Comparative Analysis between Theoretical Model and 2-Phase Method for Mining Profit Based Pattern

Vijay Kumar Verma
Department of Computer Science &Engineering, Lord Krishna College of Tech. Indore M.P.
Email: vijayvermaonline@gmail.com
KanakSaxena
Department of Computer Applications Samrat Ashok Technological Institute Vidisha M.P.
kanak.saxena@gmail.com

Abstract
Every business organization needs profit. Profit based pattern play an important role in several real life applications. Mining efficient profitbased pattern is a difficult task. Several researchers have been proposed efficient algorithms for mine profit based pattern. Each and every algorithm tries to minimize useless candidate's generation, minimizing traversal at each phase, reducing memory and execution time. In this paper we proposed a comparative analysis over two basic models Theoretical and Two phase model. Our comparison is based on some parameter like accuracy, arithmetic complexity and candidates cutting strategies.

Keywords: - Profit, Pattern, Comparative, Analysis, Accuracy.

1. Introduction

In a few past year information processing system have been rapidly changed. The use of computer based system is increasingly. Computer based system generate huge amount of data every day. To mine meaningful information from huge amount of data is crucial work. Data mining has a number of methods, algorithms and techniques to discover useful information form large data. Pattern which is based on Frequency only give information about which items reflect the transactions [1, 2].

2. Literature Review

Horizontal data format based approach was first was first proposed by Agrawal et al. (1994). Several improvements have been proposed on horizontal data format like Dividing data set technique (Savasere et al., 1995), selectionbased approach (Toivonen, 1996) DIC (Brin et al., 1997a). Counting the occurrences of item (Tiwari et al., 2009), CLOSET which was proposed in 2000 by Pei CHARM was proposed by Zaki in 2002, CLOSET was presented by Wang in 2003, FP Close (Grahne and Zhu, 2003) and AFOPT (Liu et al., 2003)[3,4,6].

Vertical data layout based approach also developed these includes vertical mining algorithms (Shenoyet al., 2000) Equivalence CLASS Transformation Zaki (2000), tress based approach by (Han et al., 2004)[7,8,9].

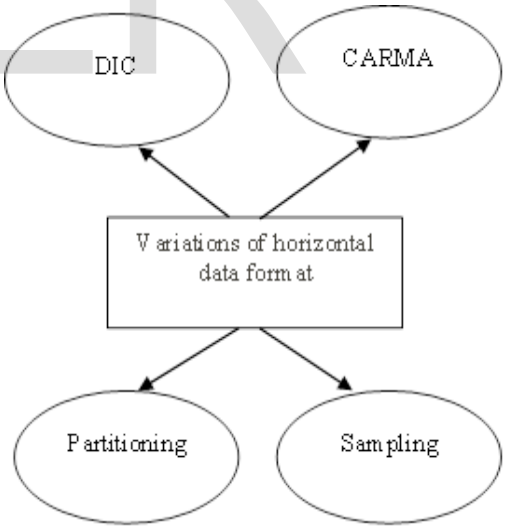

Figure 1 Variationsof horizontal data format

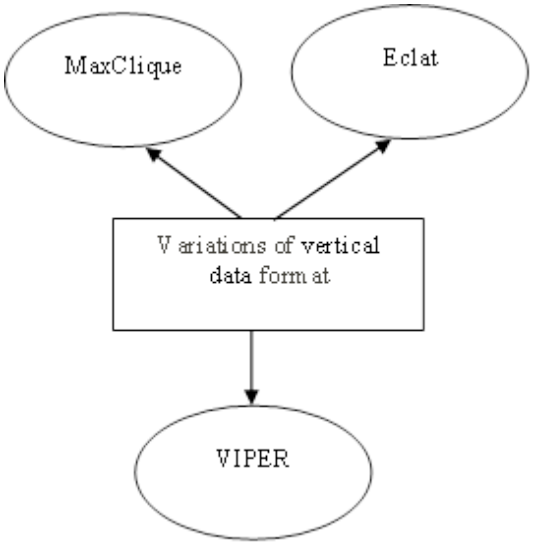

Figure 2 Variations of vertical data format

3. Comparative Analysis

We proposed a comparative analysis between theoretical model and 2-Phase model. Our comparison is based on method used for mining profit based pattern, arithmetic complexity [5,6].

Consider a simple table contains purchasing record of 10 consumers

Table 1
Purchasing records

TID	Items
1	C(18), D(0), E(1)
2	B(6), D(1), E(1)
3	A(2),C(1),E(1)
4	D(1), E(1)
5	D(4), E(2)
6	A(1), B(1)
7	B(10),D(1), E(1)
8	A(3),B(25),D(3),E(1)
9	A(1), B(1)
10	B(6),C(2),E(2)

Table 2

Profit value of each item

Item	A	B	C	D	E
Profit	3	10	1	6	5

Theoretical model is based on Support Bound Property.Suppose we want to calculate profit for item set P {B, D, E} using theoretical model. Let threshold is 120.

$$P(B,D,E) = \frac{Supmin\ (B,D,E)}{K-1} \times \left(\frac{P(B,D)}{Sup\ (B,D)} + \frac{P(B,E)}{Sup\ (B,E)}\right) + \frac{k-m}{k-1} \times \varepsilon$$

$\sup_{min}$ ({B, D, E})

= min {sup({B, D}), sup({B, E})}

= min{0.2, 0.3} = 0.2

Since p({D, E}) = 56 < ε, so left it.

$$P(B,D,E) = \frac{0.2}{3-1} \times \left(\frac{172}{0.2} + \frac{240}{0.3}\right) + \frac{1}{2} \times 120$$

=226> thresholdvalue.Now we use the same set and calculate the profit value using 2-Phase.2-Phase method is based on purchasing record based profit.

Table 3.
Record based profit

TID	Profit
1	23
2	71
3	12
4	14
5	14
6	13
7	111
8	57
9	13
10	72

To calculate the profit of item set ({B, D, E}) we need to find out the records which contains these item together.

$$PR(x) = \sum_{k=1}^{n} p(ID)i$$

PR({B, D, E})=TID(2)+TID(7)

PR({B, D, E})=71+111

PR({B, D, E})=182>threshold

BCD is high profit item set

In case of theoretical model the last term which contain k, m and termε isvery close one or has value bigger than one. Due to this useless participants may appear in the higher level. So we needs more search the space, but in case of 2-Phase we can efficiently remove useless participants so no needs extra space. If we remove the term from the calculation the accuracy the model is affected and wrong result has to be generated.

In theoretical model we needs complex calculation includes multiplication and division operations. These complex operations require extra time for executing the algorithm.

But in case of 2-phase we need simple calculations which include only a few multiplication and addition operation.

4. Conclusion

After performing some calculation we found that theoretical model is more complex in term of arithmetic calculation as compared to the 2-Phase model. The useless participants are difficult to remove incase of theoretical model because of support bound property but 2-Phase needs no extra efforts to remove use less participants. So 2-Phase model is better as compared to theoretical model.

References

[1] 2005 Hong Yao, Howard J. Hamilton, and Cory J. Butz "*A Foundational Approach to Mining Item set Utilities from Databases*" Department of Computer Science University of Regina Canada.

[2] Ying Liu Wei-keng Liao AlokChoudhary "*A Fast High Utility Itemsets Mining Algorithm*" UBDM, August 21, 2005, Chicago Copyright 2005 ACM 1-59593

[3] Alva Erwin, Raj P. Gopalan, N.R. Achuthan "*A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets*" 2007, Australian Computer Society, Inc.

[4] Alva Erwin, Raj P. Gopalan, and N.R. Achuthan "*Efficient Mining of High Utility Itemsets from Large Datasets"* PAKDD 2008, LNAI 5012, pp. 554–561, 2008. Springer-Verlag Berlin Heidelberg 2008

[5] ChowdhuryFarhan Ahmed, Syed KhairuzzamanTanbeer, *"An Efficient Candidate Pruning Technique for High Utility Pattern Mining"* T. Theeramunkong PAKDD 2009, LNAI 5476, 2009. Springer-Verlag Berlin Heidelberg 2009

[6] Vincent S. Tseng, Cheng-Wei Wu, Bai-En Shie, and Philip S. Yu "*UP-Growth: An Efficient Algorithm for High Utility Itemset Mining*" KDD'10, July 25–28, 2010, Washington, DC, USA. Copyright 2010 ACM 978-1-4503-0055-1/10/07

[7] S. Kannimuthu Dr. K. Premalatha *"iFUM - Improved Fast Utility Mining"* International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011

[8] Adinarayanareddy B. O. SrinivasaRao,"*An Improved UP-Growth High Utility Itemset Mining*" International Journal of Computer Applications (0975 – 8887) Volume 58– No.2, November 2012

[9] Arumugam P and Jose P "*Advance Mining of High Utility Item sets in Transactional Data*". International Journal of Business Management (e-ISSN: 2277-4637 and pISSN: 2231–5470) Special Issue on Role of Statistics in Management and Allied Sciences Vol. 3 No. 2 Dec. 2013, pg. 27- 40