

Greedy column subset selection for large-scale data sets

Ahmed K. Farahat · Ahmed Elgohary · Ali Ghodsi ·
Mohamed S. Kamel

Received: 10 December 2013 / Revised: 20 September 2014 / Accepted: 29 October 2014 /
Published online: 27 December 2014
© Springer-Verlag London 2014

Abstract In today's information systems, the availability of massive amounts of data necessitates the development of fast and accurate algorithms to summarize these data and represent them in a succinct format. One crucial problem in big data analytics is the selection of representative instances from large and massively distributed data, which is formally known as the Column Subset Selection problem. The solution to this problem enables data analysts to understand the insights of the data and explore its hidden structure. The selected instances can also be used for data preprocessing tasks such as learning a low-dimensional embedding of the data points or computing a low-rank approximation of the corresponding matrix. This paper presents a fast and accurate greedy algorithm for large-scale column subset selection. The algorithm minimizes an objective function, which measures the reconstruction error of the data matrix based on the subset of selected columns. The paper first presents a centralized greedy algorithm for column subset selection, which depends on a novel recursive formula for calculating the reconstruction error of the data matrix. The paper then presents a MapReduce algorithm, which selects a few representative columns from a matrix whose columns are massively distributed across several commodity machines. The algorithm first

A preliminary version of this paper appeared as [26].
This work was completed while the second author was at the University of Waterloo.

A. K. Farahat (✉) · M. S. Kamel
Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1,
Canada
e-mail: afarahat@uwaterloo.ca

M. S. Kamel
e-mail: mkamel@uwaterloo.ca

A. Elgohary
Department of Computer Science, University of Maryland, College Park, MD 20742, USA
e-mail: elgohary@cs.umd.edu

A. Ghodsi
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
e-mail: aghodsib@uwaterloo.ca

learns a concise representation of all columns using random projection, and it then solves a generalized column subset selection problem at each machine in which a subset of columns are selected from the sub-matrix on that machine such that the reconstruction error of the concise representation is minimized. The paper demonstrates the effectiveness and efficiency of the proposed algorithm through an empirical evaluation on benchmark data sets.

Keywords Column subset selection · Greedy algorithms · Distributed computing · Big data · MapReduce

1 Introduction

Recent years have witnessed the rise of the big data era in computing and storage systems. With the great advances in information and communication technology, hundreds of petabytes of data are generated, transferred, processed and stored every day. The availability of this overwhelming amount of structured and unstructured data creates an acute need to develop fast and accurate algorithms to discover useful information that is hidden in the big data. One of the crucial problems in the big data era is the ability to represent the data and its underlying information in a succinct format.

Although different algorithms for clustering and dimension reduction can be used to summarize big data, these algorithms tend to learn representatives, whose meanings are difficult to interpret. For instance, the traditional clustering algorithms such as k -means [37] tend to produce centroids, which encode information about thousands of data instances. The meanings of these centroids are hard to interpret. Even clustering methods that use data instances as prototypes, such as k -medoid [42], learn only one representative for each cluster, which is usually not enough to capture the insights of the data instances in that cluster. In addition, using medoids as representatives implicitly assumes that the data points are distributed as clusters and that the number of those clusters is known ahead of time. This assumption is not true for many data sets. On the other hand, traditional dimension reduction algorithms such as latent semantic analysis (LSA) [16] tend to learn a few latent concepts in the feature space. Each of these concepts is represented by a dense vector, which combines thousands of features with positive and negative weights. This makes it difficult for the data analyst to understand the meaning of these concepts. Even if the goal of representative selection is to learn a low-dimensional embedding of data instances, learning dimensions whose meanings are easy to interpret allows the understanding of the results of the data mining algorithms, such as understanding the meanings of data clusters in the low-dimensional space.

The acute need to summarize big data to a format that appeals to data analysts motivates the development of different algorithms to directly select a few representative data instances and/or features. This problem can be generally formulated as the selection of a subset of columns from a data matrix, which is formally known as the column subset selection (CSS) problem [3, 6, 7, 22, 29]. Although many algorithms have been proposed for tackling the CSS problem, most of these algorithms focus on randomly selecting a subset of columns with the goal of using these columns to obtain a low-rank approximation of the data matrix. In this case, these algorithms tend to select a relatively large number of columns. When the goal is to select a very few columns to be directly presented to a data analyst or indirectly used to interpret the results of other algorithms, the randomized CSS methods are not going to produce a meaningful subset of columns. On the other hand, deterministic algorithms for

CSS, although more accurate, do not scale to work on big matrices with massively distributed columns.

This paper addresses the aforementioned problem by first presenting a fast and accurate greedy algorithm for column subset selection. The algorithm minimizes an objective function, which measures the reconstruction error of the data matrix based on the subset of selected columns. The paper presents a novel recursive formula for calculating the reconstruction error of the data matrix and then proposes a fast and accurate algorithm, which selects the most representative columns in a greedy manner. The paper then presents a distributed column subset selection algorithm for selecting a very few columns from a big data matrix with massively distributed columns. The algorithm starts by learning a concise representation of the data matrix using random projection. Each machine then independently solves a generalized column subset selection problem in which a subset of columns is selected from the current sub-matrix such that the reconstruction error of the concise representation is minimized. A further selection step is then applied to the columns selected at different machines to select the required number of columns. The proposed algorithm is designed to be executed efficiently over massive amounts of data stored on a cluster of several commodity nodes. In such settings of infrastructure, ensuring the scalability and the fault tolerance of data processing jobs is not a trivial task. In order to alleviate these problems, MapReduce [15] was introduced to simplify large-scale data analytics over a distributed environment of commodity machines. Currently, MapReduce (and its open source implementation Hadoop [56]) is considered the most successful and widely used framework for managing big data processing jobs. The approach proposed in this paper considers the different aspects of developing MapReduce-efficient algorithms.

The contributions of the paper can be summarized as follows:

- The paper first presents a fast and accurate algorithm for column subset selection (CSS), which selects the most representative columns from a data matrix in a greedy manner. The algorithm minimizes an objective function, which measures the reconstruction error of the data matrix based on the subset of selected columns.
- The paper presents a novel recursive formula for calculating the reconstruction error of the data matrix based on the subset of selected columns and then uses this formula to develop a fast and accurate algorithm for greedy CSS.
- The paper proposes an algorithm for distributed CSS, which first learns a concise representation of the data matrix and then selects columns from distributed sub-matrices that approximate this concise representation.
- To facilitate CSS from different sub-matrices, a fast and accurate algorithm for generalized CSS is proposed. This algorithm greedily selects a subset of columns from a source matrix, which approximates the columns of a target matrix.
- A MapReduce-efficient algorithm is proposed for learning a concise representation using random projection. The paper also presents a MapReduce algorithm for distributed CSS, which only requires two passes over the data with a very low communication overhead.
- Medium- and large-scale experiments have been conducted on benchmark data sets in which different methods for CSS are compared.

The rest of the paper is organized as follows. Section 2 describes the notations used throughout the paper. Section 3 gives a brief background on the CSS problem and the MapReduce framework. Section 4 describes the centralized greedy algorithm for CSS. The proposed MapReduce algorithm for distributed CSS is described in detail in Sect. 5. Section 6 reviews the state-of-the-art CSS methods and their applicability to distributed data. In Sect. 7, an

empirical evaluation of the proposed method is described. Finally, Sect. 8 concludes the paper.

2 Notations

The following notations are used throughout the paper unless otherwise indicated. Scalars are denoted by small letters (e.g., m, n); sets are denoted in script letters (e.g., \mathcal{S}, \mathcal{R}); vectors are denoted by small bold italic letters (e.g., \mathbf{f}, \mathbf{g}); and matrices are denoted by capital letters (e.g., A, B). The subscript (i) indicates that the variable corresponds to the i -th block of data in the distributed environment. In addition, the following notations are used:

For a set \mathcal{S} :

$|\mathcal{S}|$ the cardinality of the set.

For a vector $\mathbf{x} \in \mathbb{R}^m$:

\mathbf{x}_i i -th element of \mathbf{x} .

$\|\mathbf{x}\|$ the Euclidean norm (ℓ_2 -norm) of \mathbf{x} .

For a matrix $A \in \mathbb{R}^{m \times n}$:

A_{ij} (i, j)-th entry of A .

A_i : i -th row of A .

$A_{:j}$ j -th column of A .

$A_{:\mathcal{S}}$ the sub-matrix of A which consists of the set \mathcal{S} of columns.

A^T the transpose of A .

$\|A\|_F$ the Frobenius norm of A : $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$.

\tilde{A} a low-rank approximation of A .

$\tilde{A}_{\mathcal{S}}$ a rank- l approximation of A based on the set \mathcal{S} of columns, where $|\mathcal{S}| = l$.

3 Background

This section reviews the necessary background on the column subset selection (CSS) problem and the MapReduce paradigm used to develop the large-scale CSS algorithm presented in this paper.

3.1 Column subset selection (CSS)

The column subset selection (CSS) problem can be generally defined as the selection of the most representative columns of a data matrix [3,6,7]. The CSS problem generalizes

the problem of selecting representative data instances, as well as the unsupervised feature selection problem. Both are crucial tasks that can be directly used for data analysis or as preprocessing steps for developing fast and accurate algorithms in data mining and machine learning.

Although different criteria for column subset selection can be defined, a common criterion that has been used in much recent work measures the discrepancy between the original matrix and the approximate matrix reconstructed from the subset of selected columns [3, 6, 7, 10, 18, 20–22, 29]. Most of the recent work either develops CSS algorithms that directly optimize this criterion or uses this criterion to assess the quality of the proposed CSS algorithms. In the present work, the CSS problem is formally defined as

Problem 1 (*Column Subset Selection*) Given an $m \times n$ matrix A and an integer l , find a subset of columns \mathcal{L} such that $|\mathcal{L}| = l$ and

$$\mathcal{L} = \arg \min_S \|A - P^{(S)} A\|_F^2,$$

where $P^{(S)}$ is an $m \times m$ projection matrix which projects the columns of A onto the span of the candidate columns $A_{:S}$.

The criterion $\mathbf{F}(S) = \|A - P^{(S)} A\|_F^2$ represents the sum of squared errors between the original data matrix A and its rank- l column-based approximation (where $l = |S|$),

$$\tilde{A}_S = P^{(S)} A. \quad (1)$$

In other words, the criterion $\mathbf{F}(S)$ calculates the Frobenius norm of the residual matrix $E = A - \tilde{A}_S$. Other types of matrix norms can also be used to quantify the reconstruction error. Some of the recent work on the CSS problem [3, 6, 7] derives theoretical bounds for both the Frobenius and spectral norms of the residual matrix. The present work, however, focuses on developing algorithms that minimize the Frobenius norm of the residual matrix.

The projection matrix $P^{(S)}$ can be calculated as

$$P^{(S)} = A_{:S} \left(A_{:S}^T A_{:S} \right)^{-1} A_{:S}^T, \quad (2)$$

where $A_{:S}$ is the sub-matrix of A , which consists of the columns corresponding to S .

It should be noted that if the subset of columns S is known, the projection matrix $P^{(S)}$ can be derived as follows. The columns of the data matrix A can be approximated as linear combinations of the subset of columns S :

$$\tilde{A}_S = A_{:S} T,$$

where T is an $l \times n$ matrix of coefficients, which can be found by solving the following optimization problem.

$$T^* = \arg \min_T \|A - A_{:S} T\|_F^2.$$

This is a least-squares problem whose closed-form solution is $T^* = (A_{:S}^T A_{:S})^{-1} A_{:S}^T A$. Substituting with T^* in \tilde{A}_S gives

$$\tilde{A}_S = A_{:S} T = A_{:S} \left(A_{:S}^T A_{:S} \right)^{-1} A_{:S}^T A = P^{(S)} A.$$

The set of selected columns (i.e., data instances or features) can be directly presented to a data analyst to learn about the insights of the data, or they can be used to preprocess the data for further analysis. For instance, the selected columns can be used to obtain a low-dimensional

representation of all columns into the subspace of selected ones. This representation can be calculated as follows.

1. Calculate an orthonormal basis Q for the selected columns,

$$Q = \text{orth}(A_{:S}),$$

where $\text{orth}(\cdot)$ is a function that orthogonalizes the columns of its input matrix and Q is an $m \times l$ orthogonal matrix whose columns span the range of $A_{:S}$. The matrix Q can be obtained by applying an orthogonalization algorithm such as the Gram-Schmidt algorithm to the columns of $A_{:S}$, or by calculating the singular value decomposition (SVD) or the QR decomposition of $A_{:S}$ [30].

2. Embed all columns of A into the subspace of Q ,

$$W = Q^T A, \quad (3)$$

where W is an $l \times n$ matrix whose columns represent an embedding of all columns into the subspace of selected ones.

The selected columns can also be used to calculate a column-based low-rank approximation of A [22]. Given a subset S of columns with $|S| = l$, a rank- l approximation of the data matrix A can be calculated as:

$$\tilde{A}_S = P^{(S)} A = A_{:S} \left(A_{:S}^T A_{:S} \right)^{-1} A_{:S}^T A. \quad (4)$$

In order to calculate a rank- k approximation of the data matrix A where $k \leq l$, the following procedure suggested by Boutsidis et al. [3] can be used.

1. Calculate an orthonormal basis Q for the columns of $A_{:S}$ and embed all columns of A into the subspace of Q :

$$\begin{aligned} Q &= \text{orth}(A_{:S}), \\ W &= Q^T A, \end{aligned}$$

where Q is an $m \times l$ orthogonal matrix whose columns span the range of $A_{:S}$ and W is an $l \times n$ matrix whose columns represent an embedding of all columns into the subspace of selected ones.

2. Calculate the best rank- k approximation of the embedded columns using singular value decomposition (SVD):

$$\tilde{W}_k = U_k^{(W)} \Sigma_k^{(W)} V_k^{(W)T},$$

where $U_k^{(W)}$ and $V_k^{(W)}$ are $l \times k$ and $n \times k$ matrices whose columns represent the leading k left and right singular vectors of W , respectively, $\Sigma_k^{(W)}$ is a $k \times k$ matrix whose diagonal elements are the leading k singular values of W , and \tilde{W}_k is the best rank- k approximation of W .

3. Calculate the column-based rank- k approximation of A as:

$$\tilde{A}_{S,k} = Q \tilde{W}_k,$$

where $\tilde{A}_{S,k}$ is a rank- k approximation of A based on the set S of columns.

This procedure results in a rank- k approximation of A within the column space of $A_{:S}$ that achieves the minimum reconstruction error in terms of Frobenius norm [3]:

$$T^* = \arg \min_{T, \text{rank}(T)=k} \|A - A_{:S} T\|_F^2.$$

Moreover, the leading singular values and vectors of the low-dimensional embedding W can be used to approximate those of the data matrix as follows:

$$\tilde{U}_k^{(A)} = QU_k^{(W)}, \quad \tilde{\Sigma}_k^{(A)} = \Sigma_k^{(W)}, \quad \tilde{V}_k^{(A)} = V_k^{(W)} \quad (5)$$

where $\tilde{U}_k^{(A)}$ and $\tilde{V}_k^{(A)}$ are $l \times k$ and $n \times k$ matrices whose columns approximate the leading k left and right singular vectors of A , respectively, and $\tilde{\Sigma}_k^{(A)}$ is a $k \times k$ matrix whose diagonal elements approximate the leading k singular values of A .

3.2 MapReduce paradigm

MapReduce [15] was presented as a programming model to simplify large-scale data analytics over a distributed environment of commodity machines. The rationale behind MapReduce is to impose a set of constraints on data access at each individual machine and communication between different machines to ensure both the scalability and fault tolerance of the analytical tasks. Currently, MapReduce has been successfully used for scaling various data analysis tasks such as regression [50], feature selection [54], graph mining [39, 58], and most recently kernel k -means clustering [23].

A MapReduce job is executed in two phases of user-defined data transformation functions, namely *map* and *reduce* phases. The input data are split into physical blocks distributed among the nodes. Each block is viewed as a list of *key-value* pairs. In the first phase, the *key-value* pairs of each input block b are processed by a single *map* function running independently on the node where the block b is stored. The *key-value* pairs are provided one-by-one to the *map* function. The output of the *map* function is another set of intermediate *key-value* pairs. The values associated with the same *key* across all nodes are grouped together and provided as an input to the *reduce* function in the second phase. Different groups of values are processed in parallel on different machines. The output of each *reduce* function is a third set of *key-value* pairs and collectively considered the output of the job. It is important to note that the set of the intermediate *key-value* pairs is moved across the network between the nodes, which incurs significant additional execution time when much data are to be moved. For complex analytical tasks, multiple jobs are typically chained together [24] and/or many rounds of the same job are executed on the input data set [25].

In addition to the programming model constraints, Karloff et al. [40] defined a set of computational constraints that ensure the scalability and the efficiency of MapReduce-based analytical tasks. These computational constraints limit the used memory size at each machine, the output size of both the *map* and *reduce* functions and the number of rounds used to complete a certain tasks.

The MapReduce algorithms presented in this paper adhere to both the programming model constraints and the computational constraints. The proposed algorithm aims also at minimizing the overall running time of the distributed column subset selection task to facilitate interactive data analytics.

4 Greedy column subset selection

The column subset selection criterion presented in Sect. 3.1 measures the reconstruction error of a data matrix based on the subset of selected columns. The minimization of this criterion is a combinatorial optimization problem whose optimal solution can be obtained in $O(n^l mnl)$ [6]. This section describes a deterministic greedy algorithm for optimizing this criterion,

which extends the greedy method for unsupervised feature selection recently proposed by Farahat et al. [27, 28]. First, a recursive formula for the CSS criterion is presented, and then the greedy CSS algorithm is described in details.

4.1 Recursive selection criterion

The recursive formula of the CSS criterion is based on a recursive formula for the projection matrix $P^{(S)}$, which can be derived as follows.

Lemma 1 *Given a set of columns S . For any $\mathcal{P} \subset S$,*

$$P^{(S)} = P^{(\mathcal{P})} + R^{(\mathcal{R})},$$

where $R^{(\mathcal{R})}$ is a projection matrix which projects the columns of $E = A - P^{(\mathcal{P})}A$ onto the span of the subset $\mathcal{R} = S \setminus \mathcal{P}$ of columns,

$$R^{(\mathcal{R})} = E_{:\mathcal{R}} \left(E_{:\mathcal{R}}^T E_{:\mathcal{R}} \right)^{-1} E_{:\mathcal{R}}^T.$$

Proof Define a matrix $B = A_{:S}^T A_{:S}$, which represents the inner product over the columns of the sub-matrix $A_{:S}$. The projection matrix $P^{(S)}$ can be written as:

$$P^{(S)} = A_{:S} B^{-1} A_{:S}^T. \quad (6)$$

Without loss of generality, the columns and rows of $A_{:S}$ and B in Eq. (6) can be rearranged such that the first sets of rows and columns correspond to \mathcal{P} :

$$A_{:S} = [A_{:\mathcal{P}} \quad A_{:\mathcal{R}}], \quad B = \begin{bmatrix} B_{\mathcal{P}\mathcal{P}} & B_{\mathcal{P}\mathcal{R}} \\ B_{\mathcal{P}\mathcal{R}}^T & B_{\mathcal{R}\mathcal{R}} \end{bmatrix},$$

where $B_{\mathcal{P}\mathcal{P}} = A_{:\mathcal{P}}^T A_{:\mathcal{P}}$, $B_{\mathcal{P}\mathcal{R}} = A_{:\mathcal{P}}^T A_{:\mathcal{R}}$ and $B_{\mathcal{R}\mathcal{R}} = A_{:\mathcal{R}}^T A_{:\mathcal{R}}$.

Let $S = B_{\mathcal{R}\mathcal{R}} - B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}}$ be the Schur complement [48] of $B_{\mathcal{P}\mathcal{P}}$ in B . Using the block-wise inversion formula [48], B^{-1} can be calculated as:

$$B^{-1} = \begin{bmatrix} B_{\mathcal{P}\mathcal{P}}^{-1} + B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1} B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} & -B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1} \\ -S^{-1} B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} & S^{-1} \end{bmatrix}$$

Substitute with $A_{:S}$ and B^{-1} in Eq. (6):

$$\begin{aligned} P^{(S)} &= [A_{:\mathcal{P}} \quad A_{:\mathcal{R}}] \begin{bmatrix} B_{\mathcal{P}\mathcal{P}}^{-1} + B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1} B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} & -B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1} \\ -S^{-1} B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} & S^{-1} \end{bmatrix} \begin{bmatrix} A_{:\mathcal{P}}^T \\ A_{:\mathcal{R}}^T \end{bmatrix} \\ &= A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T + A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1} B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T - A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1} A_{:\mathcal{R}}^T \\ &\quad - A_{:\mathcal{R}} S^{-1} B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T + A_{:\mathcal{R}} S^{-1} A_{:\mathcal{R}}^T. \end{aligned}$$

Take out $A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1}$ as a common factor from the 2nd and 3rd terms, and $A_{:\mathcal{R}} S^{-1}$ from the 4th and 5th terms:

$$\begin{aligned} P^{(S)} &= A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T - A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} S^{-1} \left(A_{:\mathcal{R}}^T - B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T \right) \\ &\quad + A_{:\mathcal{R}} S^{-1} \left(A_{:\mathcal{R}}^T - B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T \right). \end{aligned}$$

Take out $S^{-1} \left(A_{:\mathcal{R}}^T - B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T \right)$ as a common factor from the 2nd and 3rd terms:

$$P^{(S)} = A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T + \left(A_{:\mathcal{R}} - A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} \right) S^{-1} \left(A_{:\mathcal{R}}^T - B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T \right). \quad (7)$$

The first term of Eq. (7) is the projection matrix, which projects the columns of A onto the span of the subset \mathcal{P} of columns: $P^{(\mathcal{P})} = A_{:\mathcal{P}} (A_{:\mathcal{P}}^T A_{:\mathcal{P}})^{-1} A_{:\mathcal{P}}^T = A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T$. The second term can be simplified as follows. Let E be an $m \times n$ residual matrix, which is calculated as: $E = A - P^{(\mathcal{P})} A$. The sub-matrix $E_{:\mathcal{R}}$ can be expressed as:

$$E_{:\mathcal{R}} = A_{:\mathcal{R}} - P^{(\mathcal{P})} A_{:\mathcal{R}} = A_{:\mathcal{R}} - A_{:\mathcal{P}} \left(A_{:\mathcal{P}}^T A_{:\mathcal{P}} \right)^{-1} A_{:\mathcal{P}}^T A_{:\mathcal{R}} = A_{:\mathcal{R}} - A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}}.$$

Since projection matrices are idempotent, then $P^{(\mathcal{P})} P^{(\mathcal{P})} = P^{(\mathcal{P})}$ and the inner product $E_{:\mathcal{R}}^T E_{:\mathcal{R}}$ can be expressed as:

$$\begin{aligned} E_{:\mathcal{R}}^T E_{:\mathcal{R}} &= \left(A_{:\mathcal{R}} - P^{(\mathcal{P})} A_{:\mathcal{R}} \right)^T \left(A_{:\mathcal{R}} - P^{(\mathcal{P})} A_{:\mathcal{R}} \right) \\ &= A_{:\mathcal{R}}^T A_{:\mathcal{R}} - A_{:\mathcal{R}}^T P^{(\mathcal{P})} A_{:\mathcal{R}} - A_{:\mathcal{R}}^T P^{(\mathcal{P})} A_{:\mathcal{R}} + A_{:\mathcal{R}}^T P^{(\mathcal{P})} P^{(\mathcal{P})} A_{:\mathcal{R}} \\ &= A_{:\mathcal{R}}^T A_{:\mathcal{R}} - A_{:\mathcal{R}}^T P^{(\mathcal{P})} A_{:\mathcal{R}}. \end{aligned}$$

Substituting with $P^{(\mathcal{P})} = A_{:\mathcal{P}} (A_{:\mathcal{P}}^T A_{:\mathcal{P}})^{-1} A_{:\mathcal{P}}^T$ gives

$$E_{:\mathcal{R}}^T E_{:\mathcal{R}} = A_{:\mathcal{R}}^T A_{:\mathcal{R}} - A_{:\mathcal{R}}^T A_{:\mathcal{P}} \left(A_{:\mathcal{P}}^T A_{:\mathcal{P}} \right)^{-1} A_{:\mathcal{P}}^T A_{:\mathcal{R}} = B_{\mathcal{R}\mathcal{R}} - B_{\mathcal{P}\mathcal{R}}^T B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}} = S.$$

Substituting $(A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} A_{:\mathcal{P}}^T)$, $(A_{:\mathcal{R}} - A_{:\mathcal{P}} B_{\mathcal{P}\mathcal{P}}^{-1} B_{\mathcal{P}\mathcal{R}})$ and S with $P^{(\mathcal{P})}$, $E_{:\mathcal{R}}$ and $E_{:\mathcal{R}}^T E_{:\mathcal{R}}$, respectively, Eq. (7) can be expressed as:

$$P^{(\mathcal{S})} = P^{(\mathcal{P})} + E_{:\mathcal{R}} \left(E_{:\mathcal{R}}^T E_{:\mathcal{R}} \right)^{-1} E_{:\mathcal{R}}^T.$$

The second term is the projection matrix which projects the columns of E onto the span of the subset \mathcal{R} of columns:

$$R^{(\mathcal{R})} = E_{:\mathcal{R}} \left(E_{:\mathcal{R}}^T E_{:\mathcal{R}} \right)^{-1} E_{:\mathcal{R}}^T. \quad (8)$$

This proves that $P^{(\mathcal{S})}$ can be written in terms of $P^{(\mathcal{P})}$ and R as: $P^{(\mathcal{S})} = P^{(\mathcal{P})} + R^{(\mathcal{R})}$ \square

This means that projection matrix $P^{(\mathcal{S})}$ can be constructed in a recursive manner by first calculating the projection matrix which projects the columns of A onto the span of the subset \mathcal{P} of columns, and then calculating the projection matrix which projects the columns of the residual matrix onto the span of the remaining columns. Based on this lemma, a recursive formula can be developed for \tilde{A}_S .

Corollary 1 Given a matrix A and a subset of columns S . For any $\mathcal{P} \subset S$,

$$\tilde{A}_S = \tilde{A}_{\mathcal{P}} + \tilde{E}_{\mathcal{R}},$$

where $E = A - P^{(\mathcal{P})} A$, and $\tilde{E}_{\mathcal{R}}$ is the low-rank approximation of E based on the subset $\mathcal{R} = S \setminus \mathcal{P}$ of columns.

Proof Using Lemma (1), and substituting with $P^{(\mathcal{S})}$ in Eq. (1) gives:

$$\tilde{A}_S = P^{(\mathcal{P})} A + E_{:\mathcal{R}} \left(E_{:\mathcal{R}}^T E_{:\mathcal{R}} \right)^{-1} E_{:\mathcal{R}}^T A. \quad (9)$$

The first term is the low-rank approximation of A based on \mathcal{P} : $\tilde{A}_{\mathcal{P}} = P^{(\mathcal{P})} A$. The second term is equal to $\tilde{E}_{\mathcal{R}}$ as $E_{:\mathcal{R}}^T A = E_{:\mathcal{R}}^T E$. To prove that, multiplying $E_{:\mathcal{R}}^T$ by $E = A - P^{(\mathcal{P})} A$ gives:

$$E_{:\mathcal{R}}^T E = E_{:\mathcal{R}}^T A - E_{:\mathcal{R}}^T P^{(\mathcal{P})} A.$$

Using $E_{:\mathcal{R}} = A_{:\mathcal{R}} - P^{(\mathcal{P})} A_{:\mathcal{R}}$, the expression $E_{:\mathcal{R}}^T P^{(\mathcal{P})}$ can be written as:

$$E_{:\mathcal{R}}^T P^{(\mathcal{P})} = A_{:\mathcal{R}}^T P^{(\mathcal{P})} - A_{:\mathcal{R}}^T P^{(\mathcal{P})} P^{(\mathcal{P})}.$$

This is equal to 0 as $P^{(\mathcal{P})} P^{(\mathcal{P})} = P^{(\mathcal{P})}$ (an idempotent matrix). This means that $E_{:\mathcal{R}}^T A = E_{:\mathcal{R}}^T E$. Substituting $E_{:\mathcal{R}}^T A$ with $E_{:\mathcal{R}}^T E$ in Eq. (9) proves the corollary. \square

This means that the column-based low-rank approximation of A based on the subset \mathcal{S} of columns can be calculated in a recursive manner by first calculating the low-rank approximation of A based on the subset $\mathcal{P} \subset \mathcal{S}$, and then calculating the low-rank approximation of the residual matrix E based on the remaining columns.

Based on Corollary (1), a recursive formula for the column subset selection criterion can be developed as follows.

Theorem 2 *Given a set of columns \mathcal{S} . For any $\mathcal{P} \subset \mathcal{S}$,*

$$\mathbf{F}(\mathcal{S}) = \mathbf{F}(\mathcal{P}) - \|\tilde{E}_{\mathcal{R}}\|_F^2,$$

where $E = A - P^{(\mathcal{P})} A$, and $\tilde{E}_{\mathcal{R}}$ is the low-rank approximation of E based on the subset $\mathcal{R} = \mathcal{S} \setminus \mathcal{P}$ of columns.

Proof Using Corollary (1), the CSS criterion can be expressed as:

$$\begin{aligned} \mathbf{F}(\mathcal{S}) &= \|A - \tilde{A}_{\mathcal{S}}\|_F^2 = \|A - \tilde{A}_{\mathcal{P}} - \tilde{E}_{\mathcal{R}}\|_F^2 \\ &= \|E - \tilde{E}_{\mathcal{R}}\|_F^2 = \|E - R^{(\mathcal{R})} E\|_F^2. \end{aligned}$$

Using the relation between the Frobenius norm and the trace function,¹ the right-hand side can be expressed as:

$$\begin{aligned} \|E - R^{(\mathcal{R})} E\|_F^2 &= \text{trace} \left((E - R^{(\mathcal{R})} E)^T (E - R^{(\mathcal{R})} E) \right) \\ &= \text{trace} \left(E^T E - 2E^T R^{(\mathcal{R})} E + E^T R^{(\mathcal{R})} R^{(\mathcal{R})} E \right). \end{aligned}$$

As $R^{(\mathcal{R})} R^{(\mathcal{R})} = R^{(\mathcal{R})}$ (an idempotent matrix), $\mathbf{F}(\mathcal{S})$ can be expressed as:

$$\mathbf{F}(\mathcal{S}) = \text{trace} \left(E^T E - E^T R^{(\mathcal{R})} R^{(\mathcal{R})} E \right) = \text{trace} \left(E^T E - \tilde{E}_{\mathcal{R}} \tilde{E}_{\mathcal{R}} \right) = \|E\|_F^2 - \|\tilde{E}_{\mathcal{R}}\|_F^2.$$

Replacing $\|E\|_F^2$ with $\mathbf{F}(\mathcal{P})$ proves the theorem. \square

The term $\|\tilde{E}_{\mathcal{R}}\|_F^2$ represents the decrease in reconstruction error achieved by adding the subset \mathcal{R} of columns to \mathcal{P} . In the following section, a novel greedy heuristic is presented to optimize the column subset selection criterion based on this recursive formula.

4.2 Greedy selection algorithm

This section presents an efficient greedy algorithm to optimize the column subset selection criterion presented in Sect. 3.1. The algorithm selects at each iteration one column such that the reconstruction error for the new set of columns is minimized. This problem can be formulated as follows:

¹ $\|A\|_F^2 = \text{trace}(A^T A)$.

Problem 2 At iteration t , find column p such that,

$$p = \arg \min_i \mathbf{F}(S \cup \{i\}) \quad (10)$$

where S is the set of columns selected during the first $t - 1$ iterations.

A naïve implementation of the greedy algorithm is to calculate the reconstruction error for each candidate column and then select the column with the smallest error. This implementation is, however, computationally very complex, as it requires $O(m^2 n^2)$ floating-point operations per iteration. A more efficient approach is to use the recursive formula for calculating the reconstruction error. Using Theorem 2,

$$\mathbf{F}(S \cup \{i\}) = \mathbf{F}(S) - \|\tilde{E}_{\{i\}}\|_F^2,$$

where $E = A - \tilde{A}_S$ and $\tilde{E}_{\{i\}}$ is the rank-1 approximation of E based on the candidate column i . Since $\mathbf{F}(S)$ is a constant for all candidate columns, an equivalent criterion is:

$$p = \arg \max_i \|\tilde{E}_{\{i\}}\|_F^2 \quad (11)$$

This formulation selects the column p which achieves the maximum decrease in reconstruction error. Using the properties that: $\text{trace}(AB) = \text{trace}(BA)$ and $\text{trace}(aA) = a \text{trace}(A)$ where a is a scalar, the new objective function $\|\tilde{E}_{\{i\}}\|_F^2$ can be simplified as follows:

$$\begin{aligned} \|\tilde{E}_{\{i\}}\|_F^2 &= \text{trace}(\tilde{E}_{\{i\}}^T \tilde{E}_{\{i\}}) = \text{trace}(E^T R^{(i)} E) \\ &= \text{trace}\left(E^T E_{:i} (E_{:i}^T E_{:i})^{-1} E_{:i}^T E\right) \\ &= \frac{1}{E_{:i}^T E_{:i}} \text{trace}(E^T E_{:i} E_{:i}^T E) = \frac{\|E^T E_{:i}\|^2}{E_{:i}^T E_{:i}}. \end{aligned}$$

This defines the following equivalent problem.

Problem 3 (*Greedy Column Subset Selection*) At iteration t , find column p such that,

$$p = \arg \max_i \frac{\|E^T E_{:i}\|^2}{E_{:i}^T E_{:i}} \quad (12)$$

where $E = A - \tilde{A}_S$, and S is the set of columns selected during the first $t - 1$ iterations.

The computational complexity of this selection criterion is $O(n^2 m)$ per iteration, and it requires $O(nm)$ memory to store the residual of the whole matrix, E , after each iteration. In order to reduce these memory requirements, a memory-efficient algorithm can be proposed calculate the column subset selection criterion without explicitly calculating and storing the residual matrix E at each iteration. The algorithm is based on a recursive formula for calculating the residual matrix E .

Let $S^{(t)}$ denote the set of columns selected during the first $t - 1$ iterations, $E^{(t)}$ denote the residual matrix at the start of the t -th iteration (i.e., $E^{(t)} = A - \tilde{A}_{S^{(t)}}$), and $p^{(t)}$ be the column selected at iteration t . The following lemma gives a recursive formula for residual matrix at the start of iteration $t + 1$, $E^{(t+1)}$.

Lemma 2 $E^{(t+1)}$ can be calculated recursively as:

$$E^{(t+1)} = \left(E - \frac{E_{:p} E_{:p}^T}{E_{:p}^T E_{:p}} E \right)^{(t)}. \quad (13)$$

Proof Using Corollary 1, $\tilde{A}_{S \cup \{p\}} = \tilde{A}_S + \tilde{E}_{\{p\}}$. Subtracting both sides from A , and substituting $A - \tilde{A}_{S \cup \{p\}}$ and $A - \tilde{A}_S$ with $E^{(t+1)}$ and $E^{(t)}$, respectively, gives:

$$E^{(t+1)} = (E - \tilde{E}_{\{p\}})^{(t)}$$

Using Eqs. (1) and (2), $\tilde{E}_{\{p\}}$ can be expressed as $(E_{:p} (E_{:p}^T E_{:p})^{-1} E_{:p}^T) E$. Substituting $\tilde{E}_{\{p\}}$ with this formula in the above equation proves the lemma. \square

Let G be an $n \times n$ matrix which represents the inner products over the columns of the residual matrix E : $G = E^T E$. The following corollary is a direct result of Lemma 2.

Corollary 3 $G^{(t+1)}$ can be calculated recursively as:

$$G^{(t+1)} = \left(G - \frac{G_{:p} G_{:p}^T}{G_{pp}} \right)^{(t)}.$$

Proof This corollary can be proved by substituting with $E^{(t+1)T}$ (Lemma 2) in $G^{(t+1)} = E^{(t+1)T} E^{(t+1)}$, and using the fact that $R^{(\{p\})} R^{(\{p\})} = R^{(\{p\})}$ (an idempotent matrix). \square

To simplify the derivation of the memory-efficient algorithm, at iteration t , define $\delta = G_{:p}$ and $\omega = G_{:p} / \sqrt{G_{pp}} = \delta / \sqrt{\delta_p}$. This means that $G^{(t+1)}$ can be calculated in terms of $G^{(t)}$ and $\omega^{(t)}$ as follows:

$$G^{(t+1)} = (G - \omega \omega^T)^{(t)}, \quad (14)$$

or in terms of A and previous ω 's as:

$$G^{(t+1)} = A^T A - \sum_{r=1}^t (\omega \omega^T)^{(r)}. \quad (15)$$

$\delta^{(t)}$ and $\omega^{(t)}$ can be calculated in terms of A and previous ω 's as follows:

$$\begin{aligned} \delta^{(t)} &= A^T A_{:p} - \sum_{r=1}^{t-1} \omega_p^{(r)} \omega^{(r)}, \\ \omega^{(t)} &= \delta^{(t)} / \sqrt{\delta_p^{(t)}}. \end{aligned} \quad (16)$$

The column subset selection criterion can be expressed in terms of G as:

$$p = \arg \max_i \frac{\|G_{:i}\|^2}{G_{ii}}$$

The following theorem gives recursive formulas for calculating the column subset selection criterion without explicitly calculating E or G .

Theorem 4 Let $f_i = \|G_{:i}\|^2$ and $g_i = G_{ii}$ be the numerator and denominator of the criterion function for column i , respectively, $\mathbf{f} = [f_i]_{i=1,\dots,n}$, and $\mathbf{g} = [g_i]_{i=1,\dots,n}$. Then,

$$\begin{aligned} \mathbf{f}^{(t)} &= \left(\mathbf{f} - 2 \left(\boldsymbol{\omega} \circ \left(A^T A \boldsymbol{\omega} - \sum_{r=1}^{t-2} \left(\boldsymbol{\omega}^{(r)T} \boldsymbol{\omega} \right) \boldsymbol{\omega}^{(r)} \right) \right) + \|\boldsymbol{\omega}\|^2 (\boldsymbol{\omega} \circ \boldsymbol{\omega}) \right)^{(t-1)}, \\ \mathbf{g}^{(t)} &= \left(\mathbf{g} - (\boldsymbol{\omega} \circ \boldsymbol{\omega}) \right)^{(t-1)}. \end{aligned}$$

where \circ represents the Hadamard product operator.

Proof Based on Eq. (14), $\mathbf{f}_i^{(t)}$ can be calculated as:

$$\begin{aligned} \mathbf{f}_i^{(t)} &= (\|G_{:i}\|^2)^{(t)} = (\|G_{:i} - \boldsymbol{\omega}_i \boldsymbol{\omega}\|^2)^{(t-1)} \\ &= \left((G_{:i} - \boldsymbol{\omega}_i \boldsymbol{\omega})^T (G_{:i} - \boldsymbol{\omega}_i \boldsymbol{\omega}) \right)^{(t-1)} \\ &= \left(G_{:i}^T G_{:i} - 2\boldsymbol{\omega}_i G_{:i}^T \boldsymbol{\omega} + \boldsymbol{\omega}_i^2 \|\boldsymbol{\omega}\|^2 \right)^{(t-1)} \\ &= \left(f_i - 2\boldsymbol{\omega}_i G_{:i}^T \boldsymbol{\omega} + \boldsymbol{\omega}_i^2 \|\boldsymbol{\omega}\|^2 \right)^{(t-1)}. \end{aligned} \quad (17)$$

Similarly, $\mathbf{g}_i^{(t)}$ can be calculated as:

$$\begin{aligned} \mathbf{g}_i^{(t)} &= G_{ii}^{(t)} = (G_{ii} - \boldsymbol{\omega}_i^2)^{(t-1)} \\ &= (\mathbf{g}_i - \boldsymbol{\omega}_i^2)^{(t-1)}. \end{aligned} \quad (18)$$

Let $\mathbf{f} = [f_i]_{i=1,\dots,n}$ and $\mathbf{g} = [g_i]_{i=1,\dots,n}$, $\mathbf{f}^{(t)}$ and $\mathbf{g}^{(t)}$ can be expressed as:

$$\begin{aligned} \mathbf{f}^{(t)} &= \left(\mathbf{f} - 2 (\boldsymbol{\omega} \circ G \boldsymbol{\omega}) + \|\boldsymbol{\omega}\|^2 (\boldsymbol{\omega} \circ \boldsymbol{\omega}) \right)^{(t-1)}, \\ \mathbf{g}^{(t)} &= (\mathbf{g} - (\boldsymbol{\omega} \circ \boldsymbol{\omega}))^{(t-1)}, \end{aligned} \quad (19)$$

where \circ represents the Hadamard product operator, and $\|\cdot\|$ is the ℓ_2 norm.

Based on the recursive formula of G (Eq. 15), the term $G \boldsymbol{\omega}$ at iteration $(t-1)$ can be expressed as:

$$\begin{aligned} G \boldsymbol{\omega} &= \left(A^T A - \sum_{r=1}^{t-2} \left(\boldsymbol{\omega} \boldsymbol{\omega}^T \right)^{(r)} \right) \boldsymbol{\omega} \\ &= A^T A \boldsymbol{\omega} - \sum_{r=1}^{t-2} \left(\boldsymbol{\omega}^{(r)T} \boldsymbol{\omega} \right) \boldsymbol{\omega}^{(r)} \end{aligned} \quad (20)$$

Substituting with $G \boldsymbol{\omega}$ in Eq. (19) gives the update formulas for \mathbf{f} and \mathbf{g} . \square

This means that the greedy criterion can be memory efficient by only maintaining two score variables for each column, \mathbf{f}_i and \mathbf{g}_i , and updating them at each iteration based on their previous values and the columns selected so far.

Algorithm 1 shows the complete greedy CSS algorithm.

Algorithm 1 Greedy Column Subset Selection**Input:** Data matrix A , Number of columns l **Output:** Selected subset of columns \mathcal{S}

- 1: Initialize $\mathcal{S} = \{\}$
- 2: Initialize $f_i^{(0)} = \|A^T A_{:,i}\|^2$, $g_i^{(0)} = A_{:,i}^T A_{:,i}$ for $i = 1 \dots n$
- 3: Repeat $t = 1 \rightarrow l$:
- 4: $p = \arg \max_i f_i^{(t)} / g_i^{(t)}$, $\mathcal{S} = \mathcal{S} \cup \{p\}$
- 5: $\delta^{(t)} = A^T A_{:,p} - \sum_{r=1}^{t-1} \omega_p^{(r)} \omega^{(r)}$
- 6: $\omega^{(t)} = \delta^{(t)} / \sqrt{\delta_p^{(t)}}$
- 7: Update f_i 's, g_i 's (Theorem 4)

5 Distributed column subset selection on MapReduce

This section describes a MapReduce algorithm for the distributed column subset selection problem. Given a big data matrix A whose columns are distributed across different machines, the goal is to select a subset of columns \mathcal{S} from A such that the CSS criterion $\mathbf{F}(\mathcal{S})$ is minimized.

One naïve approach to perform distributed column subset selection is to select different subsets of columns from the sub-matrices stored on different machines. The selected subsets are then sent to a centralized machine where an additional selection step is optionally performed to filter out irrelevant or redundant columns. Let $A_{(i)}$ be the sub-matrix stored at machine i , the naïve approach optimizes the following function.

$$\sum_{i=1}^c \|A_{(i)} - P^{(\mathcal{L}_{(i)})} A_{(i)}\|_F^2, \quad (21)$$

where $\mathcal{L}_{(i)}$ is the set of columns selected from $A_{(i)}$ and c is the number of physical blocks of data. The resulting set of columns is the union of the sets selected from different sub-matrices: $\mathcal{L} = \cup_{i=1}^c \mathcal{L}_{(i)}$. The set \mathcal{L} can further be reduced by invoking another selection process in which a smaller subset of columns is selected from $A_{:\mathcal{L}}$.

The naïve approach, however simple, is prone to missing relevant columns. This is because the selection at each machine is based on approximating a local sub-matrix, and accordingly, there is no way to determine whether the selected columns are globally relevant or not. For instance, suppose the extreme case where all the truly representative columns happen to be loaded on a single machine. In this case, the algorithm will select a less-than-required number of columns from that machine and many irrelevant columns from other machines.

In order to alleviate this problem, the different machines have to select columns that best approximate a common representation of the data matrix. To achieve that, the proposed algorithm first learns a concise representation of the span of the big data matrix. This concise representation is relatively small, and it can be sent over to all machines. After that, each machine can select columns from its sub-matrix that approximate this concise representation. The proposed algorithm uses random projection to learn this concise representation and proposes a generalized column subset selection (CSS) method to select columns from different machines. The details of the proposed methods are explained in the rest of this section.

5.1 Random projection

The first step of the proposed algorithm is to learn a concise representation B for a distributed data matrix A . In the proposed approach, a random projection method is employed. Random projection [1, 14, 46] is a well-known technique for dealing with the curse-of-the-dimensionality problem. Let Ω be a random projection matrix of size $n \times r$, and given a data matrix X of size $m \times n$, the random projection can be calculated as $Y = X\Omega$. It has been shown that applying random projection Ω to X preserves the pairwise distances between vectors in the row space of X with a high probability [14]:

$$(1 - \epsilon) \|X_{i:} - X_{j:}\| \leq \|X_{i:}\Omega - X_{j:}\Omega\| \leq (1 + \epsilon) \|X_{i:} - X_{j:}\|, \quad (22)$$

where ϵ is an arbitrarily small factor.

Since the CSS criterion $\mathbf{F}(S)$ measures the reconstruction error between the big data matrix A and its low-rank approximation $P^{(S)}A$, it essentially measures the sum of the distances between the original rows and their approximations. This means that when applying random projection to both A and $P^{(S)}A$, the reconstruction error of the original data matrix A will be approximately equal to that of $A\Omega$ when both are approximated using the subset of selected columns:

$$\|A - P^{(S)}A\|_F^2 \approx \|A\Omega - P^{(S)}A\Omega\|_F^2. \quad (23)$$

So, instead of optimizing $\|A - P^{(S)}A\|_F^2$, the distributed CSS can approximately optimize $\|A\Omega - P^{(S)}A\Omega\|_F^2$.

Let $B = A\Omega$, the distributed column subset selection problem can be formally defined as

Problem 4 (*Distributed Column Subset Selection*) Given an $m \times n_{(i)}$ sub-matrix $A_{(i)}$ which is stored at node i and an integer $l_{(i)}$, find a subset of columns $\mathcal{L}_{(i)}$ such that $|\mathcal{L}_{(i)}| = l_{(i)}$ and

$$\mathcal{L}_{(i)} = \arg \min_S \|B - P^{(S)}B\|_F^2,$$

where $B = A\Omega$, Ω is an $n \times r$ random projection matrix, S is the set of the indices of the candidate columns and $\mathcal{L}_{(i)}$ is the set of the indices of the selected columns from $A_{(i)}$.

A key observation here is that random projection matrices whose entries are sampled *i.i.d* from some univariate distribution Ψ can be exploited to compute random projection on MapReduce in a very efficient manner. Examples of such matrices are Gaussian random matrices [14], uniform random sign (± 1) matrices [1], and sparse random sign matrices [46].

In order to implement random projection on MapReduce, the data matrix A is distributed in a column-wise fashion and viewed as pairs of $\langle i, A_{:i} \rangle$ where $A_{:i}$ is the i -th column of A . Recall that $B = A\Omega$ can be rewritten as

$$B = \sum_{i=1}^n A_{:i} \Omega_{i:} \quad (24)$$

and since the *map* function is provided one columns of A at a time, one does not need to worry about pre-computing the full matrix Ω . In fact, for each input column $A_{:i}$, a new vector $\Omega_{i:}$ needs to be sampled from Ψ . So, each input column generates a matrix of size $m \times r$, which means that $O(nmr)$ data should be moved across the network to sum the generated n matrices at m independent *reducers* each summing a row $B_{j:}$ to obtain B . To minimize that network cost, an *in-memory* summation can be carried out over the generated $m \times r$

Algorithm 2 Fast Random Projection on MapReduce**Input:** Data matrix A , Univariate distribution Ψ , Number of dimensions r **Output:** Concise representation $B = A\Omega$, $\Omega_{ij} \sim \Psi \forall i, j$

```

1: map:
2:    $\bar{B} = [0]_{m \times r}$ 
3:   foreach  $\langle i, A_{:i} \rangle$ 
4:     Generate  $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r]$ ,  $\mathbf{v}_j \sim \Psi$ 
5:      $\bar{B} = \bar{B} + A_{:i} \mathbf{v}$ 
6:   for  $j = 1$  to  $m$ 
7:     emit  $\langle j, \bar{B}_{j:} \rangle$ 

8: reduce:
9:   foreach  $\langle j, [\bar{B}_{(1)}]_{j:}, [\bar{B}_{(2)}]_{j:}, \dots, [\bar{B}_{(c)}]_{j:} \rangle$ 
10:     $B_{j:} = \sum_{i=1}^c [\bar{B}_{(i)}]_{j:}$ 
11:    emit  $\langle j, B_{j:} \rangle$ 

```

matrices at each mapper. This can be done incrementally after processing each column of A . That optimization reduces the network cost to $O(cmr)$, where c is the number of physical blocks of the matrix.² Algorithm 2 outlines the proposed random projection algorithm. The term *emit* is used to refer to outputting new $\langle \text{key}, \text{value} \rangle$ pairs from a mapper or a reducer.

5.2 Generalized column subset selection

This section presents the generalized column subset selection algorithm, which will be used to perform the selection of columns at different machines. While Problem 1 is concerned with the selection of a subset of columns from a data matrix which best represent other columns of the same matrix, Problem 4 selects a subset of columns from a source matrix which best represent the columns of a different target matrix. The objective function of Problem 4 represents the reconstruction error of the target matrix B based on the selected columns from the source matrix, and the term $P^{(S)} = A_{:S} (A_{:S}^T A_{:S})^{-1} A_{:S}^T$ is the projection matrix, which projects the columns of B onto the subspace of the columns selected from A .

In order to optimize this new criterion, a greedy algorithm can be introduced. Let $\bar{\mathbf{F}}(S) = \|B - P^{(S)}B\|_F^2$ be the distributed CSS criterion, the following theorem derives a recursive formula for $\bar{\mathbf{F}}(S)$.

Theorem 5 *Given a set of columns S . For any $\mathcal{P} \subset S$,*

$$\bar{\mathbf{F}}(S) = \bar{\mathbf{F}}(\mathcal{P}) - \left\| \tilde{F}_{\mathcal{R}} \right\|_F^2,$$

where $F = B - P^{(\mathcal{P})}B$, and $\tilde{F}_{\mathcal{R}}$ is the low-rank approximation of F based on the subset $\mathcal{R} = S \setminus \mathcal{P}$ of columns of $E = A - P^{(\mathcal{P})}A$.

Proof Using the recursive formula for the low-rank approximation of A : $\tilde{A}_S = \tilde{A}_{\mathcal{P}} + \tilde{E}_{\mathcal{R}}$, and multiplying both sides with Ω gives

$$\tilde{A}_S \Omega = \tilde{A}_{\mathcal{P}} \Omega + \tilde{E}_{\mathcal{R}} \Omega.$$

Low-rank approximations can be written in terms of projection matrices as

$$P^{(S)} A \Omega = P^{(\mathcal{P})} A \Omega + R^{(\mathcal{R})} E \Omega.$$

² The *in-memory* summation can also be replaced by a MapReduce combiner, Dean2008.

Algorithm 3 Greedy Generalized Column Subset Selection**Input:** Source matrix A , Target matrix B , Number of columns l **Output:** Selected subset of columns \mathcal{S} 1: Initialize $f_i^{(0)} = \|B^T A_{:i}\|^2$, $g_i^{(0)} = A_{:i}^T A_{:i}$ for $i = 1 \dots n$ 2: Repeat $t = 1 \rightarrow l$:3: $p = \arg \max_i f_i^{(t)} / g_i^{(t)}$, $\mathcal{S} = \mathcal{S} \cup \{p\}$ 4: $\delta^{(t)} = A^T A_{:p} - \sum_{r=1}^{t-1} \omega_p^{(r)} \omega^{(r)}$ 5: $\gamma^{(t)} = B^T A_{:p} - \sum_{r=1}^{t-1} \omega_p^{(r)} \gamma^{(r)}$ 6: $\omega^{(t)} = \delta^{(t)} / \sqrt{\delta_p^{(t)}}$, $\gamma^{(t)} = \gamma^{(t)} / \sqrt{\delta_p^{(t)}}$ 7: Update f_i 's, g_i 's (Theorem 6)Using $B = A\Omega$,

$$P^{(\mathcal{S})} B = P^{(\mathcal{P})} B + R^{(\mathcal{R})} E\Omega.$$

Let $F = E\Omega$. The matrix F is the residual after approximating B using the set \mathcal{P} of columns

$$F = E\Omega = (A - P^{(\mathcal{P})} A) \Omega = A\Omega - P^{(\mathcal{P})} A\Omega = B - P^{(\mathcal{P})} B.$$

This means that

$$P^{(\mathcal{S})} B = P^{(\mathcal{P})} B + R^{(\mathcal{R})} F$$

Substituting in $\bar{\mathbf{F}}(\mathcal{S}) = \|B - P^{(\mathcal{S})} B\|_F^2$ gives

$$\bar{\mathbf{F}}(\mathcal{S}) = \|B - P^{(\mathcal{P})} B - R^{(\mathcal{R})} F\|_F^2$$

Using $F = B - P^{(\mathcal{P})} B$ gives

$$\bar{\mathbf{F}}(\mathcal{S}) = \|F - R^{(\mathcal{R})} F\|_F^2$$

Using the relation between Frobenius norm and trace,

$$\begin{aligned} \bar{\mathbf{F}}(\mathcal{S}) &= \text{trace} \left((F - R^{(\mathcal{R})} F)^T (F - R^{(\mathcal{R})} F) \right) \\ &= \text{trace} \left(F^T F - 2F^T R^{(\mathcal{R})} F + F^T R^{(\mathcal{R})} R^{(\mathcal{R})} F \right) \\ &= \text{trace} \left(F^T F - F^T R^{(\mathcal{R})} F \right) = \|F\|_F^2 - \|R^{(\mathcal{R})} F\|_F^2 \end{aligned}$$

Using $\bar{\mathbf{F}}(\mathcal{P}) = \|F\|_F^2$ and $\tilde{F}_{\mathcal{R}} = R^{(\mathcal{R})} F$ proves the theorem. \square Using the recursive formula for $\bar{\mathbf{F}}(\mathcal{S} \cup \{i\})$ allows the development of a greedy algorithm which at iteration t optimizes

$$p = \arg \min_i \bar{\mathbf{F}}(\mathcal{S} \cup \{i\}) = \arg \max_i \|\tilde{F}_{(i)}\|_F^2 \quad (25)$$

Let $G = E^T E$ and $H = F^T E$, the objective function of this optimization problem can be simplified as follows

$$\begin{aligned}\|\tilde{F}_{(i)}\|_F^2 &= \left\| E_{:i} \left(E_{:i}^T E_{:i} \right)^{-1} E_{:i}^T F \right\|_F^2 \\ &= \text{trace} \left(F^T E_{:i} \left(E_{:i}^T E_{:i} \right)^{-1} E_{:i}^T F \right) \\ &= \frac{\|F^T E_{:i}\|^2}{E_{:i}^T E_{:i}} = \frac{\|H_{:i}\|^2}{G_{ii}}.\end{aligned}\quad (26)$$

This allows the definition of the following generalized CSS problem.

Problem 5 (*Greedy Generalized CSS*) At iteration t , find column p such that

$$p = \arg \max_i \frac{\|H_{:i}\|^2}{G_{ii}}$$

where $H = F^T E$, $G = E^T E$, $F = B - P^{(S)} B$, $E = A - P^{(S)} A$ and S is the set of columns selected during the first $t - 1$ iterations.

For iteration t , define $\boldsymbol{\gamma} = H_{:p}$ and $\boldsymbol{v} = H_{:p} / \sqrt{G_{pp}} = \boldsymbol{\gamma} / \sqrt{\delta_p}$. The vector $\boldsymbol{\gamma}^{(t)}$ can be calculated in terms of A , B and previous $\boldsymbol{\omega}$'s and \boldsymbol{v} 's as $\boldsymbol{\gamma}^{(t)} = B^T A_{:p} - \sum_{r=1}^{t-1} \boldsymbol{\omega}_p^{(r)} \boldsymbol{v}^{(r)}$.

Similarly, the numerator and denominator of the selection criterion at each iteration can be calculated in an efficient manner using the following theorem.

Theorem 6 Let $f_i = \|H_{:i}\|^2$ and $g_i = G_{ii}$ be the numerator and denominator of the greedy criterion function for column i , respectively, $\boldsymbol{f} = [f_i]_{i=1,\dots,n}$, and $\boldsymbol{g} = [g_i]_{i=1,\dots,n}$. Then,

$$\begin{aligned}\boldsymbol{f}^{(t)} &= \left(\boldsymbol{f} - 2 \left(\boldsymbol{\omega} \circ \left(A^T B \boldsymbol{v} - \sum_{r=1}^{t-2} \left(\boldsymbol{v}^{(r)T} \boldsymbol{v} \right) \boldsymbol{\omega}^{(r)} \right) \right) + \|\boldsymbol{v}\|^2 \left(\boldsymbol{\omega} \circ \boldsymbol{\omega} \right) \right)^{(t-1)}, \\ \boldsymbol{g}^{(t)} &= \left(\boldsymbol{g} - \left(\boldsymbol{\omega} \circ \boldsymbol{\omega} \right) \right)^{(t-1)},\end{aligned}$$

where \circ represents the Hadamard product operator.

As outlined in Sect. 5.1, the algorithm's distribution strategy is based on sharing the concise representation of the data B among all mappers. Then, independent $l_{(b)}$ columns from each mapper are selected using the generalized CSS algorithm. A second phase of selection is run over the $\sum_{b=1}^c l_{(b)}$ (where c is the number of input blocks) columns to find the best l columns to represent B . Different ways can be used to set $l_{(b)}$ for each input block b . In the context of this paper, the set of $l_{(b)}$ is assigned uniform values for all blocks (i.e., $l_{(b)} = \lfloor l/c \rfloor \forall b \in \{1, 2, \dots, c\}$). Algorithm 4 sketches the MapReduce implementation of the distributed CSS algorithm. It should be emphasized that the proposed MapReduce algorithm requires only two passes over the data set, and its moves a very few amount of the data across the network.

Regarding the memory requirements of the proposed method, in the map phase of the random projection stage (Algorithm 2), the needed memory per node is of $O(mr)$, which is that required to store the matrix \bar{B} . In the reduce phase, Algorithm 2 requires only $O(r)$ memory to sum the corresponding rows of \bar{B}_i . The memory complexity of the map phase of the selection stage (Algorithm 4) is that required for storing the matrices $A_{(b)}$ and B . Let q denote the number of columns loaded into the matrix $A_{(b)}$ in Algorithm 4 (lines 3:4); the memory complexity of the mapper of Algorithm 4 is of $O(m(q + r))$. Similarly, let

Algorithm 4 Distributed CSS on MapReduce**Input:** Matrix A of size $m \times n$, Concise representation B , Number of columns l **Output:** Selected columns C

```

1: map:
2:    $A_{(b)} = []$ 
3:   foreach  $\langle i, A_{:i} \rangle$ 
4:      $A_{(b)} = [A_{(b)} \ A_{:i}]$ 
5:    $\tilde{S} = \text{GeneralizedCSS}(A_{(b)}, B, l_{(b)})$ 
6:   foreach  $j$  in  $\tilde{S}$ 
7:     emit  $\langle 0, [A_{(b)}]_{:j} \rangle$ 

8: reduce:
9:   For all values  $\{[A_{(1)}]_{:\tilde{S}_{(1)}}, [A_{(2)}]_{:\tilde{S}_{(2)}}, \dots, [A_{(c)}]_{:\tilde{S}_{(c)}}\}$ 
10:   $A_{(0)} = \left[ [A_{(1)}]_{:\tilde{S}_{(1)}}, [A_{(2)}]_{:\tilde{S}_{(2)}}, \dots, [A_{(c)}]_{:\tilde{S}_{(c)}} \right]$ 
11:   $S = \text{GeneralizedCSS}(A_{(0)}, B, l)$ 
12:  foreach  $j$  in  $S$ 
13:    emit  $\langle 0, [A_{(0)}]_{:j} \rangle$ 

```

$p = \sum_{b=1}^c l_{(b)}$ denote the number of columns loaded into the matrix $A_{(0)}$ in Algorithm 4 (lines 9:10). The memory complexity of the reduce phase of Algorithm 4 is of $O(m(p+r))$. Hence, the overall memory complexity per node required by the proposed approach can be bounded by $O(m(r + \max(q, p)))$.

6 Related work

Different approaches have been proposed for selecting a subset of representative columns from a data matrix. This section focuses on briefly describing these approaches and their applicability to massively distributed data matrices. The column subset selection (CSS) methods can be generally categorized into randomized, deterministic and hybrid.

6.1 Randomized methods

The randomized methods sample a subset of columns from the original matrix using carefully chosen sampling probabilities. The main focus of this category of methods is to develop fast algorithms for column subset selection and then derive a bound for the reconstruction error of the data matrix based on the selected columns relative to the best possible reconstruction error obtained using Singular Value Decomposition (SVD).

Frieze et al. [29] was the first to suggest the idea of randomly sampling l columns from a matrix and using these columns to calculate a rank- k approximation of the matrix (where $l \geq k$). The authors derived an additive bound for the reconstruction error of the data matrix. This work of Frieze et al. was followed by different papers [20, 21] that enhanced the algorithm by proposing different sampling probabilities and deriving better error bounds for the reconstruction error. Drineas et al. [22] proposed a subspace sampling method which samples columns using probabilities proportional to the norms of the rows of the top k right singular vectors of A . The subspace sampling method allows the development of a relative error bound (i.e., a multiplicative error bound relative to the best rank- k approximation). However, the subspace sampling depends on calculating the leading singular vectors of a matrix which is computationally very complex for large matrices.

Deshpande et al. [18, 19] proposed an adaptive sampling method which updates the sampling probabilities based on the columns selected so far. This method is computationally very complex, as it depends on calculating the residual of the data matrix after each iteration. In the same paper, Deshpande et al. also proved the existence of a volume sampling algorithm (i.e., sampling a subset of columns based on the volume enclosed by their vectors) which achieves a multiplicative $(l + 1)$ -approximation. However, the authors did not present a polynomial time algorithm for this volume sampling algorithm.

Column subset selection with uniform sampling can be easily implemented on MapReduce. For non-uniform sampling, the efficiency of implementing the selection on MapReduce is determined by how easy are the calculations of the sampling probabilities. The calculations of probabilities that depend on calculating the leading singular values and vectors are time-consuming on MapReduce. On the other hand, adaptive sampling methods are computationally very complex as they depend on calculating the residual of the whole data matrix after each iteration.

6.2 Deterministic methods

The second category of methods employs a deterministic algorithm for selecting columns such that some criterion function is minimized. This criterion function usually quantifies the reconstruction error of the data matrix based on the subset of selected columns. The deterministic methods are slower, but more accurate, than the randomized ones.

In the area of numerical linear algebra, the column pivoting method exploited by the QR decomposition [30] permutes the columns of the matrix based on their norms to enhance the numerical stability of the QR decomposition algorithm. The first l columns of the permuted matrix can be directly selected as representative columns. The Rank-Revealing QR (RRQR) decomposition [2, 11, 31, 52] is a category of QR decomposition methods which permute columns of the data matrix while imposing additional constraints on the singular values of the two sub-matrices of the upper triangular matrix R corresponding to the selected and non-selected columns. It has been shown that the constraints on the singular values can be used to derive an theoretical guarantee for the column-based reconstruction error according to spectral norm [7].

Besides methods based on QR decomposition, different recent methods have been proposed for directly selecting a subset of columns from the data matrix. Boutsidis et al. [7] proposed a deterministic column subset selection method which first groups columns into clusters and then selects a subset of columns from each cluster. The authors proposed a general framework in which different clustering and subset selection algorithms can be employed to select a subset of representative columns. Çivril and Magdon-Ismail [9, 10] presented a deterministic algorithm which greedily selects columns from the data matrix that best represent the right leading singular values of the matrix. This algorithm, however accurate, depends on the calculation of the leading singular vectors of a matrix, which is computationally very complex for large matrices.

Recently, Boutsidis et al. [3] presented a column subset selection algorithm which first calculates the top- k right singular values of the data matrix (where k is the target rank) and then uses deterministic sparsification methods to select $l \geq k$ columns from the data matrix. The authors derived a theoretically near-optimal error bound for the rank- k column-based approximation. Deshpande and Rademacher [17] presented a polynomial time deterministic algorithm for volume sampling with a theoretical guarantee for $l = k$. Quite recently, Guruswami and Sinop [32] presented a deterministic algorithm for volume sampling with theoretical guarantee for $l > k$. The deterministic volume sampling algorithms are, however,

more complex than the algorithms presented in this paper, and they are infeasible for large data sets.

The deterministic algorithms are more complex to implement on MapReduce. For instance, it is time-consuming to calculate the leading singular values and vectors of a massively distributed matrix or to cluster their columns using k -means. It is also computationally complex to calculate QR decomposition with pivoting. Moreover, the recently proposed algorithms for volume sampling are more complex than other CSS algorithms as well as the one presented in this paper, and they are infeasible for large data sets.

6.3 Hybrid methods

A third category of CSS techniques is the hybrid methods which combine the benefits of both the randomized and deterministic methods. In these methods, a large subset of columns is randomly sampled from the columns of the data matrix and then a deterministic step is employed to reduce the number of selected columns to the desired rank.

For instance, Boutsidis et al. [6] proposed a two-stage hybrid algorithm for column subset selection which runs in $O(\min(n^2m, nm^2))$. In the first stage, the algorithm samples $c = O(l \log l)$ columns based on probabilities calculated using the l -leading right singular vectors. In the second phase, a Rank-revealing QR (RRQR) algorithm is employed to select exactly l columns from the columns sampled in the first stage. The authors suggested repeating the selection process 40 times in order to provably reduce the failure probability. The authors proved a good theoretical guarantee for the algorithm in terms of spectral and Frobenius term. However, the algorithm depends on calculating the leading l right singular vectors which is computationally complex for large data sets.

The hybrid algorithms for CSS can be easily implemented on MapReduce if the randomized selection step is MapReduce efficient, and the deterministic selection step can be implemented on a single machine. This is usually true if the number of columns selected by the randomized step is relatively small.

6.4 Related work on unsupervised feature selection

The column subset selection problem is a generalized variant of the unsupervised feature selection problem in which the columns of the data matrix represent features. Much research has been done on the unsupervised feature selection problem. This includes, but not limited to, (1) methods that depend on the Principal Component Analysis (PCA) method to search for the most representative features [5, 13, 38, 47], (2) methods that select features that preserve similarities between data instances, such as the Laplacian score (LS) [34] and the spectral feature selection method (a.k.a., SPEC) [59], and (3) methods that directly select features which preserve the cluster structure of the data, such as the Q - α algorithm [57] and the Multi-Cluster Feature Selection (MCFS) [8]. In addition to these categories, there are methods for unsupervised feature selection that optimize PCA-like criteria such as the Convex Principal Feature Selection (CPFS) [49], and methods that depend on clustering the features into groups and selecting representative features from each group, such as the Feature Selection using Feature Similarity (FSFS) [51].

Although a variety of methods for unsupervised feature selection can be used for column subset selection, many of these methods optimize objective functions that are different from that of Problem 1, and some of them assume that there is a cluster structure in the data. This accordingly limits the direct applicability of these methods to the column subset selection problem.

6.5 Comparison with related work

The greedy column subset selection algorithm presented in Sect. 4 belongs to the category of deterministic algorithms. In comparison with QR-based methods, the greedy CSS algorithm can be implicitly used to calculate a Q -less incomplete QR factorization of the data matrix A :

$$A = QW, A\Pi = QW\Pi = QR$$

where Π is a permutation matrix which sorts the first l columns according to their selection order. The permutation of the columns of the embedding matrix W produces an upper triangular matrix.

The greedy CSS algorithm differs from the greedy algorithm proposed by Çivril and Magdon-Ismail [9, 10] in that the latter depends on first calculating the Singular Value Decomposition of the data matrix, which is computationally complex, especially for large matrices. The proposed algorithm is also more efficient than the recently proposed volume sampling algorithms [17, 32].

In comparison with other CSS methods, the distributed algorithm proposed in this paper is designed to be MapReduce efficient. In the selection step, representative columns are selected based on a common representation. The common representation proposed in this work is based on random projection. This is more efficient than the work of Çivril and Magdon-Ismail [10] which selects columns based on the leading singular vectors. In comparison with other deterministic methods, the proposed algorithm is specifically designed to be parallelized, which makes it applicable to big data matrices whose columns are massively distributed. On the other hand, the two-step of distributed then centralized selection is similar to that of the hybrid CSS methods. The proposed algorithm, however, employs a deterministic algorithm at the distributed selection phase which is more accurate than the randomized selection employed by hybrid methods in the first phase.

In connection to the unsupervised feature selection problem, a variant of the greedy algorithm presented in this paper has previously been proposed for the unsupervised feature selection problem [27, 28] where it has shown superior performance to other state-of-the-art methods for unsupervised feature selection. However, the algorithm proposed by Farahat et al. [27, 28] is centralized and it cannot be easily extended to handle big data that are massively distributed across different machines.

7 Experiments

Two sets of experiments have been conducted. The first set of experiments has been conducted on medium-sized data sets in order to evaluate the efficiency and effectiveness of the centralized greedy CSS algorithm in comparison with state-of-the-art methods for CSS. The second set of experiments has been conducted on two big data sets to evaluate the efficiency and effectiveness of the distributed CSS algorithm on MapReduce.

Experiments have been conducted on eight benchmark data sets, whose properties are summarized in Table 1.³ The first six data sets were used to conduct the centralized experiments. The *Reuters-21578* is the training set of the *Reuters-21578* collection [44]. The

³ The data sets *Reuters-21578*, *MNIST-4K*, *PIE-20* and *YaleB-38* are available in MAT format at: <http://www.cad.zju.edu.cn/home/dengcai/Data/data.html>. *PIE-20* is a subset of *PIE-32x32* with the images of the first 20 persons.

Table 1 The properties of the medium and large data sets used to evaluate different CSS methods

Data set	Type	# Instances	# Features
Reuters-21578	Documents	5,946	18,933
Reviews	Documents	4,069	36,746
LA1	Documents	3,204	29,714
MNIST-4K	Digit images	4,000	784
PIE-20	Face images	3,400	1,024
YaleB-38	Face images	2,414	1,024
RCV1-200K	Documents	193,844	47,236
TinyImages-1M	Images	1 Million	1,024

Reviews and *LA1* are document data sets from TREC collections.⁴ The preprocessed versions of *Reviews* and *LA1* that are distributed with the CLUTO Toolkit [41] were used. The *MNIST-4K* is a subset of the MNIST data set of handwritten digits.⁵ The *PIE-20* and *YaleB-38* are preprocessed subsets of the CMU PIE [53] and Extended Yale Face [43] data sets, respectively. The *PIE-20* and *YaleB-38* data sets have been used by He et al. [35] to evaluate different face recognition algorithms. Besides, the distributed experiments were conducted on two data sets. The *RCV1-200K* is a subset of the RCV1 data set [45] which has been prepared and used by Chen et al. [12] to evaluate parallel spectral clustering algorithms. The *TinyImages-1M* data set contains 1 million images that were sampled from the 80 million tiny images data set [55] and converted to grayscale.

Similar to previous work on CSS, the different methods are evaluated according to their ability to minimize the reconstruction error of the data matrix based on the subset of selected columns. In order to quantify the reconstruction error across different data sets, a relative accuracy measure is defined as

$$\text{Relative Accuracy} = \frac{\|A - \tilde{A}_{\mathcal{U}}\|_F - \|A - \tilde{A}_{\mathcal{S}}\|_F}{\|A - \tilde{A}_{\mathcal{U}}\|_F - \|A - \tilde{A}_I\|_F} \times 100\%,$$

where $\tilde{A}_{\mathcal{U}}$ is the rank- l approximation of the data matrix based on a random subset \mathcal{U} of columns, $\tilde{A}_{\mathcal{S}}$ is the rank- l approximation of the data matrix based on the subset \mathcal{S} of columns and \tilde{A}_I is the best rank- l approximation of the data matrix calculated using the Singular Value Decomposition (SVD). This measure compares different methods relative to the uniform sampling as a baseline with higher values indicating better performance.

7.1 Evaluation of centralized greedy CSS

In the medium-scale experiments, the following CSS methods are compared.⁶

- **UniNoRep**: is uniform sampling of columns without replacement.
- **qr**: is the QR decomposition with column pivoting [30] implemented by the MATLAB *qr* function.⁷

⁴ <http://trec.nist.gov>.

⁵ <http://yann.lecun.com/exdb/mnist>.

⁶ The CSS algorithm of Boutsidis et al. [3] was not included in the comparison as its implementation is not available.

⁷ Revision: 5.13.4.7.

- **SRRQR**: is the strong rank-revealing QR decomposition [31]. Algorithm 4 of [31] was implemented in MATLAB. In this implementation, the MATLAB *qr* function is first used to calculate the QR decomposition with column pivoting and then the columns are swapped using the criterion specified by Gu and Eisenstat [31].⁸
- **ApproxSVD**: is the sparse approximation of Singular Value Decomposition (SVD) [9, 10]. The algorithm was implemented in MATLAB. The generalized CSS algorithm is used to select columns that best approximates the leading singular vectors. The use of the generalized CSS algorithm is equivalent to, but more efficient than, the algorithm proposed by Çivril and Magdon-Ismail [9, 10]. Since the calculation of exact SVD is computationally complex, the Stochastic SVD algorithm [33] is used to approximate the leading singular values and vectors of the data matrix. This significantly reduces the run time of the original algorithm proposed by Çivril and Magdon-Ismail while achieving comparable accuracy. In this experiment, the number of leading singular vectors is set to l .
- **HybridCSS**: is the hybrid column subset selection algorithm proposed by Boutsidis et al. [6]. The number of selected columns in the randomized phase is set to $l \log(l)$. The algorithm was implemented in MATLAB. In the randomized phase, the Stochastic SVD is first used to calculate the leading singular vectors, and the approximated singular vectors are then used to calculate the sampling probabilities. In the random phase, the number of leading singular vectors is set to l . In the deterministic phase, the MATLAB *qr* function is used to select columns.⁹
- **GreedyCSS**: is the greedy column subset selection method described in Algorithm 1.
- **RndGreedyCSS**: is the greedy algorithm for the generalized column subset selection in which the target matrix is a random subspace obtained using random projection. Similar to ApproxSVD and HybridCSS, the dimension of the random projection matrix is set to l .

For all the data sets, the percentage of selected columns l/n is changed from 1 to 25 % with increments of 2 % and the relative accuracies and run times are measured.¹⁰ Experiments with randomness were repeated ten times, and the average and standard deviation of measures were calculated.

Figures 1 and 2 show the relative accuracy measures and run times for different CSS methods on the six benchmark data sets.¹¹

In addition, Tables 2 and 3 show the relative accuracy measures for the best performing CSS methods (**HybridCSS**, **ApproxSVD**, **GreedyCSS**, and **RndGreedyCSS**). Each sub-table represents a data set and each column represents a percentage of selected columns. In order to test the statistical significance of the accuracy measures obtained by the proposed methods, the analysis of variance (ANOVA) procedure [36] is conducted for each sub-column.¹² In

⁸ In the implemented code, the efficient recursive formulas in Section 4 of [31] are used to implement the update of QR decomposition and the swapping criterion.

⁹ In [4] (a newer version of [6]), Boutsidis et al. suggested the use of the SRRQR algorithm [31, Algorithm 4] for the deterministic phase. Although the SRRQR algorithm achieves the theoretical guarantee presented in [6], the MATLAB *qr* function is used in the conducted experiments as it is much faster and it achieves comparable accuracy for the experimented data sets.

¹⁰ For the MNIST4K data set, the range of l/n values is smaller since the rank of the matrix is very low (i.e., less than the number of pixels).

¹¹ The **qr** and **SRRQR** methods both depend on the MATLAB *qr* function. For the document data sets, the MATLAB *qr* function takes very long times compared with other methods and accordingly they are not reported in the shown figures.

¹² The MATLAB functions *anova1* and *multcompare* were used.

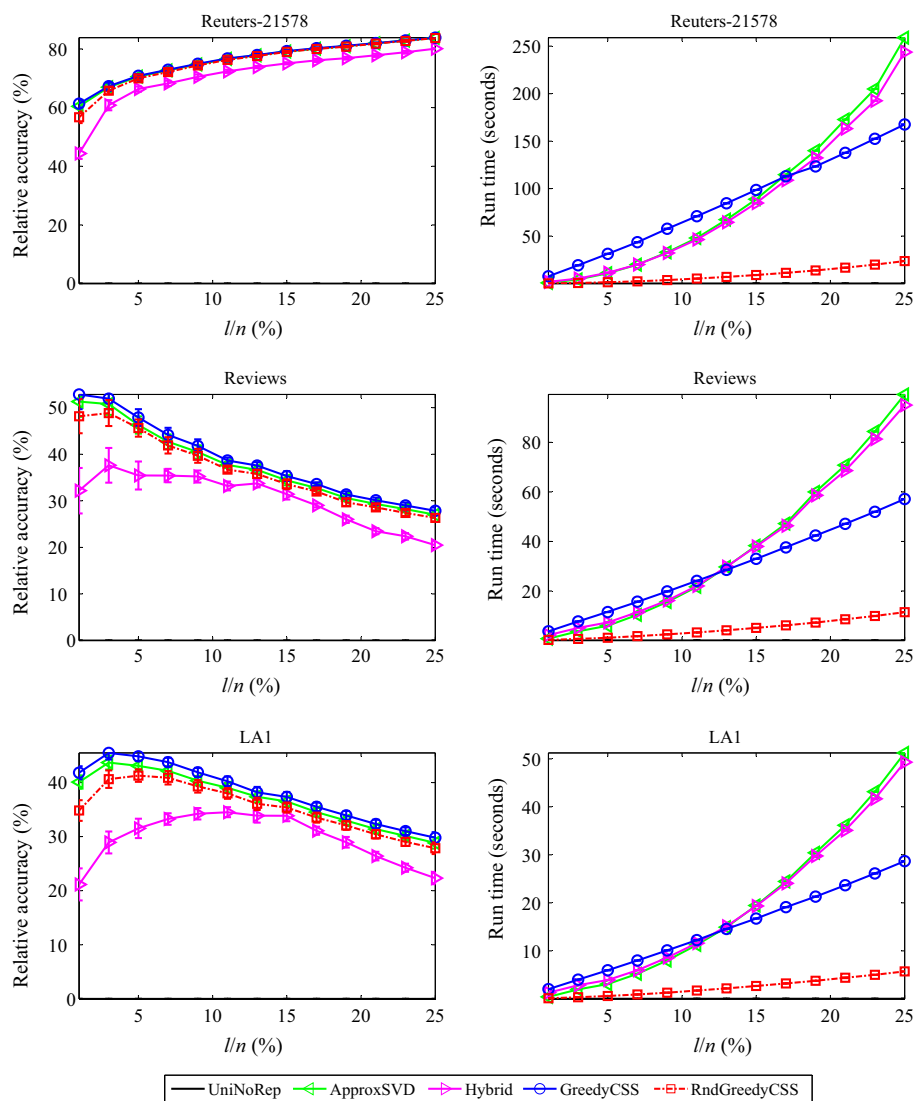


Fig. 1 The relative accuracy measures and run times of different column-based low-rank approximations \tilde{A}_S for the *Reuters-21578*, *Reviews* and *LA1* data sets

specific, the F -test is performed on all the accuracy measures in each sub-column to test the null-hypothesis that all the accuracy measures are drawn from populations with the same mean. The null-hypothesis is rejected if the p value of the test is less than a significance level of 0.05. In this case, the F -test is followed by a post hoc analysis in which Tukey's tests [36] are conducted on pairs of the best performing methods. The F -statistic, p value, and results of the Tukey's tests are reported in the table.

It can be observed from the figures and table that for all data sets, the **GreedyCSS** method significantly outperforms the **UniNoRep**, **qr**, **SRRQR**, and **HybridCSS** methods in terms of relative accuracy, and it shows comparable accuracy to the **ApproxSVD** method for almost

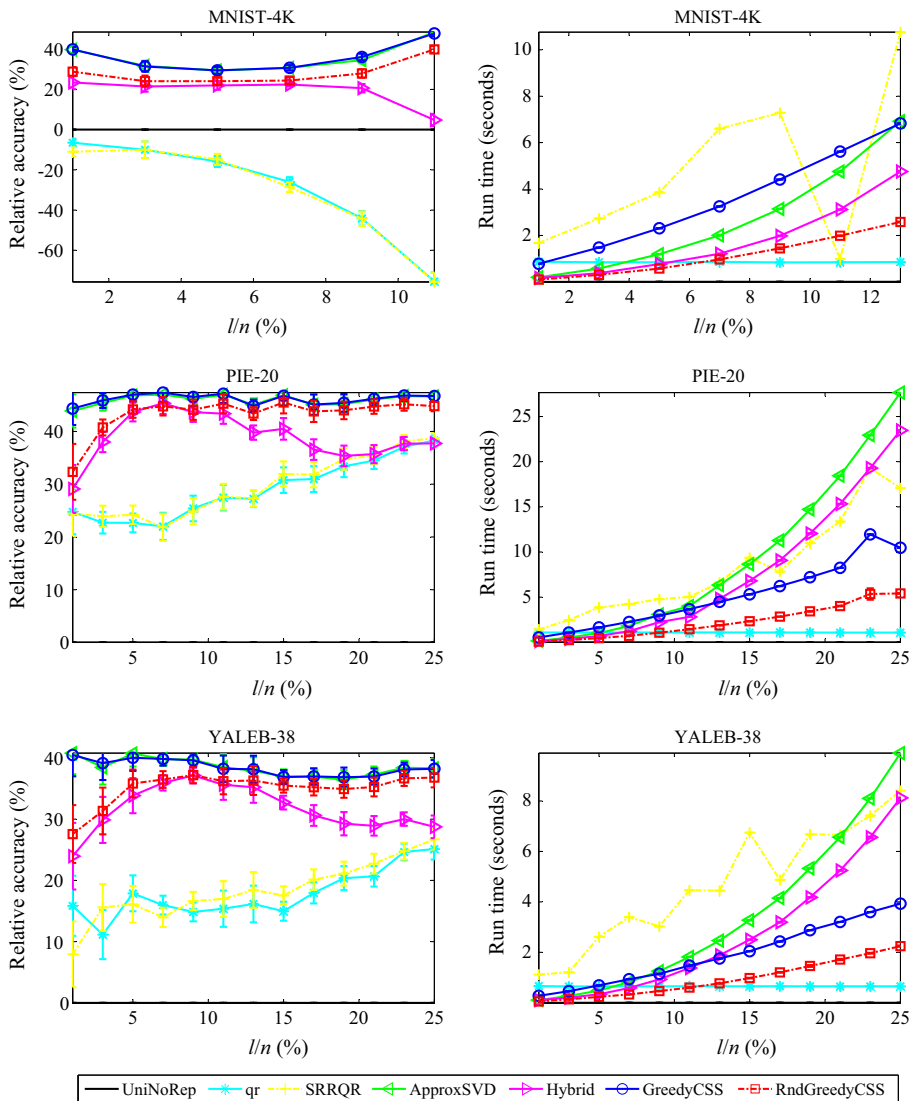


Fig. 2 The relative accuracy measures and run times of different column-based low-rank approximations \tilde{A}_S for the *MNIST-4K*, *PIE-20* and *YaleB-38* data sets

all data sets (where $a = b$ in Tables 2 and 3). In terms of run times, for most of the data sets, the **GreedyCSS** scales better than the **HybridCSS** and **ApproxSVD** methods.

On the other hand, the **RndGreedyCSS** outperforms the **UniNoRep**, **qr**, and **SRRQR** methods in terms of relative accuracy and shows comparable accuracy to the **HybridCSS** method. In terms of run times, the **RndGreedyCSS** is much more efficient than the **HybridCSS** method and other methods for all data sets. It should also be noted that the **SRRQR** method achieves comparable accuracy to the **qr** method and both methods demonstrate lower approximation accuracies than other deterministic and hybrid methods.

Table 2 The relative accuracy of the best performing CSS methods for the *Reuters-21578*, *Reviews* and *LAI* data sets

Method	$l/n = 5\%$	$l/n = 9\%$	$l/n = 13\%$	$l/n = 17\%$
Reuters-21578				
HybridCSS	66,3400.90	70,5200.81	73,8400.39	76,1100.55
ApproxSVD (a)	70,5900.65	74,7600.66	77,8600.32	80,1100.31
GreedyCSS (b)	70,8600.65	74,9600.65	77,9600.32	80,2800.31
RndGreedyCSS (c)	69,9800.61	74,4100.67	77,5800.35	80,0100.34
F-statistic	86.99	90.36	331.1	267.52
p value	$1.50\text{E}-16$	$8.10\text{E}-17$	$2.90\text{E}-26$	$1.20\text{E}-24$
Tukey's test	$a = b, a = c, b > c$	$a = b, a = c, b = c$	$a = b, a = c, b = c$	$a = b, a = c, b = c$
Reviews				
HybridCSS	35,4403.01	35,2201.30	33,7300.84	29,0200.43
ApproxSVD (a)	46,2801.86	40,4801.44	36,6400.76	32,7500.41
GreedyCSS (b)	47,8801.80	41,8001.41	37,5600.75	33,5900.40
RndGreedyCSS (c)	45,6401.86	39,6701.50	35,7900.88	31,9800.52
F-statistic	66.7	40.83	41.08	203.01
p value	$9.00\text{E}-15$	$1.10\text{E}-11$	$1.00\text{E}-11$	$1.30\text{E}-22$
Tukey's test	$a = b, a = c, b = c$	$a = b, a = c, b > c$	$a = b, a = c, b > c$	$a < b, a > c, b > c$
LAI				
HybridCSS	31,5301.79	34,2201.05	33,8601.28	31,0800.75
ApproxSVD (a)	43,0900.88	40,3400.85	37,3501.03	34,5200.79
GreedyCSS (b)	44,8400.85	41,8800.83	38,1801.02	35,5000.78
RndGreedyCSS (c)	41,2801.16	39,3001.15	36,0701.13	33,5200.92
F-statistic	234.34	114.16	28.38	54.23
p value	$1.10\text{E}-23$	$1.9\text{E}-18$	$1.40\text{E}-09$	$2.00\text{E}-13$
Tukey's test	$a < b, a > c, b > c$	$a < b, a = c, b > c$	$a = b, a = c, b > c$	$a < b, a > c, b > c$

For each sub-column, the analysis of variance (ANOVA) procedure is conducted to test the statistical significance of the results. The F-statistic, p value, and results of Tukey's tests between the three best performing methods are reported

7.2 Evaluation of distributed greedy CSS

The distributed CSS method has been compared with different state-of-the-art methods. It should be noted that most of these methods were not designed with the goal of applying them to massively distributed data, and hence, their implementation on MapReduce is not straightforward. However, the designed experiments used the best practices for implementing the different steps of these methods on MapReduce to the best of the authors' knowledge. In specific, the following distributed CSS algorithms were compared.

- **UniNoRep**: is uniform sampling of columns without replacement. This is usually the worst performing method in terms on approximation error and it will be used as a baseline to evaluate methods across different data sets.
- **Hybrid (Uniform)**, **Hybrid (Column Norms)** and **Hybrid (SVD-based)**: are different distributed variants of the hybrid CSS algorithm which can be implemented efficiently on MapReduce. In the randomized phase, the three methods use probabilities calculated based on uniform sampling, column norms and the norms of the leading singular vectors' rows, respectively. The number of selected columns in the randomized phase is set to

Table 3 The relative accuracy of the best performing CSS methods for the *MNIST-4K*, *PIE-20* and *YaleB-38* data sets

MNIST-4K				
Method	$l/n = 1\%$	$l/n = 5\%$	$l/n = 9\%$	$l/n = 13\%$
HybridCSS	23.41 ± 03.33	21.97 ± 02.18	20.60 ± 02.09	-72.06 ± 17.25
ApproxSVD (a)	39.79 ± 01.16	29.36 ± 01.66	34.59 ± 01.63	55.56 ± 04.17
GreedyCSS (b)	39.91 ± 01.16	29.47 ± 01.65	36.09 ± 01.59	56.36 ± 04.09
RndGreedyCSS (c)	28.84 ± 02.25	24.11 ± 01.72	27.90 ± 02.11	51.62 ± 04.07
F-statistic	143.63	43.39	143.88	460.6
p value	$4.40\text{E}-20$	$4.80\text{E}-12$	$4.20\text{E}-20$	$9.30\text{E}-29$
Tukey's test	$a = b, a > c, b > c$	$a = b, a > c, b > c$	$a = b, a > c, b > c$	$a = b, a = c, b = c$
PIE-20				
Method	$l/n = 5\%$	$l/n = 9\%$	$l/n = 13\%$	$l/n = 17\%$
HybridCSS	43.67 ± 01.77	43.65 ± 01.86	39.77 ± 01.37	36.55 ± 01.95
ApproxSVD (a)	46.79 ± 01.22	46.16 ± 01.75	45.07 ± 01.20	45.03 ± 01.95
GreedyCSS (b)	46.98 ± 01.22	46.53 ± 01.74	44.82 ± 01.20	45.08 ± 01.95
RndGreedyCSS (c)	44.16 ± 01.57	44.11 ± 01.92	43.38 ± 01.20	43.79 ± 02.03
F-statistic	13.88	6.31	38.41	43.1
p value	$3.60\text{E}-06$	$1.50\text{E}-03$	$2.60\text{E}-11$	$5.30\text{E}-12$
Tukey's test	$a = b, a > c, b > c$	$a = b, a = c, b > c$	$a = b, a > c, b = c$	$a = b, a = c, b = c$
YALEB-38				
Method	$l/n = 5\%$	$l/n = 9\%$	$l/n = 13\%$	$l/n = 17\%$
HybridCSS	33.83 ± 02.87	37.11 ± 01.28	35.20 ± 02.54	30.60 ± 01.69
ApproxSVD (a)	40.71 ± 02.12	39.70 ± 01.09	37.88 ± 02.22	36.92 ± 01.33
GreedyCSS (b)	40.01 ± 02.14	39.59 ± 01.09	38.14 ± 02.21	36.96 ± 01.33
RndGreedyCSS (c)	35.82 ± 02.25	37.17 ± 01.32	36.29 ± 02.35	35.19 ± 01.34
F-statistic	19.56	14.56	3.53	43.87
p value	$1.10\text{E}-07$	$2.30\text{E}-06$	$2.40\text{E}-02$	$4.10\text{E}-12$
Tukey's test	$a = b, a > c, b > c$	$a = b, a > c, b > c$	$a = b, a = c, b = c$	$a = b, a > c, b > c$

For each sub-column, the analysis of variance (ANOVA) procedure is conducted to test the statistical significance of the results. The F-statistic, p value, and results of Tukey's tests between the three best performing methods are reported

$l \log(l)$. In the deterministic phase, the centralized greedy CSS is employed to select exactly l columns from the randomly sampled columns.

- **DistApproxSVD**: is an extension of the centralized algorithm for sparse approximation of Singular Value Decomposition (SVD) [10]. The distributed CSS algorithm presented in this paper (Algorithm 4) is used to select columns that best approximate the leading singular vectors (by setting $B = U_k \Sigma_k$). The use of the distributed CSS algorithm extends the original algorithm proposed by Civril and Magdon-Ismail [10] to work on distributed matrices. In order to allow efficient implementation on MapReduce, the number of leading singular vectors is set of 100.
- **DistGreedyCSS**: is the distributed column subset selection method described in Algorithm 4. For all experiments, the dimension of the random projection matrix is set to

Table 4 The relative accuracies of different distributed CSS methods

Methods	Relative accuracy (%)		
	$l = 10$	$l = 100$	$l = 500$
RCV1—200K			
UniformNoRep	00.00 ± 10.04	00.00 ± 02.70	00.00 ± 01.55
Hybrid (Uniform)	-09.75 ± 05.00	-01.38 ± 00.99	05.95 ± 01.30
Hybrid (Column Norms)	-03.43 ± 12.91	-04.29 ± 02.14	06.22 ± 01.35
Hybrid (SVD-based)	-09.20 ± 03.12	13.95 ± 03.48	21.33 ± 01.06
DistApproxSVD (a)	30.18 ± 00.03	58.15 ± 00.35	63.61 ± 00.26
DistGreedyCSS (rnd) (b)	29.28 ± 01.24	56.82 ± 00.17	68.93 ± 00.26
DistGreedyCSS (ssgn) (c)	32.09 ± 06.82	57.60 ± 00.66	68.80 ± 00.42
F-statistic	31.62	924.65	3544.63
<i>p</i> value	1.60E-06	4.40E-15	1.40E-18
Tukey's tests	$a = b, a = c, b = c$	$a = b, a = c, b = c$	$a < b, a < c, b = c$
Tiny images—1M			
UniformNoRep	00.00 ± 09.98	00.00 ± 01.35	00.00 ± 01.27
Hybrid (Uniform)	27.71 ± 11.21	02.96 ± 02.16	07.82 ± 01.13
Hybrid (Column Norms)	29.91 ± 10.21	06.54 ± 01.93	05.67 ± 00.22
Hybrid (SVD-based)	17.45 ± 04.75	04.55 ± 02.38	10.20 ± 00.84
DistApproxSVD (a)	66.79 ± 00.00	36.47 ± 00.10	28.80 ± 00.25
DistGreedyCSS (ssgn) (b)	65.65 ± 02.45	33.65 ± 02.62	25.25 ± 00.42
F-statistic	30.82	200.3	746.21
<i>p</i> value	1.30E-05	1.70E-09	2.50E-12
Tukey's tests	$a = b$	$a = b$	$a > b$

For each sub-column, the analysis of variance (ANOVA) procedure is conducted to test the statistical significance of the results. The F-statistic, *p* value, and results of Tukey's tests between the best performing methods are reported

100. This makes the size of the concise representation the same as the DistApproxSVD method. Two types of random matrices are used for random projection: (1) a dense Gaussian random matrix (rnd), and (2) a sparse random sign matrix (ssgn).

Experiments with randomness were repeated three times, and the average and standard deviation of measures were calculated. For the methods that require the calculations of Singular Value Decomposition (SVD), the Stochastic SVD (SSVD) algorithm [33] is used to approximate the leading singular values and vectors of the data matrix. The use of SSVD significantly reduces the run time of the original SVD-based algorithms while achieving comparable accuracy. In the conducted experiments, the SSVD implementation of Mahout was used.

The distributed experiments were conducted on Amazon EC2¹³ clusters, which consist of 10 instances for the *RCV1-200K* data set and 20 instances for the *TinyImages-1M* data set. Each instance has a 7.5 GB of memory and a two-cores processor. All instances are running Debian 6.0.5 and Hadoop version 1.0.3. The data sets were converted into a binary format in the form of a sequence of *key-value* pairs. Each pair consisted of a column index as the key

¹³ Amazon Elastic Compute Cloud (EC2): <http://aws.amazon.com/ec2>.

Table 5 The run times of different distributed CSS methods

Methods	Run time (minutes)		
	$l = 10$	$l = 100$	$l = 500$
RCV1—200K			
UniformNoRep	00.53 ± 00.03	00.55 ± 00.02	00.55 ± 00.03
Hybrid (Uniform)	00.69 ± 00.01	00.74 ± 00.05	02.69 ± 00.08
Hybrid (Column Norms)	01.42 ± 00.03	01.44 ± 00.05	03.46 ± 00.05
Hybrid (SVD-based)	11.41 ± 00.04	11.38 ± 00.03	13.39 ± 00.03
DistApproxSVD	12.76 ± 00.05	12.90 ± 00.02	15.32 ± 00.06
DistGreedyCSS (rnd)	04.51 ± 00.08	04.82 ± 00.06	07.09 ± 00.02
DistGreedyCSS (ssgn)	02.85 ± 00.00	03.17 ± 00.02	05.47 ± 00.14
Tiny images—1M			
UniformNoRep	01.06 ± 00.02	00.98 ± 00.01	01.04 ± 00.04
Hybrid (Uniform)	01.19 ± 00.01	01.30 ± 00.06	06.31 ± 00.06
Hybrid (Column Norms)	02.51 ± 00.03	02.59 ± 00.01	07.48 ± 00.36
Hybrid (SVD-based)	43.09 ± 00.08	43.21 ± 00.03	47.91 ± 00.44
DistApproxSVD	60.75 ± 00.36	62.13 ± 00.86	70.31 ± 00.68
DistGreedyCSS (ssgn)	20.51 ± 00.14	21.86 ± 00.48	28.76 ± 00.81

and a vector of the column entries. That is the standard format used in Mahout¹⁴ for storing distributed matrices.

Tables 4 and 5 show the average relative accuracies and run times for distributed CSS methods, respectively. In Table 4, the analysis of variance (ANOVA) procedure was conducted on the accuracy measures in each sub-column. The statistical significance tests were performed as explained in Sect. 7.1.

It can be observed from the table that for the *RCV1-200K* data set, the **DistGreedyCSS** methods (with random Gaussian and sparse random sign matrices) outperforms uniform sampling and hybrid methods in terms of relative accuracies. In addition, the run times of both of **DistGreedyCSS** methods are relatively small compared with the **DistApproxSVD** method which achieves accuracies that are statistically equivalent to the **DistGreedyCSS** methods for $l = 10$ and $l = 100$ (where $a = b, a = c$ in the *RCV1-200K* sub-table of Table 4). Both the **DistApproxSVD** and **DistGreedyCSS** methods achieve very good approximation accuracies compared with randomized and hybrid methods. It should also be noted that using a sparse random sign matrix for random projection takes much less time than a dense Gaussian matrix, while achieving comparable approximation accuracies (where $b = c$ in the *RCV1-200K* sub-table of Table 4). Based on this observation, the sparse random matrix has been used with the *TinyImages-1M* data set.

For the *TinyImages-1M* data set, although the **DistApproxSVD** achieves slightly higher approximation accuracies than **DistGreedyCSS** (with sparse random sign matrix), they are statistically equivalent for $l = 10$ and $l = 100$ (where $a = b$ in the *TinyImages-1M* sub-table of Table 4) and the **DistGreedyCSS** selects columns in almost one-third of the time. The reason why the **DistApproxSVD** outperforms **DistGreedyCSS** for this data set is that its rank is relatively small (less than 1024). This means that using the leading 100 singular values to represent the concise representation of the data matrix captures most of the information in

¹⁴ Mahout is an Apache project for implementing Machine Learning algorithms on Hadoop. See <http://mahout.apache.org/>.

the matrix and accordingly is more accurate than random projection. The **DistGreedyCSS**, however, still selects a very good subset of columns in a relatively small time.

8 Conclusion

This paper proposes a novel algorithm, which greedily selects a subset of columns from a data matrix such that reconstruction error of the data matrix is minimized. The algorithm depends on a novel recursive formula for the reconstruction error of the data matrix, which allows a greedy selection criterion to be calculated efficiently at each iteration. This paper also presents an accurate and efficient MapReduce algorithm for selecting a subset of columns from a massively distributed matrix. The algorithm starts by learning a concise representation of the data matrix using random projection. It then selects columns from each sub-matrix that best approximate this concise approximation. A centralized selection step is then performed on the columns selected from different sub-matrices. In order to facilitate the implementation of the proposed method, a novel algorithm for greedy generalized CSS is proposed to perform the selection from different sub-matrices. In addition, the different steps of the algorithms are carefully designed to be MapReduce efficient. Experiments on medium and big data sets show that the proposed greedy CSS method significantly outperforms uniform sampling, QR-based sampling and the hybrid CSS methods in terms of relative accuracy, and it achieves comparable accuracy to that of the approximate SVD sampling. In terms of run times, the centralized and distributed greedy CSS methods are more efficient than the SVD-based hybrid method and the approximate SVD sampling.

References

1. Achlioptas D (2003) Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J Comput Syst Sci* 66(4):671–687
2. Bischof C, Quintana-Ortí G (1998) Computing rank-revealing QR factorizations of dense matrices. *ACM Trans Math Softw* 24(2):226–253
3. Boutsidis C, Drineas P, Magdon-Ismael M (2011) Near optimal column-based matrix reconstruction. In: *Proceedings of the 52nd annual IEEE symposium on foundations of computer science (FOCS'11)*, pp 305–314
4. Boutsidis C, Mahoney MW, Drineas P (2008a) An improved approximation algorithm for the column subset selection problem, *CoRR abs/0812.4293*
5. Boutsidis C, Mahoney MW, Drineas P (2008b) Unsupervised feature selection for principal components analysis. In: Li Y, Liu B, Sarawagi S (eds) *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'08)*. ACM, New York, pp 61–69
6. Boutsidis C, Mahoney MW, Drineas P (2009) An improved approximation algorithm for the column subset selection problem. In: *Proceedings of the 20th annual ACM-SIAM symposium on discrete algorithms (SODA'09)*, pp 968–977
7. Boutsidis C, Sun J, Auerousis N (2008) Clustered subset selection and its applications on it service metrics. In: *Proceedings of the 17th ACM conference on information and knowledge management (CIKM'08)*, pp 599–608
8. Cai D, Zhang C, He X (2010) Unsupervised feature selection for multi-cluster data. In: *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'10)*. ACM, New York, NY, pp 333–342
9. Çivril A, Magdon-Ismael M (2008) Deterministic sparse column based matrix reconstruction via greedy approximation of SVD. In: *Proceedings of the 19th international symposium on algorithms and computation (ISAAC'08)*. Springer, New York, pp 414–423
10. Çivril A, Magdon-Ismael M (2012) Column subset selection via sparse approximation of SVD. *Theoret Comput Sci* 421:1–14
11. Chan T (1987) Rank revealing QR factorizations. *Linear Algebra Appl* 88:67–82

12. Chen W-Y, Song Y, Bai H, Lin C-J, Chang E (2011) Parallel spectral clustering in distributed systems. *IEEE Trans Pattern Anal Mach Intell* 33(3):568–586
13. Cui Y, Dy J (2008) Orthogonal principal feature selection, the sparse optimization and variable selection workshop at the international conference on machine learning (ICML)
14. Dasgupta S, Gupta A (2003) An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct Algorithms* 22(1):60–65
15. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. *Commun ACM* 51(1):107–113
16. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. *J Am Soc Inform Sci Technol* 41(6):391–407
17. Deshpande A, Rademacher L (2010) Efficient volume sampling for row/column subset selection. In: *Proceedings of the 51st annual IEEE symposium on foundations of computer science (FOCS'10)*, pp 329–338
18. Deshpande A, Rademacher L, Vempala S, Wang G (2006a) Matrix approximation and projective clustering via volume sampling. *Theory Comput* 2(1):225–247
19. Deshpande A, Rademacher L, Vempala S, Wang G (2006b) Matrix approximation and projective clustering via volume sampling. In: *Proceedings of the 17th annual ACM-SIAM symposium on discrete algorithms (SODA'06)*. ACM, New York, NY, pp 1117–1126
20. Drineas P, Frieze A, Kannan R, Vempala S, Vinay V (2004) Clustering large graphs via the singular value decomposition. *Mach Learn* 56(1–3):9–33
21. Drineas P, Kannan R, Mahoney M (2007) Fast Monte Carlo algorithms for matrices II: computing a low-rank approximation to a matrix. *SIAM J Comput* 36(1):158–183
22. Drineas P, Mahoney M, Muthukrishnan S (2006) Subspace sampling and relative-error matrix approximation: column-based methods. Approximation, randomization, and combinatorial optimization. Algorithms and techniques. Springer, Berlin, pp 316–326
23. Elgohary A, Farahat AK, Kamel MS, Karray F (2013) Embed and conquer: scalable embeddings for kernel k-means on mapreduce. *CoRR abs/1311.2334*
24. Elsayed T, Lin J, Oard DW (2008) Pairwise document similarity in large collections with MapReduce. In: *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: short Papers (HLT'08)*, pp 265–268
25. Ene A, Im S, Moseley B (2011) Fast clustering using MapReduce. In: *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'11)*, pp 681–689
26. Farahat A, Elgohary A, Ghodsi A, Kamel M (2013) Distributed column subset selection on MapReduce. In: *Proceedings of the 13th IEEE international conference on data mining (ICDM'13)*, pp 171–180
27. Farahat AK, Ghodsi A, Kamel MS (2011) An efficient greedy method for unsupervised feature selection. In: *Proceedings of the 11th IEEE international conference on data mining (ICDM'11)*, pp 161–170
28. Farahat AK, Ghodsi A, Kamel MS (2013) Efficient greedy feature selection for unsupervised learning. *Knowl Inf Syst* 35(2):285–310
29. Frieze A, Kannan R, Vempala S (1998) Fast Monte-Carlo algorithms for finding low-rank approximations. In: *Proceedings of the 39th annual IEEE symposium on foundations of computer science (FOCS'98)*, pp 370–378
30. Golub G, Van Loan C (1996) *Matrix computations*, 3rd edn. Johns Hopkins University Press, Baltimore
31. Gu M, Eisenstat SC (1996) Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J Sci Comput* 17(4):848–869
32. Guruswami V, Sinop AK (2012) Optimal column-based low-rank matrix reconstruction. In: *Proceedings of the 21st annual ACM-SIAM symposium on discrete algorithms (SODA'12)*, pp 1207–1214
33. Halko N, Martinsson P-G, Shkolnisky Y, Tygert M (2011) An algorithm for the principal component analysis of large data sets. *SIAM J Sci Comput* 33(5):2580–2594
34. He X, Cai D, Niyogi P (2005) Laplacian score for feature selection, advances in neural information processing systems 18 (NIPS'05). MIT Press, Cambridge, MA
35. He X, Yan S, Hu Y, Niyogi P, Zhang H (2005) Face recognition using Laplacianfaces. *IEEE Trans Pattern Anal Mach Intell* 27(3):328–340
36. Hogg RV, Ledolter J (1987) *Engineering statistics*, vol 358. MacMillan, New York
37. Jain AK, Dubes RC (1988) *Algorithms for clustering data*. Prentice-Hall Inc, Upper Saddle River, NJ
38. Jolliffe I (2002) *Principal component analysis*, 2nd edn. Springer, New York
39. Kang U, Tsourakakis C, Appel A, Faloutsos C, Leskovec J (2008) Hadi: fast diameter estimation and mining in massive graphs with hadoop, CMU-ML-08-117
40. Karloff H, Suri S, Vassilvitskii S (2010) A model of computation for MapReduce. In: *Proceedings of the 21st annual ACM-SIAM symposium on discrete algorithms (SODA'10)*, pp 938–948

41. Karypis G (2003) CLUTO—a clustering toolkit, technical report #02-017. University of Minnesota, Department of Computer Science
42. Kaufman L, Rousseeuw P (1987) Clustering by means of medoids. Department of Mathematics and Informatics, Technical report, Technische Hogeschool, Delft (Netherlands)
43. Lee K, Ho J, Kriegman D (2005) Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans Pattern Anal Mach Intell* 27(5):684–698
44. Lewis D (1999) Reuters-21578 text categorization test collection distribution 1.0
45. Lewis DD, Yang Y, Rose TG, Li F (2004) Rcv1: a new benchmark collection for text categorization research. *J Mach Learn Res* 5:361–397
46. Li P, Hastie TJ, Church KW (2006) Very sparse random projections. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'06)*, pp 287–296
47. Lu Y, Cohen I, Zhou X, Tian Q (2007) Feature selection using principal feature analysis. In: *Proceedings of the 15th international conference on multimedia*. ACM, New York, NY, pp 301–304
48. Lütkepohl H (1996) *Handbook of matrices*. Wiley, New Jersey
49. Masaeli M, Yan Y, Cui Y, Fung, G, Dy J (2010) Convex principal feature selection. In: *Proceedings of SIAM international conference on data mining (SDM)*, pp 619–628
50. Meng X, Mahoney M (2013) Robust regression on mapreduce. In: *Proceedings of the 30th international conference on machine learning (ICML-13)*, pp 888–896
51. Mitra P, Murthy C, Pal S (2002) Unsupervised feature selection using feature similarity. *IEEE Trans Pattern Anal Mach Intell* 24(3):301–312
52. Pan C (2000) On the existence and computation of rank-revealing LU factorizations. *Linear Algebra Appl* 316(1):199–222
53. Sim T, Baker S, Bsat M (2003) The CMU pose, illumination, and expression database. *IEEE Trans Pattern Anal Mach Intell* 25(12):1615–1618
54. Singh S, Kubica J, Larsen S, Sorokina D (2009) Parallel large scale feature selection for logistic regression. *Proceedings of the SIAM international conference on data mining*, pp 1171–1182
55. Torralba A, Fergus R, Freeman W (2008) 80 Million tiny images: a large data set for nonparametric object and scene recognition. *IEEE Trans Pattern Anal Mach Intell* 30(11):1958–1970
56. White T (2009) *Hadoop: the definitive guide*, 1st edn. O'Reilly Media Inc, Sebastopol
57. Wolf L, Shashua A (2005) Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach. *J Mach Learn Res* 6:1855–1887
58. Xiang J, Guo C, Aboulmaga A (2013) Scalable maximum clique computation using mapreduce. *IEEE 29th international conference on data engineering (ICDE)*, 2013, pp 74–85
59. Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning. In: *Proceedings of the 24th international conference on machine learning (ICML'07)*. ACM, New York, NY, pp 1151–1157



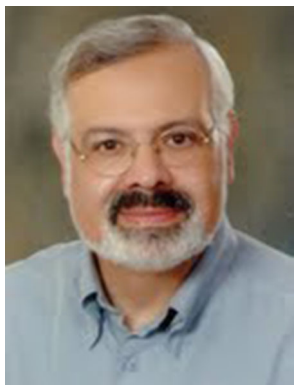
Ahmed K. Farahat holds a Ph.D. degree from the University of Waterloo and M.Sc. and B.Sc. degrees from Cairo University in Egypt, all in Computer Engineering. Since January 2013, Dr. Farahat has been a postdoctoral fellow at the University of Waterloo. He also worked as a research assistant at the University of Waterloo (2007–2012) and a research engineer at IBM Egypt (2005–2007). His research interests lie in the areas of machine learning and data mining, and their applications to big data analytics. Dr. Farahat is the recipient of the Best Paper Award Runner-Up at the 2013 IEEE International Conference on Data Mining (ICDM), and two best paper awards in specialized workshops at ICDM'09 and NIPS'13.



Ahmed Elgohary received a B.Sc. degree in Computer Engineering at Alexandria University, Egypt in 2010. He joined the University of Waterloo in 2012 where he received an M.A.Sc. degree in Electrical and Computer Engineering in 2014. In his M.A.Sc. thesis, Ahmed developed efficient algorithms for scaling kernel-based clustering on MapReduce. Ahmed is the recipient of the Best Paper Award Runner-Up at the 2013 IEEE International Conference on Data Mining (ICDM). He is currently pursuing a Ph.D. in Computer Science at the University of Maryland, College Park. Ahmed is broadly interested in data mining and large-scale analytics.



Ali Ghodsi is an Associate Professor in the Department of Statistics and Actuarial Science at the University of Waterloo. He is also a member of the Centre for Computational Mathematics in Industry and Commerce and the Artificial Intelligence Research Group at the University of Waterloo. He was a researcher at the University of Toronto with the Probabilistic and Statistical Inference Group and at the University of Alberta at the Alberta Ingenuity Centre for Machine Learning. Ghodsi's research spans a variety of areas in computational statistics. He studies theoretical frameworks and develops new machine learning algorithms for analyzing large-scale data sets, with applications to bioinformatics, data mining, pattern recognition, robotics, computer vision, and sequential decision making.



Mohamed S. Kamel received the B.Sc. (Hons) EE (Alexandria University), M.A.Sc. (McMaster University), and Ph.D. (University of Toronto). He joined the University of Waterloo, Canada, in 1985, where he is a Professor and Director of the Center for Pattern Analysis and Machine Intelligence at the Department of Electrical and Computer Engineering. He currently holds the position of University Research Chair. His research interests are in computational intelligence, pattern recognition, machine learning and cooperative intelligent systems. He has authored and coauthored over 500 papers in journals and conference proceedings, 13 edited volumes, 16 chapters in edited books, four patents, and numerous technical and industrial project reports. Under his supervision, 91 Ph.D. and M.A.Sc. students have completed their degrees. Dr. Kamel is member of ACM, PEO, Life Fellow of IEEE, Fellow of the Engineering Institute of Canada (EIC), Fellow of the Canadian Academy of Engineering (CAE), and Fellow of the International Association of Pattern Recognition (IAPR).