SPARSE BOOLEAN MATRIX FACTORIZATIONS

Pauli Miettinen
15.12.2010
BOOLEAN FACTORIZATIONS

• Input: a 0/1 (i.e. Boolean) \( n \)-by-\( m \) matrix \( \mathbf{A} \) and integer \( k \) (i.e. the rank)

• Output: 0/1 \( n \)-by-\( k \) matrix \( \mathbf{B} \) and 0/1 \( k \)-by-\( m \) matrix \( \mathbf{C} \)

• Goal: minimize \( \sum_{ij} |A_{ij} - (B \circ C)_{ij}| \)

• Boolean matrix multiplication: \((B \circ C)_{ij} = \lor_p B_{ip} C_{pj}\)

• Like normal, but addition defined as \( 1 + 1 = 1 \)
SOME EXITING PROPERTIES

• Easy to interpret

• Generalizes many data mining techniques

• Boolean rank can be exponentially smaller than normal rank
  • Boolean factorizations can have less error than SVD

• Computations become combinatorial
SOME BAD NEWS

- Computations become combinatorial
- Finding optimal Boolean factorizations is computationally hard
- Hard inapproximability results for:
  - best Boolean rank-\(k\) factorization of a given matrix
  - Boolean rank of a given matrix
    - As hard as finding graph’s minimum chromatic number
GOOD NEWS

• Sparsity helps!
SPARSE FACTORIZATIONS

• Ideally, sparse matrices have sparse factors

• Not true with many factorization methods

• Sparse Boolean matrices have sparse decompositions
Theorem 1. For any \( n \times m \) 0/1 matrix \( A \) of Boolean rank \( k \), there exist \( n \times k \) and \( k \times m \) 0/1 matrices \( B \) and \( C \) such that \( A = B \circ C \) and \( |B| + |C| \leq 2|A| \).

- Ideally, sparse matrices have sparse factors
- Not true with many factorization methods
- Sparse Boolean matrices have sparse decompositions
APPROXIMATING THE BOOLEAN RANK

• Sparsity is not enough; we need some structure in it

• An $n$-by-$m$ 0/1 matrix $A$ is $f(n)$-uniformly sparse, if all of its columns have at most $f(n)$ 1’s

**Theorem 2.** The Boolean rank of $\log(n)$-uniformly sparse matrix can be approximated to within $O(\log(m))$ in time $\tilde{O}(m^2n)$. 
NON-UNIFORMLY SPARSE MATRICES

• Uniform sparsity is very restricted; what can we do

• Trade non-uniformity with approximation accuracy
NON-UNIFORMLY SPARSE MATRICES

- Uniform sparsity is very restricted; what can we do
- Trade non-uniformity with approximation accuracy

**Theorem 3.** If there are at most $\log(m)$ columns with more than $\log(n)$ 1s, then we can approximate the Boolean rank in polynomial time to within $O(\log^2(m))$. 
**Theorem 4.** If n-by-m 0/1 matrix A is $O(\log n)$-uniformly sparse, we can approximate the best dominated $k$-cover of A by $\frac{e}{e-1}$ in polynomial time.

- Dominated $k$-cover: The rank is $k$ and if $(B \circ C)_{ij} = 1$, then $A_{ij} = 1$
- Has applications e.g. in role mining
APPROXIMATING THE RANK
SPARSITY
APPROXIMATION ERROR
CONCLUSIONS

• Sparse Boolean matrices have sparse decompositions
  • Not true with “normal” decompositions
• Sparsity helps with computational complexity
  • Requires some regularity in sparsity
• Initial work; better results to be expected.
CONCLUSIONS

• Sparse Boolean matrices have sparse decompositions
  • Not true with “normal” decompositions
• Sparsity helps with computational complexity
  • Requires some regularity in sparsity
• Initial work; better results to be expected.

Thank You!