ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

A Survey on Evolutionary Co-Clustering Formulations for Mining Time-Varying Data Using Sparsity Learning

R.Amsaveni[1], R. Suresh Kumar MCA, MPhil [2]

M.Phil Scholar, PG, Department of Computer Applications, Sree Saraswathi Thyagaraja College, Pollachi, India[1]

Assistant Professor, PG, Department of Computer Applications, Sree Saraswathi Thyagaraja College, Pollachi India[2]

ABSTRACT: The data matrix is considered as static in Traditional clustering and feature selection methods. However, the data matrices evolve smoothly over time in many applications. A simple approach to learn from these time-evolving data matrices is to analyze them separately. Such strategy ignores the time-dependent nature of the underlying data. Two formulations are proposed for evolutionary co-clustering and feature selection based on the fused Lasso regularization. The evolutionary co-clustering formulation is able to identify smoothly varying data embedded into the matrices along with the temporal dimension. Formulation allows for imposing smoothness constraints over temporal dimension of the data matrices. The evolutionary feature selection formulation can uncover shared features in clustering from time-evolving data matrices.

KEYWORDS: Time-varying data, Sparsity learning, co-clustering, feature selection, temporal smoothness.

I. INTRODUCTION

Clustering is an important explorative statistical analysis of gene expression data. It aims to identify and group genes that exhibit similar expression patterns over several conditions and also group the conditions based on the expression profiles across set of genes. The successful clustering approach should guarantee two criteria which are homogeneity high similarity between elements in the same cluster, and separation – low similarity between elements from different clusters.

Traditional clustering approaches such as k-means and hierarchical clustering put each gene in exactly one cluster based on the assumption that all genes behave similarly in all conditions. However, recent understanding of cellular processes shows that it is possible for subset of genes to be co expressed under certain experimental conditions, and at the same time; to behave almost independently under other conditions. From this context, a new two mode clustering approach called bi-clustering or co-clustering has been introduced to group the genes and conditions in both dimensions simultaneously.

As a class of powerful methods for unsupervised pattern mining, existing co-clustering methods invariably assume that the data matrices are static; that is, they do not evolve over time. However, in many real-world domains, the processes that generated the data are time-evolving. Hence, the observed data are usually dynamic. As a consequence, the block structures embedded into the time-varying data should also evolve smoothly over time. Therefore, it is desirable to incorporate the temporal smoothness constraint into the co-clustering formalism.

The proposed formulation employs sparsity-inducing regularization to identify block structures from the time-varying data matrices. More specifically, it applies fused Lasso type of regularization to encourage temporal smoothness over the block structures identified from contiguous time points. The proposed formulation is very flexible and can be applied to encourage temporal smoothness over either one or both dimensions of the data matrices.

DOI: 10.15680/IJIRCCE.2015. 0308006

II. RELATED WORK

S. Alelyani, J. Tang, and H. Liu[1] presented a review on feature selection for clustering as Nowadays data are mostly high dimensional data. Dimensionality reduction is one of the popular technique to remove noisy (i.e.) irrelevant) and redundant attributes. There are two types of dimensionality reduction that is feature selection and feature extraction. Clustering is one of the important data mining tasks. Different features affect clusters differently. Some are important for clusters while others may hinder the clustering task. Important features are selected for clustering.

D. Chakrabarti, R. Kumar, and A. Tomkins [2], described that Evolutionary clustering is the problem of processing time-tamped data to produce a sequence of clustering; that is, a clustering for each time step of the system. Each clustering in the sequence should be similar to the clustering at the previous time step, and should accurately reflect the data arriving during that time step. Every day, new data arrives for the day, and must be incorporated into a clustering.

Y. Cheng and G. M. Church[3], introduced an efficient node-deletion algorithm to find sub matrices in expression data that have low mean squared residue scores and it is shown to perform well in finding co-regulation patterns in yeast and human. This introduces "bi-clustering', or simultaneous clustering of both genes and conditions, to knowledge discovery from expression data. This approach overcomes some problems associated with traditional clustering methods, by allowing automatic discovery of similarity based on a subset of attributes, simultaneous clustering of genes and conditions, and overlapped grouping that provides a better representation for genes with multiple functions or regulated by many factors.

M. Lee, H. Shen, J. Z. Huang, and J. S. Marron[4], describes that Sparse singular value decomposition (SSVD) is proposed as a new exploratory analysis tool for bi-clustering or identifying interpretable row–column associations within high-dimensional data matrices. SSVD seeks a low-rank, checkerboard structured matrix approximation to data matrices. The desired checkerboard structure is achieved by forcing both the left- and right-singular vectors to be sparse, that is, having many zero entries. By interpreting singular vectors as regression coefficient vectors for certain linear regressions, sparsity-inducing regularization penalties are imposed to the least squares regression to produce sparse singular vectors.

H. Cho, I. S. Dhillon, Y. Guan, and S. Sra[5],says that Microarray experiments have been extensively used for simultaneously measuring DNA expression levels of thousands of genes in genome research. A key step in the analysis of gene expression data is the clustering of genes into groups that show similar expression values over a range of conditions. Since only a small subset of the genes participate in any cellular process of interest, by focusing on subsets of genes and conditions, lower the noise induced by other genes and conditions.

Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng[6] concluded that Evolutionary clustering is an emerging research area essential to important applications such as clustering dynamic Web and blog contents and clustering data streams. In evolutionary clustering, a good clustering result should fit the current data well, while simultaneously not deviate too dramatically from the recent history. To fulfill this dual purpose, a measure of temporal smoothness is integrated in the overall measure of clustering quality. Proposed frameworks incorporate temporal smoothness in evolutionary spectral clustering. Solutions to the evolutionary spectral clustering problems provide more stable and consistent clustering results that are less sensitive to short-term noises while at the same time are adaptive to long-term cluster drifts.

R. Tibshirani and M. Saunders[7] described that the lasso penalizes a least squares regression by the sum of the absolute values (L1-norm) of the coefficients. The form of this penalty encourages sparse solutions (with many coefficients equal to 0). We propose the 'fused lasso', a generalization that is designed for problems with features that can be ordered in some meaningful way. The fused lasso penalizes the L1-norm of both the coefficients and their successive differences. Thus it encourages sparsity of the coefficients and also sparsity of their differences—i.e. local constancy of the coefficient profile. The technique is also extended to the 'hinge' loss function that underlies the support vector classifier.

Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H.Liu[8] says that the rapid advance of computer based high-throughput technique has been provided unparalleled opportunities for humans to expand capabilities in production, services, communications, and research. Meanwhile, immense quantities of high-dimensional data are accumulated challenging state-of-the-art data mining techniques. Feature selection is an essential step in successful data mining applications, which can effectively reduce data dimensionality by removing the irrelevant (and the redundant) features.

L. Wasserman, M. Azizyan, and A. Singh[9] concluded thata nonparametric method is used for selecting informative features in high-dimensional clustering problems. It starts with a screening step that uses a test for multimodality. Then we apply kernel density estimation and mode clustering to the selected features. The output of the method consists of a list of relevant features, and cluster assignments. Explicit bounds on the error rate of the resulting clustering are provided. In addition, we provide the first error bounds on mode based clustering.

M. Deodhar and J. Ghosh[10] says thatFor difficult classification or regression problems, practitioners often segment the data into relatively homogeneous groups and then build a predictive model for each group. This two-step procedure usually results in simpler, more interpretable and actionable models without any loss inaccuracy. In this work, we consider problems such as predicting customer behavior across products, where the independent variables can be naturally partitioned into two sets, that is, the data is dynamic in nature. A pivoting operation now results in the dependent variable showing up as entries in a “customer by product” data matrix. We present the Simultaneous CO-clustering And Learning (SCOAL) framework, based on the key idea of interleaving co-clustering and construction of prediction models to iteratively improve both cluster assignment and fit of the models. This algorithm provably converges to a local minimum of a suitable cost function. The framework not only generalizes co-clustering and collaborative filtering to model-base clustering, but can also be viewed as simultaneous co-segmentation and classification or regression, which is typically better than independently clustering the data first and then building models. Moreover, it applies to a wide range of bi-modal or multimodal data, and can be easily specialized to address classification and regression problems.

III. PROPOSED SYSTEM

Mining streaming data has been an active re-search area to address requirements of many applications. The proposed a new popular technique for mining time-varying data continuous and fast-growing data streams based on fused Lasso regularization with tune parameters and smoothness generation with genetic algorithm.

The evolutionary co-clustering formulation is able to identify smoothly varying hidden block structures embedded into the matrices along the temporal dimension. Our formulation is very flexible and allows for imposing smoothness constraints over the whole dimensions of the data matrices. The evolutionary feature selection formulation can uncover shared features in clustering from time-evolving data matrices. The proposed systems show that the optimization problems involved are non-convex, non-smooth and non-separable.

IV. PROPOSED ALGORITHM

We propose an iterative two-step procedure to compute the solution of the general optimization problem. The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution.

```
Algorithm:  Fused Lasso Regularization with Genetic Algorithm
Input: Time series dataset I with $m \times n$ dimension, Cluster s, $s \in \hat{} R \cup C$, Covariance distribution I ( X’, Y’).
Output: Set of co-clusters

Step 1:
        Begin with a random co-clustering I(X, Y) where X and Y are which could lead to poor local minima.
Repeat
        Step (i): Update lasso parameter co-cluster models, $\forall$[g]k1, [h]l1,
                  Update statistics for co-cluster (g, h) based on basis cluster C to compute new z values
        Step (ii): if s is a column cluster then
```

```
            I(X, Y ) = I(X, Y )T
        Step (iii): Randomly split s into two clusters, s1 and s2
        Step (iv): Update the column cluster value for Genetic to fitness is assigned to each features
    Step 2: Post-process
        for all xi ∈ s do
        Assign xi to cluster s′ , where s′ = argminj=1,2 Distance (I( Y |xi)||I(Y |sj))
        Update I(g |s1), I(h |s2) and g, h accordingly
        until I converges
        return s1, s2 and _I
    return identified co-clusters.
```

V. Conclusion

In this survey paper, we surveyed most recent studies on the sparsity learning of time-varying data series clustering. These studies are classified into three major categories depending upon whether they work directly with the unsupervised data and time series data with features extracted from the raw data, or indirectly with models built from the raw data. The basics of time varying traditional clustering, including the three key components of time series clustering studies are high-lighted in this survey: the clustering algorithm, the distance measure, and the classification criterion. The application areas are summarized with a brief description of the data used. The uniqueness and limitation of past studies, and some potential topics for future study are also discussed

References

1. S. Alelyani, J. Tang, and H. Liu, "Feature selection for clustering: A review," Data Clustering: Algorithms and Applications, C. Aggarwal, and C. Reddy, Eds., Boca Raton, FL, USA: CRC Press.
2. D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary clustering," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov.Data Min., 2006, pp. 554–560.
3. Y. Cheng and G. M. Church, "Biclustering of expression data," in Proc. Eighth Int. Conf. Intell. Syst. Mol. Biol., 2000, pp. 93–103.
4. M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, "Biclustering via sparse singular value decomposition," Biometrics, vol. 66, pp. 1087–1095, 2010.
5. H. Cho, I. S. Dhillon, Y. Guan, and S. Sra, "Minimum sum-squared residue co-clustering of gene expression data," in Proc. Fourth SIAM Int. Conf. Data Min., 2004, pp. 114–125.
6. Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng, "On evolutionary spectral clustering," ACM Trans. Knowl. Discov. Data, vol. 3,pp. 17:1–17:30, Dec. 2009.
7. R. Tibshirani and M. Saunders, "Sparsity and smoothness via the fused lasso," J. Royal Statist. Soc. B, vol. 67, no. 1, pp. 91–108, 2005.
8. Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H.Liu, "Advancing feature selection research-asu feature selectionrepository," School Comput., Informatics Decision Syst. Eng., ArizonaState Univ., Tempe, AZ, USA, 2010.
9. L. Wasserman, M. Azizyan, and A. Singh, "Feature selection for high-dimensional clustering," arXiv preprint arXiv:1406.2240, 2014.
10. M. Deodhar and J. Ghosh, "SCOAL: A framework for simultaneous co-clustering and learning from complex data," ACM Trans.Knowl. Discov. Data, vol. 4, no. 3, pp. 11:1–11:31, 2010.