2012 IEEE 28th International Conference on Data Engineering

Boolean Matrix Decomposition Problem: Theory, Variations and Applications to Data Engineering

Jaideep Vaidya

Management Science & Information Systems Department, Rutgers University
1 Washington Park, Newark, NJ, 07102, USA
jsvaidya@business.rutgers.edu

***Abstract*—With the ubiquitous nature and sheer scale of data collection, the problem of data summarization is most critical for effective data management. Classical matrix decomposition techniques have often been used for this purpose, and have been the subject of much study. In recent years, several other forms of decomposition, including Boolean Matrix Decomposition have become of significant practical interest. Since much of the data collected is categorical in nature, it can be viewed in terms of a Boolean matrix. Boolean matrix decomposition (BMD), wherein a boolean matrix is expressed as a product of two Boolean matrices, can be used to provide concise and interpretable representations of Boolean data sets. The decomposed matrices give the set of meaningful concepts and their combination which can be used to reconstruct the original data. Such decompositions are useful in a number of application domains including role engineering, text mining as well as knowledge discovery from databases. In this seminar, we look at the theory underlying the BMD problem, study some of its variants and solutions, and examine different practical applications.**

I. Introduction

Today, data collection and storage is proceeding at an unrivaled pace. Indeed, the estimated information added to the digital universe each year should approach 1 Zettabyte (10^{21} bytes) [1]. Given this unimaginable increase in data collection, data summarization is critical. In data mining, matrix decompositions are often employed to produce concise representations of data. In matrix decomposition, an input matrix is represented as a product of two factor matrices, which can be easier to analyze. Matrix decomposition techniques such as LU, SVD[1], LDA[2], and Eigen value decompositions have been the subject of study for quite a long time. However, one major problem with such decomposition is their lack of interpretability. Indeed, while the conciseness of factor matrices is important, for knowledge discovery, their interpretability is also critical. To address this, the concept of non-negative matrix factorization (NMF)[3] has been proposed. In NMF, the decomposed matrices are restricted to be non-negative. While this is definitely more interpretable than the standard decompositions, in many cases even this is not sufficiently interpretable (for example, what do fractional values mean?). Furthermore, standard numerical matrix decomposition does not enable analysis of categorical data. Since much of the real data such as market basket data, word-document data and gene expression data is categorical, or even Boolean in nature, a new matrix decomposition method called Boolean matrix decomposition (BMD) has attracted much attention lately from the data mining and knowledge discovery community.

The goal of BMD is to decompose a Boolean matrix (A) into two Boolean matrices (B and C), where one of the matrices, considered the concept matrix, can be viewed as a set of meaningful concepts (e.g. interesting itemsets, topics, etc.), while the second matrix, called the combination matrix, describes how each observed record (i.e., each row of A), can be expressed as a union of a subset of the concepts. The concepts can be restricted to be a subset of the original records, or unrestricted to simply be any subset of the original items. In either case, the summarization shows both how the original data can be reconstructed in a logical sense, as well as identify the critical underlying concepts which may be overlapping. As such, BMD is a fundamental data analysis technique that can be used in domains as diverse as information security[4], to text mining, to bioinformatics[5]. Many variations of BMD have been recently proposed, with interesting applications – for example, BMD can give the best set of topics summarizing a document, or the best set of roles providing security in role based access control, or even the best set of mod/resc genes required to explain a compatibility dataset for an insect host. BMD is typically a tough problem, with the basic problem being NP-hard, and variations being even hard to approximate (inapproximable unless $P = NP$). In this seminar, we plan to cover the basic theory behind BMD, present different problem variations and extensions, solution approaches formulated from different areas, and finally, cover various applications in diverse domains. We will also cover some of the challenges and avenues for future work in this growing area.

The rest of this paper is structured as follows. In Section II we give a formal overview of the basic BMD problem. Section III identifies some of the variants of the BMD, along with some recently developed solutions. Finally IV revisits some of the application areas and looks at future directions for further research in this area.

II. The Boolean Matrix Decomposition Problem

As discussed above, in Boolean matrix decomposition, a single boolean matrix is decomposed into two boolean matrices which can be recombined appropriately to reconstruct the

[1]Based on the *HMI? How Much Information* report at http://hmi.ucsd.edu/pdf/HMI_2010_EnterpriseReport_Jan_2011.pdf

1084-4627/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDE.2012.144

original matrix. Formally, the boolean matrix $A \in \{0,1\}^{m \times n}$ can be decomposed into the boolean matrices $B \in \{0,1\}^{m \times k}$ and $C \in \{0,1\}^{k \times n}$. This can be represented as $A = B \otimes C$, where $\otimes$ represents the Boolean matrix product such that $a_{ij} = \bigvee_{s=1}^{k}(b_{is} \wedge c_{sj})$. By letting the value of each matrix cell denote the presence/absence of the item corresponding to its column, each row of a matrix can be viewed as a subset of items. For example, a row (1011) represents the itemset $\{1,3,4\}$. For real data, each column corresponds to an attribute, such as a word in textual document data or a particular product in market basket data. The decomposed matrices B and C can be viewed as the combination matrix and the concept matrix respectively. Thus, the concept matrix gives the k "meaningful" concepts (subsets of the items), while the combination matrix specifies which of the concepts are present in each row of the original matrix. Semantically, each column of the concept matrix C could be assigned a meaning based on the actual items contained in it. Viewing the Boolean row as an itemset, the Boolean matrix decomposition can also be represented as $A_i = \cup_{b_{ij}=1} C_j$, where A_i and C_j denote the ith row of A and the jth row of C respectively.

The following presents an illustrative BMD example:

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}. \quad (1)$$

The input Boolean matrix containing four records can be summarized by only three concepts. The combination matrix shows how to reconstruct the input matrix with the three concepts – for example, the first row of the input matrix is the union of the first and third concepts. In the context of text mining, each column of the original matrix would correspond to a word while each row represents a document (with a 1 in cell (i,j) representing the fact that document i contains word j. Now, the BMD describes the four documents using three topics corresponding to each row of the concept matrix respectively. If the input matrix represents movie feedback, with rows corresponding to users, columns corresponding to movies, and 1 denoting "like", the columns of the concept matrix can be interpreted as movie types and thus a person's preference can be described as a combination of movie types. Figure 1 gives a graphical illustration of the same example.

It should be clear that there are many possible decompositions that can give the original matrix. A trivial example is where $k = m$, B is the identity matrix, and $C = A$ (thus $A = I \otimes A$) or where $k = n$, C is the identity matrix, and $B = A$ (thus $A = A \otimes I$). Neither of these is of any use. However alternative decompositions (such as the one presented in Equation 1 could indeed be of interest. Therefore, to achieve a meaningful BMD, it is necessary to find a decomposition that meets a certain objective. For example, we may choose to find the decomposition that minimizes the number of concepts (topics in the above example). Now, the BMD problem is the following optimization problem: for the given boolean matrix $A \in \{0,1\}^{m \times n}$, find boolean matrices $B \in \{0,1\}^{m \times k}$ and $C \in \{0,1\}^{k \times n}$, such that $A = B \otimes C$ and k is minimal. As shown by Vaidya et al.[4], this problem actually corresponds to the Minimum Tiling Problem originally studied by Geerts et al. and is proven to be NP-hard[6].

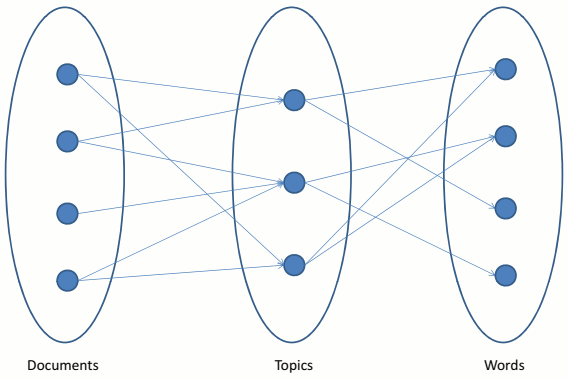

Fig. 1. Boolean matrix decomposition depicted graphically

III. Problem Variants and Solution Approaches

Along with the basic BMD problem, many practical variants of the problem also abound. Miettinen et al.[7] first studied the relaxation of this problem, wherein the best approximate decomposition was desired when the number of concepts was fixed (i.e., find B, C such that $\| A - B \otimes C \|_1$ is minimum, where $\| \cdot \|_1$ represents the L_1-norm (effectively the number of unequal elements for boolean matrices). Instead of finding both the set of concepts and their combination, we may want to only find the combination given the concepts. This can be done in polynomial time if an exact solution is desired, but is only fixed parameter tractable if approximation is possible[7]. Vaidya et al.[8], [4] study the related problem of minimizing the number of concepts while restricting the approximation to be of a certain quality (i.e., limiting the number of errors within a threshold, and perhaps of a certain type). This has great application in the area of role mining and role engineering[9].

Variants of this problem can be of several forms. First, it is possible to change the objective function. For example, instead of minimizing the number of concepts, it may make more sense to minimize the number of assignments or the sum of the number of assignments and the number of concepts[10], [11], [12]. It is also possible to restrict the solution space. For example, the boolean Column and Column-Row Matrix Decompositions [13] restrict the concepts to be from among the original records. This may make more sense in summarization through representatives. Sparse BMD [14] studies the case where the input data is sparse, and the factor matrices are also required to be sparse. Alternatively, the basic model itself can be extended, since basic BMD only incompletely represents real data semantics. For example, the extended BMD problem [15] looks at the case where one of the factor matrices may include '-1's (i.e., each element is from the ternary set

$\{1, 0, -1\}$. The -1 now corresponds to the set difference semantics, thus allowing exclusion of either items directly or through concepts. While a BMD solution only allows an input record to be described as a union of a subset of concepts, an EBMD solution allows the input record to be described as an inclusion of some concepts with exclusion of other concepts. For example, consider a long presidential speech that covers all topics except "Foreign Policy". With only the set union operation, to describe that speech, we have to list all topics that appear in it. If the set difference operation can be utilized, we can create a topic called "ALL-TOPICS" and represent the presidential speech by "ALL-TOPICS $\setminus$ Foreign Policy". Thus, this can enable more succint representation of the original records, as well a smaller overall number of concepts required to reconstruct the records.

Other variants include the minimal perturbation problem where a set of concepts are given along with the original input records, and the best decomposition is required, where best includes one of the standard objectives as well as the requirement of being similar (in some fashion) to the given set of concepts. Rank-one decompositions [16] have also been recently studied to enable quick and dirty approximations of the original data.

Since most of the problems in this area are NP-hard, and in certain cases, even hard to approximate, many heuristic solutions have been proposed. Many of these follow the greedy approach [17], [4], [18], [19], [20], [21], [6], though the specific way of creating candidate concepts may change. There has also recently been work on modeling this problem as an Integer Programming Model[22] and solving it using branch and bound solvers, or through SAT solvers. Probabilistic modeling of BMD[23] is another approach that has been taken to solve the problem as well. [14] comes up with novel computational techniques to solve the sparse BMD problem.

IV. Application Areas and Future Directions

As discussed earlier, boolean matrix decomposition has numerous applications in diverse areas, from text mining, access control, to data engineering in bioinformatics. Different problem domains have different constraints and can lead to meaningful variant of the problem being studied. Solution approaches can also be promulgated using the expertise from the various domains. In the future, we expect this to be a multidisciplinary field, where core problems can be identified, and practical solutions can have lasting impact due to the heterogeneity of applications. More work is necessary, both for identifying new problems and constraints, and also creating efficient solutions for specific sub-problems.

References

[1] G. Golub and C. Loan, *Matrix computations*, ser. Johns Hopkins studies in the mathematical sciences. Johns Hopkins University Press, 1996. [Online]. Available: http://books.google.com/books?id=mlOa7wPX6OYC

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, March 2003. [Online]. Available: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999. [Online]. Available: http://dx.doi.org/10.1038/44565

[4] J. Vaidya, V. Atluri, and Q. Guo, "The role mining problem: A formal perspective," *ACM Trans. Inf. Syst. Secur.*, vol. 13, no. 3, pp. 1–31, 2010.

[5] I. Nor, D. Hermelin, S. Charlat, J. Engelstadter, M. Reuter, O. Duron, and M.-F. Sagot, "Mod/resc parsimony inference," in *Proceedings of the 21st annual conference on Combinatorial pattern matching*, ser. CPM'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 202–213.

[6] F. Geerts, B. Goethals, and T. Mielikainen, "Tiling databases," in *Discovery Science*, ser. Lecture Notes in Computer Science. Springer-Verlag, 2004, pp. 278 – 289. [Online]. Available: http://www.springerlink.com/content/31ahky75yecgtwlu

[7] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila, "The discrete basis problem," *IEEE Trans. on Knowl. and Data Eng.*, vol. 20, pp. 1348–1362, October 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1442800.1442809

[8] J. Vaidya, V. Atluri, and Q. Guo, "The role mining problem: Finding a minimal descriptive set of roles," in *The Twelth ACM Symposium on Access Control Models and Technologies*, Sophia Antipolis, France, Jun.20-22 2007, pp. 175–184.

[9] E.J.Coyne, "Role-engineering," in *1st ACM Workshop on Role-Based Access Control*, 1995.

[10] J. Vaidya, V. Atluri, Q. Guo, and H. Lu, "Edge-rmp: Minimizing administrative assignments for role-based access control," *Journal of Computer Security*, vol. 17, pp. 211–235, 2009.

[11] A. Ene, W. Horne, N. Milosavljevic, P. Rao, R. Schreiber, and R. Tarjan, "Fast exact and heuristic methods for role minimization problems," in *The ACM Symposium on Access Control Models and Technologies*, June 2008.

[12] D. Zhang, K. Ramamohanarao, and T. Ebringer, "Role engineering using graph optimisation," in *SACMAT '07: Proceedings of the 12th ACM symposium on Access control models and technologies*. New York, NY, USA: ACM, 2007, pp. 139–144.

[13] P. Miettinen, "The boolean column and column-row matrix decompositions," *Data Min. Knowl. Discov.*, vol. 17, pp. 39–56, August 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1401694.1401756

[14] ——, "Sparse boolean matrix factorizations," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ser. ICDM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 935–940.

[15] H. Lu, J. Vaidya, V. Atluri, and Y. Hong, "Extended boolean matrix decomposition," in *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ser. ICDM '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 317–326.

[16] H. Lu, J. Vaidya, V. Atluri, H. Shin, and L. Jiang, "Weighted rank-one binary matrix factorization," in *SDM*. SIAM / Omnipress, 2011, pp. 283–294.

[17] V. Snásel, J. Platos, and P. Krömer, "On genetic algorithms for boolean matrix factorization," in *Proceedings of the 2008 Eighth International Conference on Intelligent Systems Design and Applications - Volume 02*, ser. ISDA '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 170–175.

[18] I. Molloy, H. Chen, T. Li, Q. Wang, N. Li, E. Bertino, S. Calo, and J. Lobo, "Mining roles with semantic meanings," in *SACMAT '08: Proceedings of the 13th ACM symposium on Access control models and technologies*. New York, NY, USA: ACM, 2008, pp. 21–30.

[19] A. Colantonio, R. D. Pietro, and A. Ocello, "Leveraging lattices to improve role mining," in *Proceedings of The IFIP TC-11 23rd International Information Security Conference (IFIP SEC '08)*, 2008, pp. 333–347.

[20] A. Colantonio, R. Di Pietro, and A. Ocello, "A cost-driven approach to role engineering," in *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2008, pp. 2129–2136.

[21] Q. Guo, J. Vaidya, and V. Atluri, "The role hierarchy mining problem: Discovery of optimal role hierarchies," in *Proceedings of the 24th Annual Computer Security Applications Conference*, Dec.8-12 2008, pp. 237 – 246.

[22] H. Lu, J. Vaidya, and V. Atluri, "Optimal boolean matrix decomposition: Application to role engineering," in *Proceedings of the 24th International Conference on Data Engineering (ICDE)*, 2008, pp. 297–306.

[23] M. Frank, D. Basin, and J. M. Buhmann, "A class of probabilistic models for role engineering," in *CCS '08: Proceedings of the 15th ACM conference on Computer and communications security*. New York, NY, USA: ACM, 2008, pp. 299–310.