Simultaneous Clustering: A Survey

Malika Charrad and Mohamed Ben Ahmed

National School of Computer Science
Manouba University, Tunisia
{malika.charrad,mohamed.benahmed}@riadi.rnu.tn
http://www.riadi.rnu.tn/

Abstract. Although most of the clustering literature focuses on one-sided clustering algorithms, simultaneous clustering has recently gained attention as a powerful tool that allows to circumvent some limitations of classical clustering approach. Simultaneous clustering methods perform clustering in the two dimensions simultaneously. In this paper, we introduce a large number of existing simultaneous clustering approaches applied in bioinformatics as well as in text mining, web mining and information retrieval and classify them in accordance with the methods used to perform the clustering and the target applications.

Keywords: Simultaneous clustering, Biclusters, Block clustering.

1 Introduction

Simultaneous clustering, usually designated by biclustering, co-clustering, 2-way clustering or block clustering, is an important technique in two-way data analysis. A number of algorithms that perform simultaneous clustering on rows and columns of a matrix have been proposed to date. The goal of simultaneous clustering is to find sub-matrices, which are subgroups of rows and subgroups of columns that exhibit a high correlation. This type of algorithms has been proposed and used in many fields, such as bioinformatique [23], web mining [8], text mining [3] and social network analysis [18]. A wide range of different articles were published dealing with different kinds of algorithms and methods of simultaneous clustering. Comparisons of several biclustering algorithms can be found, in [23] and [27]. However, these comprehensive surveys focus only on algorithms used to genetic data analysis. In this paper, we give a brief description of a large number of existing approaches to biclustering including approaches based on mixture model, and those based on information theory.

2 Simultaneous Clustering Problem

Clustering is the grouping together of similar subjects. Standard clustering techniques consider the value of each point in all dimensions, in order to form group of similar points. This type of one-way clustering techniques is based on similarity between subjects across all variables.

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 370–375, 2011.

Simultaneous clustering algorithms seeks "blocks" of rows and columns that are interrelated. They aim to identify a set of biclusters $B_k(I_k, J_k)$, where I_k is a subset of the rows X and J_k is a subset of the columns Y. I_k rows exhibit similar behavior across J_k columns, or vice versa and every bicluster B_k satisfies some criteria of homogeneity. A biclustering method may assume a specific structure and data type. Madeira and Oliveira introduce in their survey [23] some biclustering structures defined by : single bicluster, exclusive rows biclusters, exclusive columns biclusters, nonoverlapping biclusters with tree structure, and arbitrarily positioned overlapping biclusters. Biclusters can be with constant values, with constant values on rows or columns, with coherent values or with coherent evolution. There are many advantages in a simultaneous rather than one way clustering (table 1). In fact, simultaneous clustering may highlight the association between the row and column clustering that appears from the data analysis as a linked clustering. Furthermore, it allows the researcher to deal with sparse and high dimensional data matrices [2]. Simultaneous clustering is also an interesting paradigm for unsupervised data analysis as it is more informative, has less parameters, is scalable and is able to effectively interwine row and column information.

Table 1. Comparison between Clustering and Simultaneous clustering

Clustering	**Simultaneous Clustering**
- applied to either the rows or the columns of the data matrix **separately** ⇒ **global model**.	- performs clustering in the two dimensions **simultaneously** ⇒ **local model**.
- produce **clusters** of rows **or** clusters of columns.	seeks **blocks** of rows **and** columns that are interrelated.
- Each subject in a given subject cluster is defined using **all** the variables. Each variable in a variable cluster characterizes **all** subjects.	- Each subject in a bicluster is selected using **only** a subset of the variables and each variable in a bicluster is selected using **only** a subset of the subjects.
- Clusters are **exhaustive**	- The clusters on rows and columns **should not be exclusive** and/or **exhaustive**

3 Simultaneous Clustering Approaches

A survey of simultaneous clustering algorithms applied on biological data has been given by Madeira and Oliveira in 2004 [23]. These alogorithms are based on five approaches : Iterative Row and Column Clustering Combination (IRCCC), Divide and Conquer (DC), Greedy Iterative Search (GIS), Exhaustive Bicluster Enumeration (EBE) and Distribution Parameter Identification (DPI). The IRCCC approach consists to apply clustering algorithms to the rows and columns of the data matrix, separately, and then to combine results using some sort of iterative procedure. The algorithms based on DC approach begin with the entire data in one block (bicluster) and identifies biclusters at each iteration by splicing a given block into two pieces. GIS approach creates biclusters by

adding or removing rows/columns from them, using a criterion that maximizes the local gain. EBE approach identifies biclusters using an exhaustive enumeration of all possible biclusters in the data matrix. DPI approach assumes that the biclusters are generated using a given statistical model and tries to identify the distribution parameters that fit the available data, by minimizing a certain criterion through an iterative approach. All the algorithms presented in this survey analyze biological data from gene expression matrices. Given that there are a number of algorithms based on bipartite graph model ([12] [3]), mixture model [26] and information theory ([13] [30]), which are applied in other fields such as text mining, web mining and information retrieval, we propose to categorize simultaneous clustering methods into five categories : bipartite Graph methods, variance minimization methods, two-way clustering methods, motif and pattern recognition methods and probabilistic and generative methods.

- The bipartite graph methods consists in modeling rows and columns as a weighted bipartite graph and assigning weights to graph edges using similarity measure techniques. The created bipartite graph is then partitioned in a way that minimizes the cut of the partition, i.e. the sum of the weights of the crossing edges between parts of the partition. In [38], the authors created a word-document bipartite graph. The graph was partitioned using a partial singular value decomposition of the associated edge weight matrix of the bipartite graph. Dhillon [12] used the spectral method for partitioning the bipartite graph constructed in the same way as in [38]. [28] proposed an isoperimetric co-clustering algorithm (ICA) for partitioning the word-document matrix. ICA used the same model than spectral partitioning but instead of searching the solutions of the singular word-document system of linear equations, it converts the system to a nonsingular system of equations which is easier to solve. The bipartite graph methods are also used for gene expression analysis. One example is Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) [31].

- The variance minimization methods define clusters as blocks in the matrix with minimal deviation of their elements. This definition has been already considered by Hartigan (1972) [20] and extended by Tibshirani et al. [33]. Some examples are the δ-cluster methods, such as δ-ks clusters [7], δ-pClusters [35] and δ-biclusters [10], which search for blocks of elements having a deviation below δ. FLexible Overlapped biClustering (FLOC) introduced by [36] extend Cheng and Church δ-biclusters by dealing with missing values.

- Two-way clustering methods use one-way clustering such as kmeans [32] [17] [9] [26], Self Organizing Maps [5], Expectation-Minimization algorithm [16] or an hierarchical clustering algorithm [15] to produce clusters on both dimensions of the data matrix separately. One-dimension results are then combined to produce subgroups of rows and columns called biclusters. These methods identify clusters on rows and columns but not directly biclusters.

– Motif and pattern recognition methods define a bicluster as samples sharing a common pattern or motif. To simplify this task, some methods discretize the data such as xMOTIF [25] or binarize the data such as Bimax [27]. Order-Preserving SubMatrices (OPSM) [4] searches for blocks having the same order of values in their columns. Spectral clustering (SPEC) [37] performs a singular value decomposition of the data matrix after normalization. Contiguous column coherent (CCC biclustering) [24] is a method for gene expression time series, which finds patterns in contiguous columns.

Table 2. Simultaneous clustering algorithms

Algorithm	Application	Approach	Data type
Two-way splitting [20]	Other	Variance minimization	Continuous
CROEUC [17]	Other	Two-way clustering	Continuous
CROKI2 [17] [9]	Other	Two-way clustering	Categorical
CROBIN [17]	Other	Two-way clustering	Binary
CTWC [15]	Bioinformatique	Two-way clustering	Continuous
Plaid Models [22]	Bioinformatique	Probabilistic and generative	Continuous
δ-biclusters [10]	Bioinformatique	Variance minimization	Continuous
δ-ks patterns [7]	Bioinformatique	Variance minimization	Continuous
ITWC [32]	Bioinformatique	Two-way clustering	Continuous
DCC [5]	Bioinformatique	Two-way clustering	Continuous
OPSM [4]	Bioinformatique	Motif and pattern recognition	Continuous
SAMBA [31]	Bioinformatique	Probabilistic and generative	Continuous
FLOC [36]	Bioinformatique	Variance minimization	Continuous
Spectral [37]	Bioinformatique	Motif and pattern recognition	Continuous
IT [13]	Text Mining	Probabilistic and generative	Continuous
BSGP [12]	Text Mining	Bi-partite Graph	Categorical
cHawk [1]	Bioinformatique	Bi-partite Graph	Continuous
[30]	Other	Probabilistic and generative	Categorical
Block-EM [16]	Other	Two-way clustering	Continuous binary
Block-CEM [16]	Other	Two-way clustering	Continuous binary
Cemcroki2 [26]	Other	Two-way clustering	Categorical

– Probabilistic and generative methods use model-based techniques to define biclusters [21]. Probabilistic Relational Models (PRMs) [14] and their extension ProBic [34] are fully generative models that combine probabilistic modeling and relational logic. cMonkey [29] is a generative approach which models biclusters by Markov chain processes. Gu and Liu [19] generalized the plaid models proposed in [22] to fully generative models called Bayesian BiClustering model (BBC). The latter models introduced in [6] and [19] are generative models which have the advantage that they select models using well-understood model selection techniques such as maximum likelihood.

Costa et al. [11] introduced a hierarchical model-based co-clustering algorithm. In their method the co-occurrence matrix is characterized in probabilistic terms, by estimating the joint distribution between rows and columns.

The table 2 presents main simultaneous clustering algorithms dealing with continuous, binary or categorical data, the approach they are based on and the domain of application.

4 Conclusion

The survey presented in this work can be used by the interested researcher as a good starting point to learn and apply some of the many techniques proposed in the last few years, and some of the older ones. Many interesting directions for future research have been uncovered by this review work, like the validation of biclustering methods and the statistical significance of biclusters.

References

1. Ahmad, W., Khokhar, A.: cHawk: an efficient biclustering algorithm based on bipartite graph crossing minimization. VLDB. ACM, New York (2007)
2. Balbi, S., Miele, R., Scepi, G.: Clustering of documents from a two-way viewpoint. In: 10th Int. Conf. on Statistical Analysis of Textual Data (2010)
3. Bichot, C.E.: Co-clustering documents and words by minimizing the normalized cut objective function. JMMA 9, 131–147 (2010)
4. Ben-Dor, A., Chor, B., Karp, R.: Discovering local structure in gene expression data: The order–preserving submatrix problem. J. of Comput. Biol. 10, 373–384 (2003)
5. Busygin, S., Jacobsen, G., Kramer, E.: Double conjugated clustering applied to leukemia microarray data. In: 2nd SIAM Int. Conf. on Data Mining (2002)
6. Caldas, J., Kaski, S.: Bayesian biclustering with the plaid model. In: IEEE Intern. Workshop on Machine Learning for Signal Processing, pp. 291–296 (2008)
7. Califano, A., Stolovitzky, G., Tu, Y.: Analysis of gene expression microarays for phenotype classification. In: Int. Conf. on Computational Molecular Biology (2000)
8. Charrad, M., Lechevallier, Y., Ahmed, M.b., Saporta, G.: Block Clustering for Web Pages Categorization. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 260–267. Springer, Heidelberg (2009)
9. Charrad, M.: une approche generique pour l'analyse croisant usage et contenu de sites Web par des methodes de bipartitionnement. PhD Thesis, Paris (2010)
10. Cheng, Y., Church, G.M.: Biclustering of expression data. In: 8th Int. Conf. on Intelligent Systems for Molecular Biology, pp. 93–103 (2000)
11. Costa, G., Manco, G., Ortale, R.: A hierarchical model-based approach to co-clustering high-dimensional data. In: ACM sym. on App. comput., pp. 886–890 (2008)
12. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: 7th ACM SIGKDD 2001, California, pp. 269–274 (2001)
13. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: ACM SIGKDD, pp. 89–98. ACM, Washington DC (2003)
14. Getoor, L., Friedman, N., Koller, D., Taskar, B.: Learning probabilistic models of link structure. J. Mach. Learn. Res. 3, 679–707 (2002)

15. Getz, G., Levine, E., Domany, E.: Coupled two-way clustering analysis of gene microarray data. Proc. of the Natural Academy of Sciences USA (2000)
16. Govaert, G., Nadif, M.: Clustering with block mixture models. J. of the Pattern Recognition, 463–473 (2003)
17. Govaert, G.: Classification croisee. Th. de doctorat d'Etat, Paris (1983)
18. Grimal, C., Bisson, G.: Classification a partir d'une collection de matrices. CAp2010 (2010)
19. Gu, J.: Bayesian biclustering of gene expression data. BMC Genomics (2008).
20. Hartigan, J.A.: Direct clustering of a data matrix. J. of American Statistical Association 67(337), 123–129 (1972)
21. Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A.: FABIA: factor analysis for bicluster acquisition. Bioinformatics journal 26(12), 1520–1527 (2010)
22. Lazzeroni, L., Owen, A.: Plaid models for gene expression data. Technical report, Stanford University (2002)
23. Madeira, S.C., Oliveira, A.L.: Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Trans. on Comp. Biol. and Bioinfor., 24–45 (2004)
24. Madeira, S.C., Teixeira, M.C.: Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. IEEE ACM (2010)
25. Murali, T.M., Kasif, S.: Extracting conserved gene expression motifs from gene expression data. In: Pacific Sym. on Biocomputing, Hawaii, USA, pp. 77–88 (2003)
26. Nadif, M., Govaert, G.: Block clustering of contingency table and mixture model. In: Famili, A.F., Kok, J.N., Peña, J.M., Siebes, A., Feelders, A. (eds.) IDA 2005. LNCS, vol. 3646, pp. 249–259. Springer, Heidelberg (2005)
27. Prelic, A., Bleuler, S., Zimmermann, P.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics, 122–129 (2006)
28. Rege, M., Dong, M.: Co-clustering Documents and Words Using Bipartite Isoperimetric Graph Partitioning. In: 6th IEEE Int. Conf. on Data Mining, pp. 532–541 (2006)
29. Reiss, D.J.: Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinfor., 280–302 (2006)
30. Robardet, C.: Contribution à la classification non supervisee : proposition d'une methode de bi-partitionnement, PhD Thesis, Claude Bernard University (2002).
31. Tanay, A., Sharan, R., Shamir, R.: Biclustering Algorithms: A Survey. In: Aluru, S. (ed.) Handbook of Comp. Molecular Biology, Chapman, Boca Raton (2004)
32. Tang, C., Zhang, L.A.: Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In: IEEE Int. Sym. on Bioinfo. and Bioeng. (2001)
33. Tibshirani, R., Hastie, T., Eisen, M.: Clustering methods for the analysis of DNA microarray data. Technical report, Stanford University (1999)
34. Van den, B.T.: Robust Algorithms for Inferring Regulatory Networks Based on Gene Expression Measurements. PhD Thesis (2009)
35. Wang, H., Wang, W., Yang, J., Yu, P.S.: Clustering by pattern similarity in large data sets. In: ACM SIGMOD Int. Conf. on Management of Data, pp. 394–405 (2002)
36. Yang, J., Wang, H., Wang, W., Yu, P.S.: An improved biclustering method for analyzing gene expression profiles. Int. J. on Art. Int. Tools, 771–790 (2005)
37. Klugar, Y., Basri, R., Chang, J.T., Gerstein, M.: Spectral biclustering of microarray data: coclustering genes and conditions. Genome Research 13, 703–716 (2003)
38. Zha, H., He, X., Ding, C., Simon, H., Gu, M.: Bipartite Graph Partitioning and Data Clustering. In: ACM Conf. on Inf. and Knowledge Management, pp. 25–32 (2001)