The Boolean Column and Column-Row Matrix Decompositions*

Pauli Miettinen

Helsinki Institute for Information Technology
University of Helsinki
pauli.miettinen@cs.helsinki.fi

Matrix decompositions are used for many data mining purposes. One of these purposes is to find a concise but interpretable representation of a given data matrix. Different decomposition formulations have been proposed for this task, many of which assume a certain property of the input data (e.g., nonnegativity) and aim at preserving that property in the decomposition.

In this paper we propose two new decomposition formulations for binary matrices, namely the Boolean CX and CUR decompositions. They are natural combinations of two previously-presented decomposition formulations. The Boolean CX (BCX) decomposition assumes a binary input matrix A and decomposes it into two binary factor matrices, C and X, with matrix C containing a subset of A's columns. Matrix A is represented using the Boolean matrix product of C and X, $A \approx C \circ X$. The Boolean matrix product $\circ$ is like the normal matrix product, but with addition defined as $1 + 1 = 1$. In the Boolean CUR (BCUR) decomposition, A is decomposed into three matrices, C, U, and R, with matrix C containing a subset of A's columns and matrix R containing a subset of A's rows. Matrix A is represented as $A \approx C \circ U \circ R$.

We also study two subproblems of these decompositions, the Basis Usage (BU) problem and the Mixing Matrix (MM) problem. In the former we are given the matrices A and C and our goal is to find the matrix X such that $A \approx C \circ X$. In the latter we are given the matrices A, C, and R and our goal is to find the matrix U such that $A \approx C \circ U \circ R$. We give lower and upper bounds for the approximability of the BU and MM problems and use the results to show the NP-completeness of the BCX problem.

We give algorithms for the problems and study the performance of the algorithms via extensive experimental evaluation. Our results show that, despite the high theoretical complexity of the problems, even simple algorithms can perform well with both synthetic and real-world data.

References

1. Miettinen, P.: The Boolean column and column-row matrix decompositions. Data Mining and Knowledge Discovery 17(1), 39–56 (August 2008)

* This is an extended abstract of an article published in the Data Mining and Knowledge Discovery journal [1].

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part I, LNAI 5211, p. 17, 2008.