Comparative Study on Order-Preserving Submatrices with repeated measurements and Bucket Order-Preserving SubMatrices

Pratiksha Kalbhor
MCA, University of Mumbai

Bhagyashree Nehete
MCA, University of Mumbai

***Abstract*--- The Order-Preserving SubMatrices (OPSMs) are employed to discover significant biological associations between genes and have been shown useful in capturing concurrent patterns in data when the relative magnitudes of data items are more important than their exact values. Bucket OPSM (BOPSM) model is a relaxed OPSM model by considering the linearity relaxation and OPSM-RM (OPSM with repeated measurements) is a robust version of OPSM, where each data item is represented by a set of values obtained from replicated experiments.**

Here we consider two mining algorithms: *HT-Bound algorithm* used to mine the OPSM-RM patterns and *APRIBOPSM* is developed to exhaustively mine BOPSM patterns. We show the effectiveness and efficiency of algorithms by comparing the result of these two algorithms by conducting series of experiments on real microarray data and presents the result. According to the result obtained we find that HT algorithm performs better than the APRIBOPSM algorithm.

***Keywords*-- Data mining , BOPSM, OPSM-Rm, Mining methods,Frequent patterns**

I. INTRODUCTION

Order-preserving SubMatrices (OPSMs) are mainly useful for finding unexpected change in the noisy data. The data is mostly presented in the form matrix of rows and columns. Rows in the matrix represent the set of data and columns in the matrix represent the set of conditions. OPSM is applied on this matrix and main aim is to find out the similar pattern of the values in the matrix.

First OPSM model was proposed by Ben-Dor et al. [1]. Main objective was to find out the data that follows same pattern under the certain set of conditions, but they don't have any relation with the other set of conditions.

Consider the matrix without Repeated Measurements as follows:

ISSN: 2278 – 1323

Conditions / Data	C1	C2	C3	C4	C5
R1	9	7	16	2	20
R2	10	8	20	6	1
R3	5	16	13	4	10
R4	6	3	9	2	6

The above matrix contains 4 rows and 5 columns. The values in rows R1, R2 and R3 follows the same sequential pattern C4< C2<C1<C3. So ({R1, R2, R3}, (C4, C2, C1, C3)) is an OPSM. For better understanding consider all values in the rows are unique.

Relaxed models of OPSM are as follows:

1. BOPSM (Bucket OPSM)

The BOPSM model proposed by Fang et al. (2012) [2] requires that all the data in a BOPSM pattern support a consensus bucket order of a set of conditions, in the sense that the condition values of a data in different buckets should maintain the ordering relationship between the buckets, and the condition values in the same bucket should be similar enough.

Here we consider APRIBOPSM Algorithm used to find out frequent patterns. The APRIBOPSM Algorithm is based on Apriori framework.

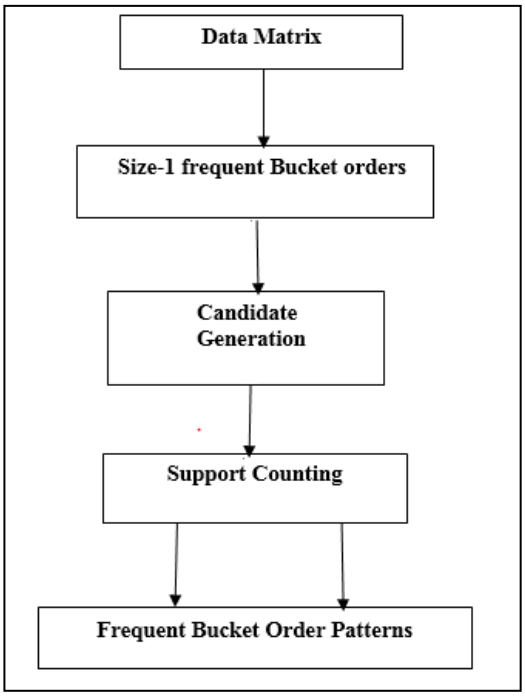


Fig.1 Architecture of APRIBOPSM

2. OPSM-RM (OPSM with Repeated Measurements)

Basic OPSM-RM is no strong against the noisy data. To deal with the errors, repeated experiments are carried out but the OPSM is failed to deal with the repeated measurements. So to deal with the repeated measurement's OPSM-RM is required. OPSM-RM is robust version of OPSM, where each data item is represented by a set of values obtained from replicated experiments.

Consider the matrix with Repeated Measurements as follows:

ISSN: 2278 – 1323

	A 1	A 2	A3	B1	B2	B 3	C1	C2	C3
R 1	49	55	80	38	51	81	11 5	10 1	79
R 2	67	54	13 0	96	85	82	12 4	92	94
R 3	65	49	62	67	39	28	13 2	11 9	83
R 4	81	83	10 5	11 5	11 0	87	13 3	10 8	10 5

In the above matrix contains that each column has 3 replicates (e.g. experiment in column "a" is repeated 3 times and therefore generating 3 sub columns : a1, a2, a3)

In OPSM-RM we consider HT-Bound algorithm to find out frequent patterns. Ben Kao states that HT-Bound is efficient and scalable algorithm [1].

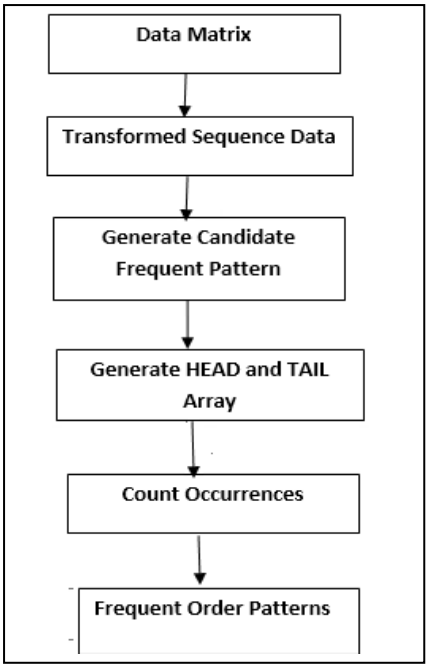


Fig.2 Architecture of HTBOUND

II. ALGORITHMS

1. APRIBOPSM Algorithm

In APRIBOPSM Algorithm,

1) In the first step, we find all size-1 frequent bucket orders which are simply all the single columns.
2) Then, generate candidate bucket order of size – k from frequent bucket order.
3) The number of supporting rows for each candidate bucket order is then counted.
4) The size-k candidate bucket orders that are supported by minimum support form the size-k frequent bucket order set.
5) Finally, we generate all the frequent bucket order patterns.

2. HT-Bound Algorithm

In HT-Bound algorithm,

1) In the first step, the data matrix by sorting each columns in each row w.r.t. their values in ascending order and replaces the entries by the column label.
2) Then the next task is to generate the frequent pattern occurred in the row and to obtain the number of occurrences of each label in the row.

ISSN: 2278 – 1323

3) Next Assuming the head and tail pattern from the frequent pattern considering there should be common pattern (mid pattern) between head and tail.
4) Now count the occurrences of mid pattern after the first occurrence of the column label of the head pattern. Repeat this step till last column label in the head pattern.
5) Now count the number of entries (occurrences) of tail pattern after the first occurrence of mid pattern. Continue this process till the last occurrence of the mid pattern.
6) Entries in the head array are associated with the rightmost entries in the tail array. To find the occurrences of frequent pattern in the row, add the entries in the tail array with respect to the entries in the head array.

OPSM Related Models

A Summary of OPSM Related Models are as follows:

Models	Matrix Types	Pattern Characteristics	References
OPSM	Real-valued matrices	Strict order-preserving	[Ben-Dor et al. 2002]
BOPSM	Real-valued matrices	Relaxed order-preserving	[Fang et al. 2012]
OPSM-RM	Set-valued matrices	Fractional support to order	[Yip et al. 2013]

III. Comparison between APRIBOPSM and HT-Bound Algorithm

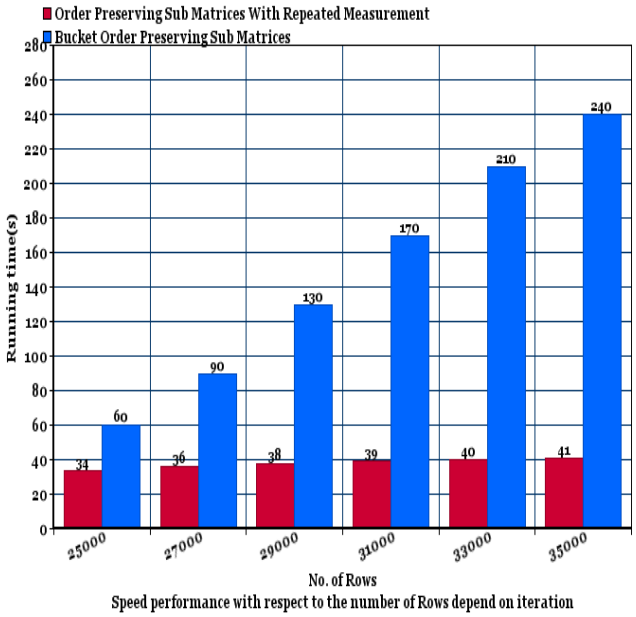
- APRIBOPSM and HT-Bound Algorithm are mainly used in data mining process to identify patterns and analyzing large amount of data.
- APRIBOPSM and HT-Bound Algorithm are based on generation of frequent patterns.
- APRIBOPSM Algorithm improves the quality of the mined patterns compared to the strict OPSM model.
- APRIBOPSM Algorithm is more efficient than the basic OPSM mining method.
- APRIBOPSM is more robust than the basic OPSM.
- APRIBOPSM uses information from previous steps to produce the frequent patterns.
- APRIBOPSM is easy to implement but still it has some limitations. In case of large dataset, APRIBOPSM algorithm is not efficient.

ISSN: 2278 – 1323

- This algorithm requires many database scans. In case of large dataset, APRIBOPSM algorithm produce large number of candidate patterns.
- Algorithm scan database repeatedly for searching frequent patterns, so more time and resource are required in large number of scans so it is inefficient in large datasets.
- In APRIBOPSM as the number of column increases, number of candidate patterns increases then it leads to time consuming process.
- APRIBOPSM is also does not deal with the repeated measurements of data.
- The original OPSM definition is not robust against noisy data.
- It also fails to take advantage of the additional information provided by replicates.
- HT-Bound is used to deal with errors, noisy data and experimental repeated values.
- HT-Bound algorithm is very efficient than the APRIBOPSM.
- HT-Bound is very scalable with respect to number of rows and columns than the APRIBOPSM.
- HT-Bound is faster even though there is large dataset available.
- HT-Bound is very robust than the APRIBOPSM and basic OPSM.

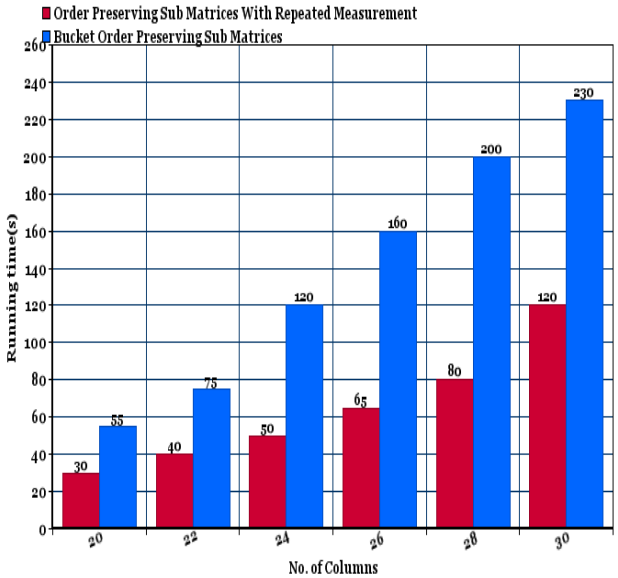
Comparison of APRIBOPSM and HTBound algorithm on the bases of speed performance with respect to rows and columns are as follows :

1) With respect to rows



Speed performance with respect to the number of Rows depend on iteration

2) With respect to Columns



Speed performance with respect to the number of columns depend on iteration

ISSN: 2278 – 1323

IV. Applications

- **Market Basket Analysis:** given a database of customer transactions, where each transaction is a set of items the goal is to find groups of items which are frequently purchased together.
- **Telecommunication** (each customer is a transaction containing the set of phone calls)
- **Credit Cards/ Banking Services** (each card/account is a transaction containing the set of customer's payments)
- **Medical Treatments** (each patient is represented as a transaction containing the ordered set of diseases)
- **Basketball-Game Analysis** (each game is represented as a transaction containing the ordered set of ball passes)

V. Conclusion

In this author tried to find the best data mining algorithm. From a broad variety of efficient data mining algorithms the most important ones are compared. The algorithms are systemized and their performance is analyzed based on runtime and theoretical considerations. On the bases of results author gives the advantages and limitations of HT-Bound and APRIBOPSM algorithm. And author compared the data mining algorithms i.e. APRIBOPSM and HT-Bound algorithm. after comparing Speed performance and number of frequent patterns with respect to number of Rows and Columns by these two data mining algorithms, author finds that HT-Bound algorithm is faster than APRIBOPSM algorithm.

Future Scope: Hence this algorithm can be used in other projects to bring out interestingness among the data present in the depository. HTBOUND algorithm can be combined for better results for any real life application like Hospital, Medical system etc. Algorithms can also be combined to form an efficient algorithm.

VI. References

[1] "Mining Order-Preserving Submatrices from Data with Repeated Measurements " Kevin Y. Yip, Ben Kao, Xinjie Zhu, Chun Kit Chui,Sau Dan Lee, and David W. Cheung, Senior Member, IEEE. VOL.25, NO. 7, JULY 2013

[2] "Mining Bucket Order-Preserving SubMatrices in Gene Expression Data " Qiong Fang, Student Member, IEEE, Wilfred Ng, Jianlin Feng, Member, IEEE, and Yuliang Li. VOL. 24, NO. 12, DECEMBER 2012.

ISSN: 2278 – 1323

[3]http://www.ncbi.nlm.nih.gov/sites/GDS browser/

[4] "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and
Comparative of Various Association Rule Algorithms " Ms Shweta,Dr. Kanwal Garg Volume 3, Issue 6, June 2013

[5] "Mining Order-Preserving SubMatrices from Probabilistic Matrices" QIONG FANG,
WILFRED NG, JIANLIN FENG Hong Kong University of Science and Technology

[6]" Discovering Local Structure in Gene Expression Data:
The Order-Preserving Submatrix Problem" Amir Ben-Dor_ Benny Chory Richard Karpz Zohar Yakhini

[7] S. Bleuler and E. Zitzler, "Order Preserving Clustering Over Multiple Time Course Experiments,"
Proc.Third European Conf. Applications of Evolutionary Computing (EC '05), pp. 33-43, 2005.

About Author

Pratiksha Hanumant Kalbhor currently pursuing MCA Final year from ASM's Institute of Management & Computer Studies, IMCOST, Thane which belongs to University of Mumbai, has an interest in learning new technologies. Research interests is in the area of data mining

Bhagyashree Vasudeo Nehete, currently pursuing MCA Final year from ASM's Institute of Management & Computer Studies (IMCOST), Thane, which belongs to University of Mumbai, has an interest in Software Engineering.

ISSN: 2278 – 1323