COLUMN SUBSET SELECTION FOR APPROXIMATING DATA MATRICES

By

Ali Çivril

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Malik Magdon-Ismail, Thesis Adviser

Petros Drineas, Member

Mark Goldberg, Member

John E. Mitchell, Member

Rensselaer Polytechnic Institute
Troy, New York

December 2009
(For Graduation December 2009)

ABSTRACT

In this thesis, we study the problem of selecting a subset of columns of a matrix so that they capture the important information contained in the matrix. We present complexity results and algorithms. The problem, in a very broad sense, asks for a "good" subset of columns of a given real matrix that provides a performance guarantee in terms of an objective function related to the spectrum of the matrix.

We first present a linear-time spectral graph drawing algorithm as a motivation which is a vast improvement over the standard quadratic-time method Classical Multidimensional Scaling (CMDS). To guarantee a fast implementation of the algorithm, it is desirable to quickly select a subset of columns of a distance matrix associated with the graph. Intuitively, in order to obtain a well conditioned submatrix, one has to choose a subset of column vectors -in a geometrical sense- such that they are as "far away" from each other as possible. We consider formalizations of this notion by studying the problem of selecting a subset of columns of size k such that it satisfies some certain orthogonality conditions. We establish the NP-hardness of a few such problems and further show that two of them do not admit PTAS. For the problem of choosing the maximum volume sub-matrix, which we call MAX-VOL, we analyze a greedy algorithm and show that it provides a $2^{-O(k \log k)}$ approximation. Our analysis of the greedy heuristic is tight to within a logarithmic factor in the exponent, which we show by explicitly constructing an instance for which the greedy heuristic is $2^{-\Omega(k)}$ from optimal. Further, we show that no efficient algorithm can appreciably improve upon the greedy algorithm by proving that MAX-VOL is NP-hard to approximate within 2^{-ck} for some constant c. Our proof is via a reduction from the Label-Cover problem.

Our last result is a constructive solution to the low-rank matrix approximation problem which asks for a subset of columns of a matrix that captures "most" of its spectrum. Our main result is a simple greedy deterministic algorithm with guarantees on the performance while choosing a small number of columns. Specifically,

our greedy algorithm chooses c columns from A with $c = \tilde{O}\left(\frac{k^2 \log k}{\epsilon^2}\mu^2(A)\right)$ such that

$$\|A - CC^+A\|_F \leq (1+\epsilon)\|A - A_k\|_F,$$

where C is the matrix composed of the c columns, C^+ is the pseudo-inverse of C (CC^+A is the best reconstruction of A from C), and $\mu(A)$ is a measure of the *coherence* in the normalized columns of A. To the best of our knowledge, this is the first deterministic algorithm with performance guarantees on the number of columns and a $(1+\epsilon)$ approximation ratio in Frobenius norm. Numerical results suggest that the performance of the algorithm might be far better than the theoretical bounds suggest.