International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Survey of Mining Order-Preserving Submatrices from Data with Repeated Measurements

Swati Gaikwad

Pune University, Zeal Educational Socity's, Dyanganga College of Engineering and Research, Pune, Maharashtra India

Abstract: ***There some situations, where relative magnitude of data items have more importance than their accurate values. In those situations, Order-preserving submatrices (OPSM's) are proving to be successful in detecting concurrent patterns in data. For example: relative magnitudes have importance in the process of analyzing gene expressions profiles which are extracted from microarray experiments. There are two reasons for it: 1. Relative magnitudes describe changes in gene activities among whole experiment. 2. In these situations, accurate values cannot be trusted since high level of noise is present. As data noise generating problems, Number of experiments has been carried out for collecting multiple measurements to address this issue. An advanced model of OPSM can be experimented and studied in which set of values obtained from replicated experiments are used to represent each data item. New problem can be called OPSM-RM (OPSM with Repeated Measurements). Depending on number of practical requirements, OPSM-RM can be characterized. Generic mining algorithm can be implemented by studying challenges of OPSM-RM.***

Keywords: Order-preserving submatrices, Data noise, Relative Magnitude, Data Mining and its methods and algorithm.

1. Introduction

Order-Preserving Submatrix has been a useful technique to identify groups of genes that have some common functions. The aim of OPSM is to discover a subset of genes that exhibit concurrent rises and falls of their expression values across different experiments. For example, genes with concurrent changes of mRNA expression levels across different time points may share the same cell-cycle related properties [2]. Instead of comparing their absolute values, it is more convenient to compare relative expression levels of different genes at different time intervals, because normal microarray data contains high noise. The relative magnitudes of data are considered in OPSM.OPSM find correspondent genes that share same functions. The original OPSM problem was first proposed by Ben-Dor and company [3]

Definition 1: Given an nXm matrix (data set) D, an order-preserving submatrix is a pair (R,P), where R is a subset of then rows (represented by a set of row ids)and P is a permutation of a subset of the m columns (represented by a sequence of column ids) such that for each row in R, the data values are monotonically increasing with respect to P, i.e.,DiPj< DiPj [9].

Table 1: A data set

Margin	*a*	*b*	*C*	*d*
R1	45	37	114	80
R2	65	97	122	45
R3	66	68	133	93
R4	80	112	131	63

For example, Table 1 shows a dataset with 4 rows and 4 columns. The values of rows 2, 3 and 4 increase from a to b, so ({2, 3, 4}, <a, b>) is an OPSM. In this study we assume that all values in a row are unique. We say that a row supports a permutation if its values increase monotonically with respect to the permutation.

In the above example, rows 2, 3 and 4 support the permutation <a, b>, but row 1 does not. For a fixed dataset, the rows that support a permutation can be unambiguously identified. In the following discussion, we will refer to an OPSM simply by its variation which will also be called a pattern. An OPSM is said to be frequent if the number of supporting rows is not less than a support threshold, ρ. An OPSM mining problem is that it is vulnerable to noisy data. In our above example, if the value of column a is slightly increased in row 3, say from 66 to 69, then row 3 will no longer support the pattern <a, b >, but will support <b,a> instead.

1. An error can cause a value of replicate to deviate from that of other replicates. Hence, the support for given pattern should not be affected by that replicate.
2. The overall support should reflect the uncertainty when replicates hugely disagree on their support of pattern.

Here, initial two requirements can be accomplished by following steps:

1. Summarize replicates by strong statistics like medians
2. Mining of resulting data set for utilizing original definitions of OPSM.

But last requirement cannot be accomplished by single summarizing statistics. The reason behind it, a row can fully support or fully not support a pattern. And so information of uncertainty is vanished.

To overcome this issue a concept of fractional support can be used. All of the three requirements can be satisfied if Fractional support is used for indicating how much a row supports to OPSM. The fractional support will maximum, if all possible replicates groupings of row support a particular pattern.

To reduce errors, multiple experiments are taken and multiple measurements are recorded. Better approximation of real physical quantity is allowed by those replicates. In some of the microarray datasets, each experiment is repeated three times to produce three measurements of every data

point. Researchers are obtaining replicates for getting higher quality of data. This is all because cost of microarray experiments is dropping. In usual study of microarray data sets, every experiment is carried out three times to obtain three measurements of each data inputs. Recent studies have stated that to improve quality of data, multiple replicates are important and necessary [5]. There may be chances of different replicates supporting different OPSM.

Basic OPSM is not strong against noisy data. Even it has failed to make use of extra data that has been provided by replicates. This shows that definition of OPSM should be improved so that it can deal with repeated measurements. There are few conditions or requirements which should be followed by new improved OPSM:

3. Row should contribute high support for a particular pattern if that pattern is supported by all grouping of available replicates. The small change will be done in fractional support if one of the replicates of column deviates from others. The fractional support will be fuzzy that depicts uncertainty, if only fractional of replicate groupings support a particular pattern.

2. Literature Review

An OPSM is specified by a cluster of rows and a cluster of columns, where the cluster of genes display synchronous rises and falls of their expression values along the cluster of experiments. In recent years there exist many studies that address the issue of mining sets of genes that share compatible expression patterns over sets of experiments. In this section we review a few classic clustering methods of gene expression data. We go through several previous OPSM approaches.

2.1 Previous OPSM Approaches

Most existing clustering methods for micro-array data cluster genes by comparing genes' expression levels in all experiments, or cluster experiments by comparing experiments' expression levels for all genes. Going beyond such global approaches, Ben-Dor et al. originally motivated and introduced the conventional order-preserving submatrices (OPSM) mining problem [3]. They sought local patterns, in which the expression levels of a subset of genes induce the same linear ordering of a subset of experiments. The OPSM mining problem was proven to be NP-hard in their work. By defining a probabilistic model, they developed a greedy heuristic algorithm for finding the hidden OPSM's in the random matrix. But their algorithm did not guarantee retrieval of all OPSM's or the best OPSM's. The measurements of different experiments are often mixed together in a single gene expression matrix. The mixture of different types of experiments may blur the clustering outcome, and does not allow study into similarities and differences between the distinct experiment groups. The problem caused by the mixture of different experiment groups was solved by Bleuler et al. [6]. In their work, they proposed a flexible cluster scoring scheme that allows to arbitrarily scaling the degree of orderedness required for a cluster. Based on the scoring scheme, they presented an evolutionary algorithm for mining patterns that are consistent over multiple time course experiments.

In the study by Gao et al. [7], they pointed out that OPSM mining is reducible to a special case of the sequential pattern mining problem, in which a pattern and its supporting sequences uniquely specify an OPSM cluster. They called the clusters specified by long patterns but few supporting sequences twig clusters. Biologists may be interested in the twig clusters. Due to the high computational cost and low support, those twig clusters are usually omitted by most existing approaches. So they introduced the KiWi framework for massive datasets, focusing on twig cluster mining. KiWi exploits two parameters k and w to perform a biased testing on a bounded number of candidates, keeping only highly promising seeds that will likely produce twig clusters and significant clusters.

Cheung et al. [8] proved the monotonic and transitivity properties of OPSM's. Based on the properties, they proposed an iterative generation-and-verification framework to find OPSM's. In the generation phase of each iteration, head-tail trees are employed to generate candidate patterns. Every head/tail tree is a prefix tree, where each leaf contains a reference to some frequent patterns.

The ROPSM model proposed by Fang et al. (2010) [9] is a further relaxation of the OPSM model. IT takes a set of OPSMs as input. It expands those seed OPSMs by adopting different growing strategies until maximal ROPSMs are reached.

The BOPSM model proposed by Fang et al. (2012) [10] requires that all the genes in a BOPSM pattern support a consensus bucket order of a set of conditions, in the sense that the condition values of a gene in different buckets should maintain the ordering relationship between the buckets, and the condition values in the same bucket should be similar enough.

However, none of the above studies on OPSM considered data with repeated measurements. Chui et al. [11] proposed the concept of OPSM-RM (OPSM with repeated measurements) to mine OPSM's from data with replicates. In their study they discussed the computational challenges of OPSM-RM and presented a generic mining algorithm.

3. Conclusion

In this survey paper, we discussed problem and already implemented techniques of Mining Order-Preserving Submatrices. High Noise is major issue in mining of OPSM problem. Traditional OPSM is not strong against noisy data. The solution is repeated measurements with OPSM. The definition of OPSM should be improved so that it can deal with repeated measurements.

References

[1] C.K. Chui, B. Kao, K.Y. Yip, and S.D. Lee, “Mining Order-Preserving Submatrices from Data with Repeated Measurements,” Proc. IEEE Eighth Int’l Conf. Data Mining (ICDM ’08), pp. 133-142,2008.
[2] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B.

Futcher, “Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization,” Molecular Biology of the Cell, vol. 9, no. 12, pp. 3273-3297, 1998.

[3] A. Ben-Dor, B. Chor, R.M. Karp, and Z. Yakhini, “Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem,” J. Computational Biology, vol. 10, nos. 3/4, pp. 373-384, 2003..

[4] J. Liu and W. Wang, “OP-Cluster: Clustering by Tendency in High Dimensional Space,” Proc. IEEE Third Int’l Conf. Data Mining, pp. 187-194, 2003.

[5] M.-L.T. Lee, F.C. Kuo, G.A. Whitmore, and J. Sklar, “Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence From Repetitive Cdna Hybridizations,” Proc. Nat’l Academy of Sciences USA, vol. 97, no. 18, pp. 9834-9839, 2000.

[6] 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.

[7] B.J. Gao, O.L. Griffith, M. Ester, and S.J.M. Jones, “Discovering Significant Opsm Subspace Clusters in Massive Gene Expression Data,” Proc. 12th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 922-928, 2006.

[8] L. Cheung, K.Y. Yip, D.W. Cheung, B. Kao, and M.K. Ng, “On Mining Micro-Array Data by Order-Preserving Submatrix,” Int’l J. Bioinformatics Research and Applications, vol. 3, no. 1, pp. 42-64, 2007.

[9] Kelvin Yip, Ben Kao, Xinjie Zhu, Chun Kit Chui, Sau D. Lee and David W. Cheung, Senior Member, IEEE"Mining Order Preserving Submatrices with Repeated Measurements(2013)

[10] H. Wang, W. Wang, J. Yang, and P.S. Yu, “Clustering by Pattern Similarity in Large Data Sets,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’02), pp. 394-405, 2002.

[11] C.K.Chui, B. Kao, K.Y. Yip,"Mining Oreder Preserving Submatrices feom data with Reapeted Measurments", Proc. IEEE Eigth Int'l Conf. Data Mining (ICDM' 08), pp. 133-142,2008.