], the genMax algorithm.

Support difference constraint (SDC)[2]. In the improved algorithm the Down Closure Property principle is not satisfied. So after sorting the itemsets by their values of the MIS, SDC is introduced to generate frequent itemsets from the itemsets with the frequent, rare but potentially profitable ones. Suppose sup(i) is the real support of an item in itemsets, to each itemset the SDC is defined as

$\max_{i \in s}\{\text{sup(i)}\} - \min_{i \in s}\{\text{sup(i)}\} \leq \varphi$ [2]. (4)

THE ALGORITHM

C is the set of items sorted in ascending order of their minimum supports.
L is the set of maximal frequent itemsets.
S_x is the support count of item x.

Input: transactions, I, C
Output L
EnGenMax(I, C)

1. $I_{L+1} = I \cup \{x_1: x_1 \in C\}$
2. C_{L+1}=C-x_1
3. $I'_{L+1} = I \cup \{x_2: x_2 \in C\}$
4. $C'_{l+1} = C_{l+1} - x_2$
5. $if\ S_{x_1} \leq minSup(I)$ // minumum of all the MIS of elements in I
6. $if\ C_{L+1} \neq \emptyset$
7. EnGenMax(I_{L+1}, C_{L+1})
8. $if\ I_{L+1}\ has\ no\ superset\ in\ L\ then\ L = L \cup I_{L+1}$
9. $if\ S_{x_2} \leq minSup(I)$ //minimum of all the MIS of elements in I
10. $if\ C'_{L+1} \neq \emptyset$
11. EnGenMax(I'_{L+1}, C'_{L+1})
12. $if\ I'_{L+1}$ has no superset in L then
13. L:= L $\cup$ I'_{L+1}

The algorithm is best described following the solution in section 4.2

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

IV. EXPERIMENT AND ANALYSIS

the experiment make use of an e-cart data obtained some times in May, 2013 from an e-commerce website as shown in table 1.

Table 1 shows the transactions obtained. And table 2 indicates the items with there respective minimum support and there count in the transaction shown is a sorted ascending order of their minimum support.

Table 1: transactions

s-cart_ID	itemset
941518	CD, DVD player
710495	CD, laptop
643426	CD, DVD player, laptop
732963	Laptop, TV
987213	CD, Laptop
782362	DVD, laptop, TV
948628	DVD, laptop
535429	Laptop, TV
358215	CD, DVD, laptop, TV
682725	DVD, laptop, TV

table 2: items with MIS and support count

ID	itemset	MIS	support
1	{laptop}	0.2	0.8
2	{CD}	0.3	0.5
3	{DVD}	0.5	0.6
4	{TV}	0.6	0.6

For easy sorting of the items, the items are represented by numbers, 1, 2, 3, 4 according to there minimum supports.

Table 3: items sorted by MIS

Number representation	Item	Minimum support
1	Laptop	0.2
2	CD	0.3
3	DVD	0.5
4	TV	0.6

4.1 USING ALGORITHM FROM SECTION 2.1

Following the algorithm proposed in 3. The above transactions is mined in this way

Table 4: Frequent 1-items

ID	itemset	MIS	support
1	{1}	0.2	0.8
2	{2}	0.3	0.5
3	{3}	0.5	0.6
4	{4}	0.6	0.6

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Table 5: candidate 2-items

ID	itemset	MIS(c[1])	support
1	{1, 2}	0.2	0.4
2	{1, 3}	0.2	0.4
3	{1, 4}	0.2	0.5
4	{2, 3}	0.3	0.3
5	{2, 4}	0.3	0.1
6	{3, 4}	0.5	0.4

table 6: frequent 2-items

ID	itemset	MIS(c[1])	support
1	{1, 2}	0.2	0.4
2	{1, 3}	0.2	0.4
3	{1, 4}	0.2	0.5
4	{2, 3}	0.3	0.3

Table 7: Candidate 3-items

ID	itemset	MIS(c[1])	support
1	{1,2,3}	0.2	0.2
2	{1,2,4}	0.2	0.1
3	{1,3,4}	0.2	0.3

table 8: frequent 3-items

ID	itemset	MIS(c[1])	support
1	{1,2,3}	0.2	0.2
2	{1,3,4}	0.2	0.3

4.2 USING THE PROPOSED ALGORITHM OF SECTION 3.0

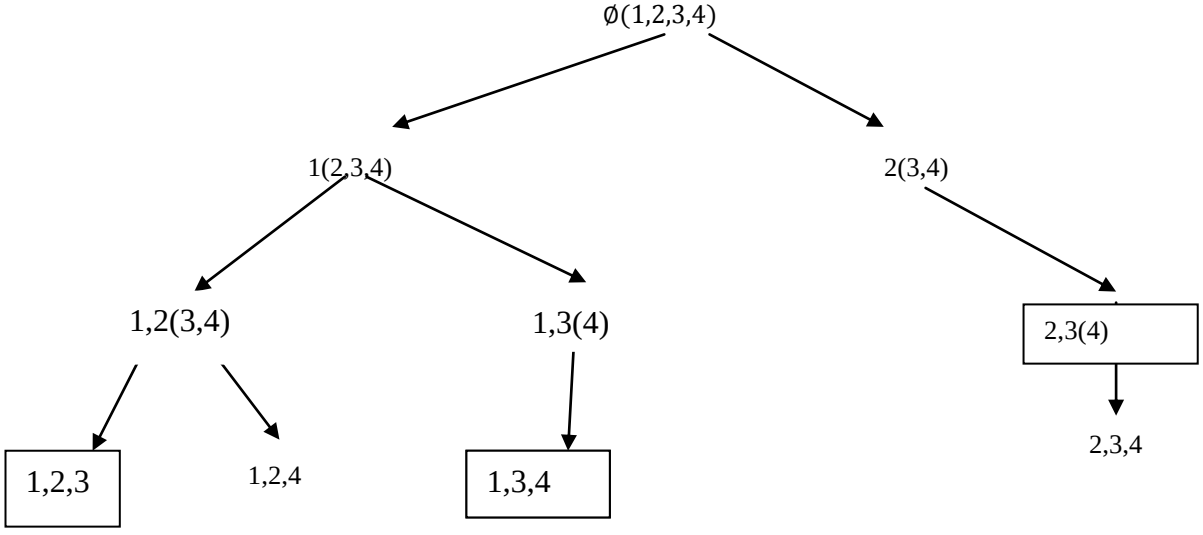

Fig 1. Pictorial representation of a problem solved using the proposed algorithm.
Items in the box represent the maximal frequent items sets.
Thus, the maximal frequent item sets are [{2,3}, {1,2,3}, {1,3,4}]

The two frequent 3-items sets corresponds to those obtained from 3.1 above.
To find the remaining frequent item sets, we first find all the possible subsets of the above maximal frequent item-sets.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Table 9: 2-items subsets

itemsets	MIS(c[1])	support
1,2	0.2	0.4
1,3	0.2	0.4
1,4	0.2	0.5
2,3	0.3	0.3
3,4	0.5	0.4

Table 10: 1-item subsets

item	MIS	support
1	0.2	0.8
2	0.3	0.5
3	0.5	0.6
4	0.6	0.6

From the above subsets, in table 9, only one subset is, {3,4} not frequent.

4.3 COMPARISON OF 3.1 AND 3.2

Under the approach in [3], the 2-item candidate set is the largest candidate set. This is always the case both in [1] and [3]. From the above examples, it is evident to see that the number of candidate sets generated to find the maximal frequent item sets is much lower than the number of candidates generated directly to find all the frequent item sets. And while both approaches arrives at the same frequent item sets, the number of candidate sets in the proposed algorithm is lower than the approach used in [3].

Based on this advantage and the fact that the algorithm in [3] follows the principle: the (n-1) frequent itemsets is used to generate n-frequent candidate sets, the number of transaction traversing in [3] is greater than the number of traversing is the proposed algorithm.

Below is a graph that shows the above advantages more clearly:

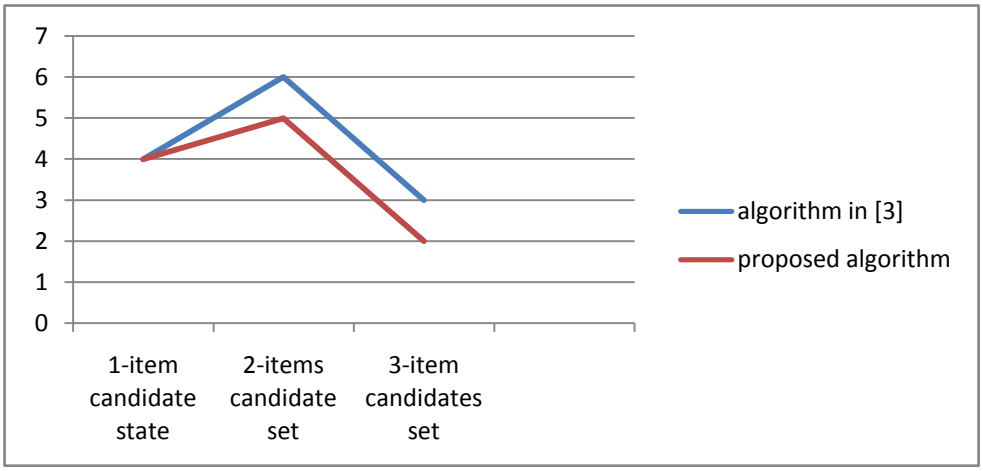

Fig 2: a graph of algorithm in [3] and the proposed algorithm. The vertical axis shows the size of the n-item candidate sets, while the horizontal shows the n-items candidate set.

V. CONCLUSION

While due to the nature of items in real life, using a single minimum item support results in rare items problem. The issue is addressed with multiple minimum item supports as proposed in [3]. The algorithm proposed in this paper uses the same technique of multiple minimum item support but is proved to be more efficient than the algorithm in [3] which is an enhancement of the conventional apriori algorithm.

The proposed algorithm reduced the number of candidate item sets, thus reduced the number of transaction traversing which in effect will reduce the programming timing and the memory cost.

Since the algorithm find directly the maximal frequent itemsets, it can be used where only the maximal frequent item sets is required. Maximal frequent item sets are used in several data mining tasks such as incremental mining of dynamic databases.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

REFERENCES

[1] R. Agrawal, and R. Srikant. Fast algorithms for mining association rules in large database. Technical Report FJ9839, IBM Almaden Research Center , San Jose, CA,(Jun. 1994),p.2-6

[2] Matthew Eric, Oteyy Adriano, Velosoyz Chao, Wangy Srinivasan, Parthasarathyy Wagner, Meira Jr.z, "Incremental Techniques for Mining Dynamic and Distributed Databases" Computer and Information Science Department The Ohio-State University. (n.d.)

[3] Yujun Tong**,** Jun Zhou**,** Wen'ge Xie**,** Dan Jia, "Research and Application of an Enhanced Data Mining Algorithm".ICMM2011, 841-844. 2011,Jinzhou, China.

[4] J, Han and Y. Fu. Discovery of multiple-level association rules from large databases. In Proc. 1995 Int. Conf. Very Large Databases(VLDB'95)(Sep. 1995), p.402-431.

[5] Shilpa, Sunita Parashar, "Performance Analysis of Apriori Algorithm with Progressive Approach for Mining Data" *International Journal of Computer Applications (0975 – 8887*) Volume 31– No.1, October 2011.

[6] Karam Gouda, Mohammed J. Zaki, "GenMax: An Efficient Algorithm for Mining Maximal Frequent Itemsets" Data Mining and Knowledge Discovery, 11, 1–20, 2005

[7] Bing L, Wynne H , Yiming M. Mining Association Rules with Multiple Minimum Supports. In: KDD-99.San Diego,USA(1999).

[8] D. Cheung, J. Han, V. Ng, and C. Wong. Maintenance of discovered association rules in large databases: An incremental update technique. In Proc. of the 12th Int. Conf. on Data Engineering (ICDE'96), March 1996.

[9] J, Han and Y. Fu. Discovery of multiple-level association rules from large databases. In Proc. 1995 Int. Conf. Very Large Databases(VLDB'95)(Sep. 1995), p.402-431.

[10] R. Agrawal, et al. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD Conference on Management of data(May 1993), p. 207-216.

[11] SIGMOD-98, 1998. [11] Park, J. S. Chen, M. S. and Yu, P. S. "An effective hash based algorithm for mining association rules." SIGMOD-95, 1995, pp. 175-186.

[12] Rastogi, R. and Shim, K. "Mining optimized association rules with categorical and numeric attributes." ICDE –98.

[13] Liu, B., Hsu, W. and Ma, Y. Mining association rules with multiple minimum supports. SoC technical report, 1999.

[14] Piatetsky-Shapiro, Gregory (1991), *Discovery, analysis, and presentation of strong rules*, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., *Knowledge Discovery in Databases*, AAAI/MIT Press, Cambridge, MA.

[15] Salleb-Aouissi, Ansaf; Vrain, Christel; and Nortet, Cyril (2007). "QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules". *International Joint Conference on Artificial Intelligence (IJCAI)*: 1035–1040.

[16] Bayardo, Roberto J., Jr.; Agrawal, Rakesh; Gunopulos, Dimitrios (2000). "Constraint-based rule mining in large, dense databases". *Data Mining and Knowledge Discovery* **4** (2): 217–240.doi:10.1023/A:1009895914772

[17] Rauch, Jan; Logical calculi for knowledge discovery in databases, in Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, Springer, 1997, pp. 47-57

[18] Bayardo Jr, Roberto J. "Efficiently mining long patterns from databases." ACM Sigmod Record. Vol. 27. No. 2. ACM, 1998.

BIOGRAPHY

Saifullahi Aminu Bello: is working with department of computer science, Kano University of Science and Technology, Kano-Nigeria. He obtained his B.sc computer science from the same university. And currently studies M.sc Computer Science and Technology at Liaoning university of Technology, Jinzhou-China. His area of interests is Data mining and Distributed databases. He received an award as the best graduating student in B.sc Computer Science 2008/2009 by Kano University of Science and Technology, Nigeria.

Abubakar Ado: is working with department of Computer Science, Northwest University, Kano-Nigeria. He Obtained his B. Tech Computer Science from Abubakar Tafawa Balewa, Bauchi-Nigeria. He is currently a masters student at Liaoning University of technology, Jinzhou-China. His area of interests is Database and information system. He has published a paper titled "An Integrated Framework For Detecting and prevention of Trojarn Horse (BINGHE) in a Client-Server Network".

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Tong Yujun: is lecturer at School of electrical/Electronic Engineering Liaoning University of Technology he is currently an Associate Professor, his Research area include: Web Data Mining, Distributed Database and Software Engineering.

Abubakar Sulaiman Gezawa: is a system analyst at the department of MIS, Bayero University Kano. He is currently a master student of Computer science and Technology, Liaoning University of Technology, Jinzhou, Liaoning Province, China.

Abduarra'uf Garba: is working with the computer science department, North-West University Kano, Nigeria. He is currently a master student of Computer science and Technology, Liaoning University of Technology, Jinzhou, Liaoning Province, China