The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Solving Generalized Column Subset Selection with Heuristic Search

Swair Shah, Baokun He, Ke Xu, Crystal Maung, Haim Schweitzer
{swair,baokun.he,ke.xu5,ktm016100,hschweitzer}@utdallas.edu

Abstract

We address the problem of approximating a matrix by the linear combination of a column sparse matrix and a low rank matrix. Two variants of a heuristic search algorithm are described. The first produces an optimal solution but may be slow, as these problems are believed to be NP-hard. The second is much faster, but only guarantees a suboptimal solution. The quality of the approximation and the optimality criterion can be specified in terms of unitarily invariant norms.

1 Introduction

Approximating a matrix by a linear combination of a small number of vectors is an important problem encountered in many applications of numerical linear algebra. Let X and V be two matrices. The approximation of X in the column subspace of V can be written as:

$$X \approx VA, \quad \text{approximation error} = \Theta(X - VA) \tag{1}$$

Here A is the coefficients matrix and Θ is an error criterion. When the columns of V are restricted to be the columns of X, the problem is the well known Column Subset Selection Problem (CSSP); when all the columns of V are unrestricted, the problem is the classic Principal Component Analysis (PCA).

The generalization of these two problems is the case where some of the columns of the V are required to be columns of X, and the rest are unrestricted. In this case the approximation (1) can be written as:

$$X \approx SA_1 + VA_2 \tag{2}$$

Here S consists of columns from X, and V is unrestricted. We refer to this representation as the "double low-rank representation" (**DLRR**). The algorithms described in this paper are for the DLRR, and thus can be applied to the CSSP.

The approximation errors of the three representations considered here are shown in (3). The notation $|S|$ is used for the number of columns of the matrix S, and the notation $S \subset X$ indicates that the columns of S are a also columns of X. We propose heuristic search algorithms for the DLRR that are optimal for all unitarily invariant norms (*e.g.* Spectral, Nuclear, Frobenius).

$$\boxed{\begin{aligned}
E_{\text{PCA}}(X,r) &= \min_{V,A} \Theta(X - VA) \\
&\qquad \text{subject to } |V| = r \\
E_{\text{CSSP}}(X,r) &= \min_{S,A} \Theta(X - SA) \\
&\qquad \text{subject to } S \subset X,\, |S| = r \\
E_{\text{DLRR}}(X,r_1,r_2) &= \min_{S,A_1,V,A_2} \Theta(X - SA_1 - VA_2) \\
&\qquad \text{subject to } S \subset X,\, |S| = r_1,\, |V| = r_2
\end{aligned}} \tag{3}$$

2 The algorithm

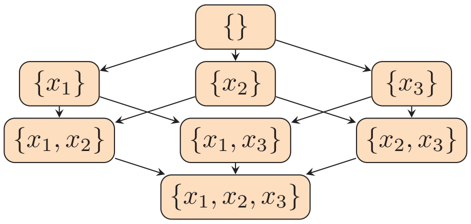
Our method conducts a heuristic search on the column subsets graph that was originally described in (Arai, Maung, and Schweitzer 2015). The graph nodes correspond to column subsets, and there is an edge from subset S_i to subset S_j if adding one column to S_i creates S_j. The graph generated for the matrix $X = (x_1, x_2, x_3)$ is depicted in the top part of Fig 1.

The heuristic search algorithm is shown at the bottom of Fig 1. The algorithm keeps a fringe list L of nodes that need to be examined, and a list C of closed nodes, containing nodes that will not be visited again. The nodes are selected from the fringe according to the value of the heuristic function f'. The algorithm terminates at the first time that a node n_i is selected which has a column subset of size r_1.

This algorithm is essentially the same as the classic A^* algorithm (Pearl 1984). However, the standard heuristic functions d, f, g, h that are used by the classic A^* do not have a trivial equivalent in our case. We proceed to define heuristic functions with similar notation to the ones used in the classic A^* algorithm.

2.1 Heuristic functions

The DLRR is defined in terms of X, r_1, r_2. At each node n_i the subset S_i and its size k_i are known. Define e^* to be the smallest error of approximating X by a selection of r_1 columns and the best possible additional r_2 unrestricted vectors. The heuristic functions are defined at each node n_i. The function d is defined as the smallest error of approximating X by a selection of r_1 columns that include S_i and the best possible additional r_2 unrestricted vectors. The function g is defined as the smallest error of approximating X by the



```
   Input: $X$, $r_1$, $r_2$, and a heuristic function $f'(n)$.
   Each node $n_i$ has a subset $S_i$ of size $k_i$.
   Initialization: Put an empty subset into $L$.
1  while $L$ is nonempty do
2      Pick $n_i$ with the smallest $f'(n_i)$ from $L$.
3      if $k_i = r_1$ then
4          Stop and return $n_i$ as the solution node.
5      else
6          Add $n_i$ to $C$.
7          for all children $n_j$ of $n_i$ do
8              if $n_j$ is not in $C$ or $L$ then
9                  put $n_j$ in $L$.
10             end
11         end
12     end
13 end
14 Here $L$ is empty. Solution was not found.
```

Figure 1: The subset graph and the search algorithm

selection S_i and the best possible additional r_2 unrestricted vectors. The function f is defined as the smallest error of approximating X by the selection S_i and the best possible additional r_1+r_2− k_i unrestricted vectors.

$$\begin{aligned}
e^*(X, r_1, r_2) &= \min_{S, A_1, V, A_2} \Theta(X - SA_1 - VA_2) \\
&\quad \text{subject to } \; S \subset X, |S| = r_1, |V| = r_2 \\
d(n_i, r_1, r_2) &= \min_{S, A_1, V, A_2} \Theta(X - SA_1 - VA_2) \\
&\quad \text{subject to } \; S_i \subset S \subset X, |S| = r_1, |V| = r_2 \\
g(n_i, r_2) &= \min_{A_1, V, A_2} \Theta(X - S_iA_1 - VA_2) \\
&\quad \text{subject to } \; |V| = r_2 \\
f(n_i, r_1, r_2) &= \min_{A_1, V, A_2} \Theta(X - S_iA_1 - VA_2) \\
&\quad \text{subject to } \; |V| = r_1 + r_2 - k_i
\end{aligned} \tag{4}$$

We could not find an analog for the heuristic function h from the classic A^*. Clearly, the best heuristic choice for the algorithm is f'=d. But since d cannot be calculated efficiently we consider other choices using f and g. Observe that both f and g can be viewed as approximations of d. The important theoretical characterizations of our results are stated below.

Proposition: For each node n_i:

$$f(n_i, r_1, r_2) \leq d(n_i, r_1, r_2) \leq g(n_i, r_2)$$

and if $k_i = r_1$ then the inequalities become equalities.

The heuristic $f' = g$ leads to a greedy algorithm. To see this observe that g is monotonically decreasing along any path of the graph. Therefore, when a node n_i is picked in the Line 2 of the algorithm its children will be examined before all other nodes currently in the fringe.

The heuristic $f' = f$ gives an algorithm similar to the classic A^*. In particular it can be proved that this choice guarantees that the algorithm finds an optimal solution.

The heuristic $f' = f + \epsilon g$ gives an algorithm similar to the classic Weighted A^* algorithm (Pearl 1984; Arai et al. 2016). For $\epsilon > 0$ the algorithm is much faster than $f' = f$, and gives better accuracy than the case where $f' = g$. The sub-optimality guarantee is stated in the theorem below:

Theorem: Let n_* be an optimal solution node for the DLRR. Define:

$$f'(n_i, r_1, r_2) = f(n_i, r_1, r_2) + \epsilon g(n_i, r_2), \quad \epsilon \geq 0$$

Then the algorithm in Fig 1 will terminate with a sub-optimal solution node n_{**} with the corresponding values $S^{**}, V^{**}, A_1^{**}, A_2^{**}$, satisfying:

$$\begin{aligned}
&\Theta(X - S^{**}A_1^{**} - V^{**}A_2^{**}) \\
&\qquad \leq e^*(X, r_1, r_2) + \epsilon E_{\text{PCA}}(X, r_2)
\end{aligned} \tag{5}$$

where E_{PCA} is defined in Eq. 3. The proofs can be found in the full version of this paper.

The results above hold for arbitrary error function Θ. Though Θ should be selected to allow efficient calculations of the heuristic f'. The full paper shows that this can be achieved for all Unitarily Invariant norms. These include Frobenius, Spectral, Nuclear etc.

3 Experimental Results

We implemented and tested our algorithms on various datasets from the UCI Machine Learning repository. The results for the CSSP were compared to those obtained by (Nie, Huang, and Ding 2012) and (Gu and Eisenstat 1996) (for spectral norm). The accuracy of our algorithms compared favorably with those algorithms. The DLRR can also be computed by first applying CSSP to select r_1 columns and then PCA for the remaining r_2 vectors. Our experiments confirmed the hypothesis that solving DLRR with our method (which is optimal) gives significantly better accuracy than the optimal CSSP followed by PCA. These experimental results will be shown in the full version of the paper.

References

Arai, H.; Xu, K.; Maung, C.; and Schweitzer, H. 2016. Weighted A* algorithms for unsupervised feature selection with provable bounds on suboptimality. In *AAAI'16*.

Arai, H.; Maung, C.; and Schweitzer, H. 2015. Optimal column subset selection by A-star search. In *AAAI'15*.

Gu, M., and Eisenstat, S. C. 1996. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing* 17(4):848–869.

Nie, F.; Huang, H.; and Ding, C. H. 2012. Low-rank matrix recovery via efficient schatten p-norm minimization. In *AAAI'12*.

Pearl, J. 1984. Heuristics: intelligent search strategies for computer problem solving.