Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 8, August 2014, pg.597 – 601

RESEARCH ARTICLE

Proximate Mining Order- Sub Matrices Maintaining from Probabilistic Matrices

Anugu Rahul Reddy[1], Balakrishna Bangaru[2]
[1]M.Tech 2nd Year, Dept. of CSE, JNTU, HYDERABAD, INDIA
[2]Assistant Professor, Dept. of CSE, JNTU, HYDERABAD, INDIA
[1] anugu.rahulreddy@gmail.com; [2] balakrishna.bangaru@gmail.com

Abstract— Order-preserving submatrices (OPSMs) capture consensus trends over columns shared by rows in a data matrix. Mining OPSM patterns discovers important and interesting local correlations in many real applications, such as those involving biological data or sensor data. The prevalence of uncertain data in various applications, however, poses new challenges for OPSM mining, since data uncertainty must be incorporated into OPSM modeling and the algorithmic aspects.

Keywords— Data mining, Classification, bioinformatics, mining methods and Algorithms

I. INTRODUCTION

In gene expression analysis, the Order-Preserving Sub-Matrices (OPSMs) are employed to discover significant biological associations between genes and experiment conditions. Mining OPSMs has been extensively studied as a biclustering problem in the area of gene expression analysis. The gene expression data are usually presented as a matrix in which the rows correspond to a set of genes, the columns correspond to a set of experiment conditions, and the entries represent the expression levels of the genes under the conditions. The OPSM model, first proposed by Ben-Dor et al. [2], aims to capture the fact that the expression levels of a set of genes follow the same trend under a set of conditions, but show no obvious correlation under other conditions. Specifically, in a gene expression matrix, an OPSM consists of a subset of genes and a subset of conditions such that the expression levels of every gene induce the same linear order of the conditions. This linear order of the conditions represents the consensus trend that all the genes in an OPSM follow.

There are two basic principles that can be adopted to relax the OPSM model.

1) The induced orders of all the genes in the relaxed OPSM patterns are allowed to be only similarly ordered instead of identically ordered. We call this relaxation strategy similarity relaxation.
2) The induced orders of the genes are allowed to be an identical bucket order, where each bucket contains a set of conditions. We call this relaxation strategy linearity relaxation.

The main contributions of this paper are twofold.

1) We propose a novel BOPSM model and develop an efficient breadth-first method called APRIBOPSM, which exhaustively mines BOPSM patterns. Empirical studies show that the adoption of the BOPSM model improves the quality of the mined patterns compared to the strict OPSM model. The APRIBOPSM method is also significantly more efficient than the state-of-the-art OPSM mining method OPC-Tree, when it is used to mine strict OPSM patterns. This shows the robustness of the APRIBOPSM algorithm.
2) We further generalize the BOPSM model and propose the GeBOPSM model. Experiments show that, compared to all current relaxed OPSM models, the GeBOPSM model better captures the characteristics of noise-contaminated OPSM patterns in real gene expression data.
3) We develop an efficient mining method called SEEDGROWTH, which is able to mine sufficient GeBOPSM patterns by growing seed BOPSMs into maximal GeBOPSMs.

II. EXISTING SYSTEM

To synthesize additional replicates, for each gene and each experiment, we follow standard practice to model the values by a Gaussian distribution with the mean and variance equal to the sample mean and variance of the 4 replicates. The expression values of new replicates were then sampled from the Gaussian. New columns are synthesized by randomly drawing an existing column, fitting Gaussians as described, and sampling values from it. This way of construction mimics the addition of knockout experiments of genes in the same sub pathways of the original ones.

III. RELATED WORK

The OPSM model was originally proposed by Ben-Dor et al , which captures the fact that the expression levels of a set of genes exhibit similar trend under a set of experiment conditions. Experiments in show that, compared to several other types of biclustering models such as CC and Bimax, the OPSM model promotes the discovery of a larger fraction of biologically significant patterns based on a real biological dataset. However, it has also been recognized that the OPSM model may be too strict to be practical, since the real gene expression data are noisy. Thus, different relaxation approaches on the model are studied.

The AOPC model:

AOPC model relaxes the condition that all the rows in an OPSM should induce the same linear order of columns, and it requires only that a pre-specified fraction of rows induce the same linear or-der, while the induced orders of other rows only need to be “similar enough”. The ROPSM model proposed by Fang et al. [5] is a further relaxation of the AOPC model also based on “similarity”. The ROPSM model uses the backbone order to capture the consensus trend of an ROPSM pattern, and only requires that the induced orders of all the rows in the pattern be similar enough to the backbone order. Basically, both the AOPC model and the ROPSM model only consider the similarity relaxation approach.

ROPSM Model:

Mining OPSM patterns is shown to be an NP-complete problem in [2], and mining relaxed OPSMs accordingly becomes more difficult. BenDor et al. Proposed a model-based method, which aims to mine the best OPSM in terms of the statistical significance. Their method keeps a limited number of partial models which are smaller OPSMs, and then expands the partial models into larger OPSMs.

The AOPC mining method proposed in takes a set of OPSMs as input, and merges pairs of OPSMs into AOPCs in a greedy way until no more AOPCs can be generated. The ROPSM mining method proposed in similarly takes a set of OPSMs as input. Instead of merging OPSMs, it expands those seed OPSMs by adopting different growing strategies until maximal ROPSMs are reached.

IV. ALGORITHMS

Algorithm 1: APRIBOPSM

Input: Matrix M(G; T), d_{max}, g_{min}, c_{min}, r_{min}

1. F_1 =f size-1 frequent bucket orders g;
2. **for** (k = 2; $F_{k\ 1}$ 6= ; k + +) **do**
3. C_k= GENCAND($F_{k\ 1}$) ;
4. COUNTSUP(M; C_k);
5. F_k = f j 2 C_k; supp( ) r_{min}g ;
6. **if** k c_{min} && F_k 6= **then**
7. Output BOPSMs ;
8. **end**

Algorithm 2: GENCAND

1. **while** there are untraversed nodes in BPT ($F_{k\ 1}$) **do**
2. Traverse to item node p s.t. j [p]j = k 2;
3. **for** item node t_1, t_2 of p's children **do**
4. insert t_2 as a child of t_1;
5. **end**
6. **if** p has a boundary node q as child **then**
7. **for** p's child item node t_1 and q's child item node t_2 **do**
8. insert branch q ! t_2 under t_1 if t_2 6= t_1;
9. **end**
10. **for** item nodes t_1, t_2 of q's children **do**
11. insert branch q ! t_2 under t_1;
12. insert branch q ! t_1 under t_2;
13. **end**
14. **end**
15. **for** all newly inserted leaf nodes t **do**
16. **if** any size-(k 1) sub-order of [t] 2= $F_{k\ 1}$ **then**
17. undo insertion of node t;
18. **end**

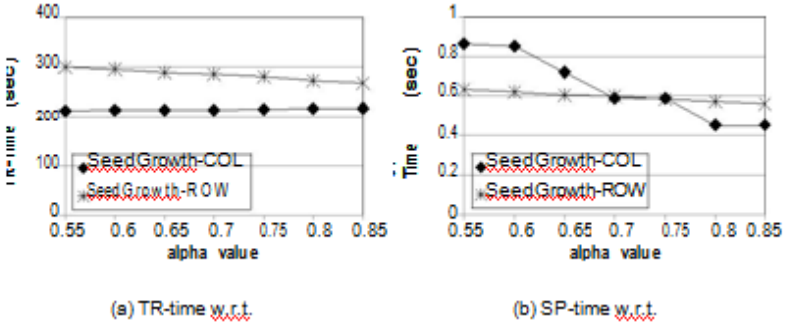
V. ALGORITHMS COMPARISON

We compare our algorithms with the state-of-the-art mining methods of other relaxed models, which include OPC-Tree, AOPC and OPSM-Growth as follows.

OPC-Tree is a tree-based OPSM mining method [11], which exhaustively mines OPSM patterns that satisfy some size thresholds.

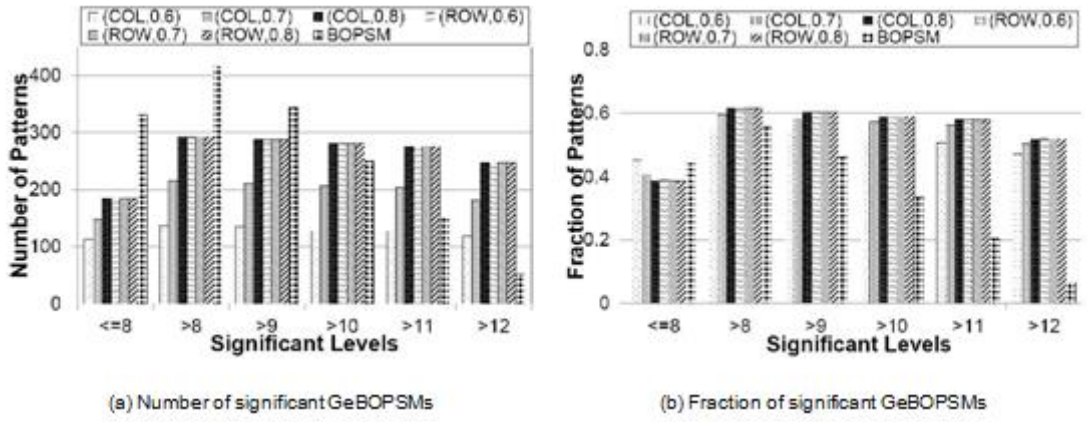
The AOPC method [20] mines AOPC patterns, which takes a set of OPSM patterns as input, and merges them into AOPCs in a greedy way until no more valid AOPCs can be generated.

The OPSM-Growth method [5] takes seed OPSMs as input and expands them into maximal ROPSM patterns.



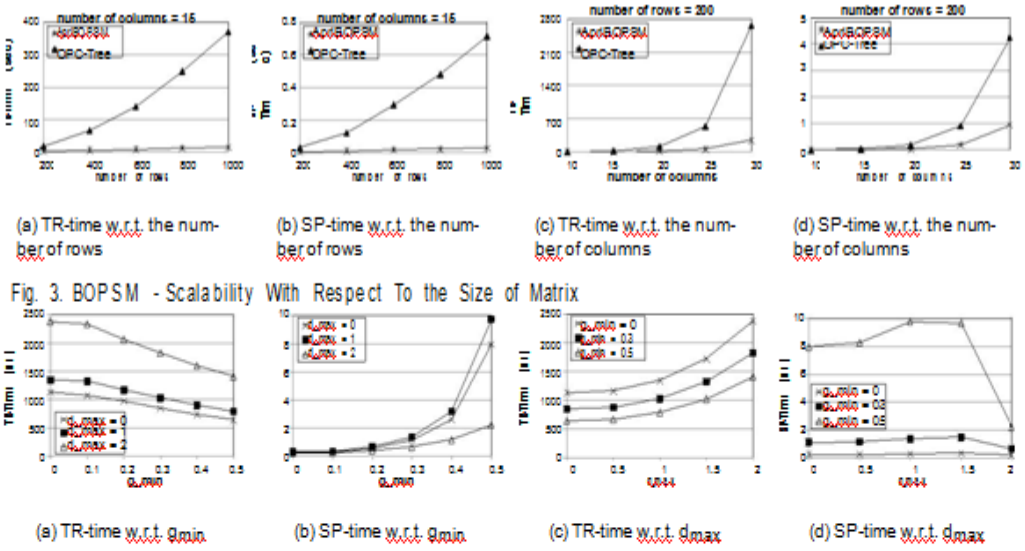
(a) TR-time w.r.t. (b) SP-time w.r.t.

FIGURE 1: GeBOPSM - Execution Time of Mined GeBOPSMs (g_{min} = 0:3; d_{max} = 0)



(a) Number of significant GeBOPSMs (b) Fraction of significant GeBOPSMs

FIGURE 2: GeBOPSM - Quality of Mined GeBOPSMs (g_{min} = 0:3; d_{max} = 0)

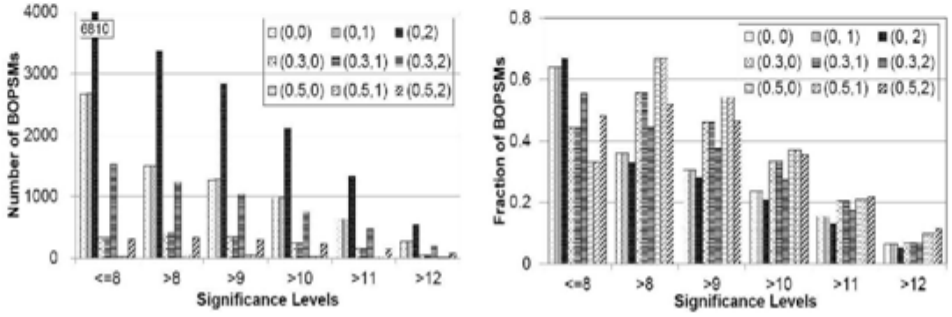


(a) TR-time w.r.t. the number of rows (b) SP-time w.r.t. the number of rows (c) TR-time w.r.t. the number of columns (d) SP-time w.r.t. the number of columns

Fig. 3. BOPSM - Scalability With Respect To the Size of Matrix

(a) TR-time w.r.t. g_{min} (b) SP-time w.r.t. g_{min} (c) TR-time w.r.t. d_{max} (d) SP-time w.r.t. d_{max}

Fig. 4. BOPSM - Execution Time With Respect To g_{min} and d_{max}



(a) Number of significant BOPSMS (b) Fraction of significant BOPSMs

FIGURE 3: BOPSM - Quality of BOPSMs with Respect To g_{min} and d_{max}

CONCLUSION

We propose the BOPSM model that captures the biological fact that some correlated genes follow a consensus trend identified as a bucket order. The model requires that the rows in a BOPSM support a backbone bucket order. The structure of the backbone bucket order can be monitored by the intra-bucket difference and inter-bucket gap thresholds. We also develop a BOPSM mining method called APRIBOPSM, which makes use of an Apriori, based framework and adopts a novel BucketPrefixTree structure to mine BOPSM patterns. Experiments on both synthetic and real datasets confirm that, the BOPSM model facilitates much better the discovery of quality but noisy OPSM patterns than the strict OPSM model. Our mining method is also significantly more efficient than OPC-Tree.

We propose the GeBOPSM model, which allows that bucket orders are similar enough to the backbone bucket order in an OPSM pattern. The GeBOPSM model generalizes all the existing relaxed OPSM models. We also develop the GeBOPSM mining method SEEDGROWTH and propose two different pattern growing strategies, i.e., column-centric or row-centric, to grow seed BOPSMs into GeBOPSM patterns. Experimental studies show that the GeBOPSM model outperforms all the current relaxed OPSM models, and importantly, it leads to the discovery of more quality patterns in terms of both fraction and number.

	(GeBOPSM,COL,0.7)	(ROPSM,ROW,0.6)	(ROPSM,COL,0.7)	(AOPC,8,0.6)	BOPSM	OPSM
Pattern number	459	134	252	294	81	747
TR-time	329:56 sec	209:64 sec	51:78 sec	55:66 sec		
SP-Time	0:72 sec	1:56 sec	0:21 sec	0:19 sec		
the number (fraction) of patterns that reach each significance level						
> 12	270(58:8%)	70(52:2%)	123(48:8%)	79(26:9%)	8(9:9%)	51(6:8%)
> 11	292(63:6%)	82(61:2%)	135(53:6%)	113(38:4%)	17(21:0%)	153(20:5%)
> 10	293(63:8%)	85(63:4%)	140(55:6%)	127(43:2%)	30(37:0%)	250(33:5%)
> 9	299(65:1%)	87(64:9%)	146(57:9%)	149(50:7%)	44(54:3%)	344(46:1%)
> 8	305(66:5%)	89(66:4%)	153(60:7%)	160(54:4%)	54(66:7%)	416(55:7%)
> 7	312(68:0%)	93(69:4%)	160(63:5%)	169(57:5%)	66(81:5%)	467(62:5%)
7	147(32:0%)	41(30:6%)	92(36:5%)	125(42:5%)	15(18:5%)	280(37:5%)

TABLE 1

Comparison of GeBOPSM, ROPSM, AOPC, BOPSM, and OPSM

REFERENCES

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In ICDE '95, pages 3–14, 1995.

[2] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. In RECOMB '02, pages 49–57, 2002.

[3] Y. Cheng and G. M. Church. Biclustering of expression data. In Proc. Int. Conf. Intell. Syst. Mol. Biol., pages 93–103, 2000.

[4] C. K. Chui, B. Kao, K. Y. Yip, and S. D. Lee. Mining order-preserving submatrices from data with repeated measure-ments. In ICDM '08, pages 133–142, 2008.

[5] Q. Fang, W. NG, and J. Feng. Discovering significant relaxed order-preserving submatrices. In SIGKDD '10, pages 433–442, 2010.

[6] B. J. Gao, O. L. Griffith, M. Ester, and S. J. M. Jones. Dis-covering significant opsm subspace clusters in massive gene expression data. In SIGKDD '06, pages 922–928, 2006.

[7] N. Gupta and S. Aggarwal. Mib: Using mutual information for biclustering gene expression data. Pattern Recognition, 2010.

[8] R. Gupta, N. Rao, and V. Kumar. Discovery of error-tolerant biclusters from noisy gene expression data. In BIOKDD '10, 2010.

[9] H.-P. Kriegel, P. Kroger,¨ and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Discov. Data, 3(1):1–58, 2009.

[10] G. Li, Q. Ma, H. Tang, A. Paterson, and Y. Xu. Qubic: a qual-itative biclustering algorithm for analyses of gene expression data. Nucleic acids research, 37(15), 2009.

[11] J. Liu and W. Wang. Op-cluster: Clustering by tendency in high dimensional space. In ICDM '03, 2003.

[12] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM TCBB, 1(1):24–45, 2004.

[13] G. Pandey, G. Atluri, M. Steinbach, C. L. Myers, and V. Kumar. An association analysis approach to biclustering. In SIGKDD '09, pages 677–686, 2009.

[14] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, and et al. Prefixs-pan: Mining sequential patterns efficiently by prefix-projected pattern growth. In ICDE '01, 2001.

[15] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, and et al. Mining se-quential patterns by pattern-growth: The prefixspan approach. IEEE TKDE, 16:1424–1440, 2004.

