OPT2025: 17th Annual Workshop on Optimization for Machine Learning

A Theoretical Analysis for CUR Decomposition based Active Learning and Feature Selection

Zhong Chen ZHONG.CHEN@SIU.EDU
School of Computing, Southern Illinois University, Carbondale, IL 62901, United States

Chen Zhao CHEN_ZHAO@BAYLOR.EDU
Department of Computer Science, Baylor University, Waco, TX 76798, United States

Yi He YIHE@WM.EDU
Department of Data Science, William & Mary, Williamsburg, VA 23185, United States

Abstract

This study comprehensively investigated CUR decomposition-based active learning and feature selection from an optimization perspective, especially provided the complexity analysis.

1. Introduction

The primary goal of active learning [4, 12, 15, 21, 24] and feature selection [6, 9, 11, 13, 25, 26, 28, 30] is to address a fundamental challenge in machine learning and online learning: the high cost of labeling data and the curse of dimensionality. Modern datasets often contain a vast number of both instances (rows) and features (columns), making full annotation prohibitively expensive and models computationally heavy and prone to overfitting. The paper introduces a unified solution that tackles both problems simultaneously by leveraging the principles of matrix decomposition, specifically the CUR matrix decomposition [3, 5, 12, 14, 23], to select the most informative data points to label (active learning) and the most relevant features (feature selection).

CUR decomposition is chosen over other techniques such as PCA [22] or SVD [1] because it provides a low-rank approximation of the original data matrix using actual rows and columns from the dataset itself. This crucial property ensures that the selected features and instances are interpretable and maintain their original meaning, as opposed to the transformed, abstract components produced by PCA. The method works by minimizing the residual of CUR with sparse row and column constraints, which identify the rows (feature) and columns (instances) that exert the most influence on the structure of the data matrix. These high-leverage points are deemed the most representative and informative.

By applying CUR decomposition, the framework efficiently selects a small subset of instances for an expert to label, drastically reducing annotation costs in active learning. Concurrently, it selects a subset of features that best capture the data's variance, improving model efficiency and generalization. This creates a powerful synergy: the model is trained on maximally informative data points described by the most relevant features, leading to a more robust, interpretable, and computationally efficient machine learning pipeline compared to handling each problem separately.

2. Active Learning and Feature Selection via CUR Decomposition

From an algorithmic perspective, the matrices $\boldsymbol{C}$, $\boldsymbol{U}$, and $\boldsymbol{R}$ can be obtained by minimizing the approximation error $\|\boldsymbol{W}-\boldsymbol{CUR}\|_F^2$. Here we make a key observation that the above definition is closely related to the problem of simultaneous sample and feature selection. More specifically, the matrix $\boldsymbol{UR}$ can be regarded as a reconstruction coefficient matrix, and $\boldsymbol{C}$ denotes the selected m samples, thus minimizing means that the total reconstruction error $\|\boldsymbol{W}-\boldsymbol{CUR}\|_F^2$ is minimized, which can make the data points in $\boldsymbol{C}$ be the most representative. The reconstruction coefficients $\boldsymbol{UR}$ are related to an r-dimensional feature subset of the dataset. The reconstruction coefficients of each reconstructed data point $\boldsymbol{w}_i$ $(i=1,2,\cdots,n)$ are formed by a linear combination of its r features. In the meantime, the matrix $\boldsymbol{CU}$ can also be regarded as a reconstruction coefficient matrix, and $\boldsymbol{R}$ is the new low-dimensional representation of $\boldsymbol{W}$, so minimizing $\|\boldsymbol{W}-\boldsymbol{CUR}\|_F^2$ also indicates that the selected r features can represent the whole dataset most precisely. The construction of the coefficient matrix $\boldsymbol{CU}$ depends on a sample subset of $\boldsymbol{W}$. Clearly, active learning and feature selection can be conducted simultaneously in such a joint framework via CUR factorization.

Let $\boldsymbol{p}=[p_1,p_2,\cdots,p_n]^T\in\{0,1\}^n$ and $\boldsymbol{q}=[q_1,q_2,\cdots,q_d]^T\in\{0,1\}^d$ denote two indicator variables to represent whether a sample and a feature is selected or not, respectively. Specifically, $p_i=1$ (or 0) $(i=1,2,\cdots,n)$ indicates that the i-th sample is selected (or not), and $q_j=1$ (or 0) $(j=1,2,\cdots,d)$ means that the j-th feature is selected (or not). Then, minimizing $\|\boldsymbol{W}-\boldsymbol{CUR}\|_F^2$ can be rewritten as

$$\begin{aligned}\min_{\boldsymbol{p}\in\{0,1\}^n,\boldsymbol{V}\in\mathbb{R}^{n\times d},\boldsymbol{q}\in\{0,1\}^d}&\|\boldsymbol{W}-\boldsymbol{W}\operatorname{diag}(\boldsymbol{p})\boldsymbol{V}\operatorname{diag}(\boldsymbol{q})\boldsymbol{W}\|_F^2\\ s.t.\ \ &\mathbf{1}_n^T\boldsymbol{p}=m,\boldsymbol{p}\in\{0,1\}^n\\ &\mathbf{1}_d^T\boldsymbol{q}=r,\boldsymbol{q}\in\{0,1\}^d\end{aligned}\tag{1}$$

where $\boldsymbol{V}\in\mathbb{R}^{n\times d}$, $\operatorname{diag}(\boldsymbol{p})=\operatorname{diag}\{p_1,p_2,\cdots,p_n\}\in\mathbb{R}^{n\times n}$ is a diagonal matrix with $\boldsymbol{p})$ on its diagonal, $\operatorname{diag}(\boldsymbol{q})=\operatorname{diag}\{q_1,q_2,\cdots,q_d\}\in\mathbb{R}^{d\times d}$ is a diagonal matrix with $\boldsymbol{q})$ on its diagonal, $\mathbf{1}_n=[1,1,\cdots,1]^T\in\mathbb{R}^n$ is an n-dimensional vector with all components being 1, and $\mathbf{1}_d=[1,1,\cdots,1]^T\in\mathbb{R}^d$ is a d-dimensional vector with all components being 1. The term $\boldsymbol{W}\operatorname{diag}(\boldsymbol{p})$ in Eq. (1) aims to make m columns of $\boldsymbol{W}$ unchanged, and resets the rest $(n-m)$ columns to zero vectors. While the term $\operatorname{diag}(\boldsymbol{q})\boldsymbol{W}$ in Eq. (1) tends to keep r rows of $\boldsymbol{W}$ unchanged, and resets the rest $(d-r)$ row to zero vectors.

3. A Convex Formulation

The objective function (1) is hard to be solved directly, since it is an NP-hard problem. After a careful observation to (1), we find that we can utilize the $\ell_{2,0}$ mixed norm of a matrix to reduce the number of the parameters. Defining $\boldsymbol{X}=\operatorname{diag}(\boldsymbol{p})\boldsymbol{V}\operatorname{diag}(\boldsymbol{q})\in\mathbb{R}^{n\times d}$, we can rewritten Eq. (1) as

$$\begin{aligned}\min_{\boldsymbol{X}\in\mathbb{R}^{n\times d}}&\|\boldsymbol{W}-\boldsymbol{WXW}\|_F^2\\ s.t.\ \ &\|\boldsymbol{X}\|_{2,0}=m\\ &\|\boldsymbol{X}^T\|_{2,0}=r\end{aligned}\tag{2}$$

where the quasi-norm $\ell_{2,0}$ norm of a matrix $\boldsymbol{X}$ is defined as the number of the non-zero rows of $\boldsymbol{X}$, denoted by $\|\boldsymbol{X}\|_{2,0}$. Based on Eq. (2), we propose to optimize the following objective function:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times d}} \|\boldsymbol{W}-\boldsymbol{W}\boldsymbol{X}\boldsymbol{W}\|_F^2+\alpha\|\boldsymbol{X}\|_{2,0}+\beta\|\boldsymbol{X}^T\|_{2,0} \tag{3}$$

where $\alpha>0$ and $\beta>0$ are two regularization parameters.

However, Eq. (3) is still an NP-hard problem due to the $\ell_{2,0}$ mixed norm of a matrix. Fortunately, there exists theoretical progress that $\|\boldsymbol{X}\|_{2,1}$ is the minimum convex hull of $\|\boldsymbol{X}\|_{2,0}$. The result of minimizing $\|\boldsymbol{X}\|_{2,1}$ is the same as that of minimizing $\|\boldsymbol{X}\|_{2,0}$, as long as $\boldsymbol{X}$ is row-sparse enough. Therefore, Eq. (3) can be relaxed to the following convex optimization problem:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times d}} \|\boldsymbol{W}-\boldsymbol{W}\boldsymbol{X}\boldsymbol{W}\|_F^2+\alpha\|\boldsymbol{X}\|_{2,1}+\beta\|\boldsymbol{X}^T\|_{2,1} \tag{4}$$

where $\|\boldsymbol{X}\|_{2,1}=\sum_{i=1}^n\sqrt{\sum_{j=1}^d x_{i,j}^2}=\sum_{i=1}^n\|\boldsymbol{x}_i\|_2$, $x_{i,j}$ is the i-th row j-th column element of $\boldsymbol{X}$, $\boldsymbol{x}_i\in\mathbb{R}^d$ is the i-th row of $\boldsymbol{X}$.

4. Optimization Algorithm

To solve Eq. (4), we first introduce two variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$, to convert to the following equivalent objective function:

$$\begin{aligned}\min_{\boldsymbol{X}\in\mathbb{R}^{n\times d},\boldsymbol{Y}\in\mathbb{R}^{n\times d},\boldsymbol{Z}\in\mathbb{R}^{d\times n}} & \|\boldsymbol{W}-\boldsymbol{W}\boldsymbol{X}\boldsymbol{W}\|_F^2+\alpha\|\boldsymbol{Y}\|_{2,1}+\beta\|\boldsymbol{Z}\|_{2,1}\\ s.t.\ \ & \boldsymbol{Y}=\boldsymbol{X}\\ & \boldsymbol{Z}=\boldsymbol{X}^T\end{aligned} \tag{5}$$

The augmented Lagrange function of Eq. (5) is:

$$\begin{aligned}\mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2)=&\|\boldsymbol{W}-\boldsymbol{W}\boldsymbol{X}\boldsymbol{W}\|_F^2+\alpha\|\boldsymbol{Y}\|_{2,1}+\beta\|\boldsymbol{Z}\|_{2,1}\\ &+\langle\boldsymbol{\Lambda}_1,\boldsymbol{X}-\boldsymbol{Y}\rangle+\langle\boldsymbol{\Lambda}_2,\boldsymbol{X}^T-\boldsymbol{Z}\rangle\\ &+\frac{\rho_1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2+\frac{\rho_2}{2}\|\boldsymbol{X}^T-\boldsymbol{Z}\|_F^2\end{aligned} \tag{6}$$

where $\boldsymbol{\Lambda}_1\in\mathbb{R}^{n\times d}$ and $\boldsymbol{\Lambda}_2\in\mathbb{R}^{d\times n}$ are the Lagrange multipliers, $\langle\boldsymbol{\Lambda}_1,\boldsymbol{X}-\boldsymbol{Y}\rangle=Trace(\boldsymbol{\Lambda}_1^T(\boldsymbol{X}-\boldsymbol{Y}))$, $Trace(\cdot)$ is the trace norm of a square matrix, which is the sum of diagonal entries of a square matrix, ρ_1 and ρ_2 are the constraint violation penalty parameters. From the augmented Lagrangian function, we find that the subproblems about $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are fully separable, as a result we can introduce the classical two-block ADMM [16], while considering $\boldsymbol{X}$ and $(\boldsymbol{Y},\boldsymbol{Z})$ as two-block variables. In an ADMM-type algorithm [17], the basic Gauss-Seidel structure [2] in $(t+1)$-th iteration is

$$\begin{aligned}\boldsymbol{X}^{(t+1)}&=\text{argmin}_{\boldsymbol{X}\in\mathbb{R}^{n\times d}}\,\mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X},\boldsymbol{Y}^{(t)},\boldsymbol{Z}^{(t)},\boldsymbol{\Lambda}_1^{(t)},\boldsymbol{\Lambda}_2^{(t)})\\ \boldsymbol{Y}^{(t+1)}&=\text{argmin}_{\boldsymbol{Y}\in\mathbb{R}^{n\times d}}\,\mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X}^{(t+1)},\boldsymbol{Y},\boldsymbol{Z}^{(t)},\boldsymbol{\Lambda}_1^{(t)},\boldsymbol{\Lambda}_2^{(t)})\\ \boldsymbol{Z}^{(t+1)}&=\text{argmin}_{\boldsymbol{Z}\in\mathbb{R}^{d\times n}}\,\mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X}^{(t+1)},\boldsymbol{Y}^{(t+1)},\boldsymbol{Z},\boldsymbol{\Lambda}_1^{(t)},\boldsymbol{\Lambda}_2^{(t)})\\ \boldsymbol{\Lambda}_1^{(t+1)}&=\boldsymbol{\Lambda}_1^{(t)}+\rho_1(\boldsymbol{X}^{(t+1)}-\boldsymbol{Y}^{(t+1)})\\ \boldsymbol{\Lambda}_2^{(t+1)}&=\boldsymbol{\Lambda}_2^{(t)}+\rho_2((\boldsymbol{X}^{(t+1)})^T-\boldsymbol{Z}^{(t+1)})\end{aligned} \tag{7}$$

Next, we will introduce how to solve these subproblems in detail.

(1) Compute the subproblem about $\boldsymbol{X}^{(t+1)}$:

When the other variables are fixed with the former iteration result $(\boldsymbol{Y}^{(t)}, \boldsymbol{Z}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)})$, the subproblem about $\boldsymbol{X}^{(t+1)}$ is as

$$\begin{aligned}\boldsymbol{X}^{(t+1)} &= \operatorname{argmin}_{\boldsymbol{X}\in\mathbb{R}^{n\times d}} \mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X}, \boldsymbol{Y}^{(t)}, \boldsymbol{Z}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)}) \\ &= \operatorname{argmin}_{\boldsymbol{X}\in\mathbb{R}^{n\times d}} \{\|\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{W}\|_F^2 \\ &+ \frac{\rho_1}{2}\|\boldsymbol{X} - \boldsymbol{Y}^{(t)} + \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)}\|_F^2 \\ &+ \frac{\rho_2}{2}\|\boldsymbol{X}^T - \boldsymbol{Z}^{(t)} + \frac{1}{\rho_2}\boldsymbol{\Lambda}_2^{(t)}\|_F^2\}\end{aligned} \tag{8}$$

The necessary optimality condition further follows as

$$\frac{\partial \mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X}, \boldsymbol{Y}^{(t)}, \boldsymbol{Z}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)})}{\partial \boldsymbol{X}} = 0 \tag{9}$$

This implies

$$\begin{aligned}&2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^T + (\rho_1+\rho_2)\boldsymbol{X} \\ &= 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{W}^T + \rho_1(\boldsymbol{Y}^{(t)} - \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)}) + \rho_2(\boldsymbol{Z}^{(t)} - \frac{1}{\rho_2}\boldsymbol{\Lambda}_2^{(t)})^T\end{aligned} \tag{10}$$

For writing conveniently, let $\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} \in \mathbb{R}^{n\times n}$, $\boldsymbol{N} = \boldsymbol{W}\boldsymbol{W}^T \in \mathbb{R}^{d\times d}$, and $\boldsymbol{H} = 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{W}^T + \rho_1(\boldsymbol{Y}^{(t)} - \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)}) + \rho_2(\boldsymbol{Z}^{(t)} - \frac{1}{\rho_2}\boldsymbol{\Lambda}_2^{(t)})^T \in \mathbb{R}^{n\times d}$, then, the equation above becomes

$$2\boldsymbol{M}\boldsymbol{X}\boldsymbol{N} + (\rho_1+\rho_2)\boldsymbol{X} = \boldsymbol{H} \tag{11}$$

Since $\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{N} = \boldsymbol{W}\boldsymbol{W}^T \in \mathbb{R}^{d\times d}$ are both positive semi-definite and symmetric, we can perform eigenvalue decomposition with all non-negative eigenvalues, obtaining

$$\begin{aligned}\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{P}\boldsymbol{\Phi}\boldsymbol{P}^T \in \mathbb{R}^{n\times n} \\ \boldsymbol{N} = \boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{Q}\boldsymbol{\Psi}\boldsymbol{Q}^T \in \mathbb{R}^{d\times d}\end{aligned} \tag{12}$$

where $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ are two orthogonal matrices with all eigenvectors, $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_n \in \mathbb{R}^{n\times n}$, $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_d \in \mathbb{R}^{d\times d}$, and $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{d\times d}$ are two diagonal matrices with all non-negative eigenvalues.

Plugging Eq. (12) into Eq. (11), we obtain

$$2\boldsymbol{P}\boldsymbol{\Phi}\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q}\boldsymbol{\Psi}\boldsymbol{Q}^T + (\rho_1+\rho_2)\boldsymbol{X} = \boldsymbol{H} \tag{13}$$

That is

$$2\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{\Phi}\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q}\boldsymbol{\Psi}\boldsymbol{Q}^T\boldsymbol{Q} + (\rho_1+\rho_2)\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q} = \boldsymbol{P}^T\boldsymbol{H}\boldsymbol{Q} \tag{14}$$

That is

$$2\boldsymbol{\Phi}\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q}\boldsymbol{\Psi} + (\rho_1+\rho_2)\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q} = \boldsymbol{P}^T\boldsymbol{H}\boldsymbol{Q} \tag{15}$$

Let $\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q} = \boldsymbol{D} \in \mathbb{R}^{n\times d}$, then Eq. (15) becomes

$$2\boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Psi} + (\rho_1+\rho_2)\boldsymbol{D} = \boldsymbol{P}^T\boldsymbol{H}\boldsymbol{Q} \tag{16}$$

Thus, the i-th row and j-th column element $D_{i,j}$ of the matrix $\boldsymbol{D}$ can be obtained by the following closed-form solution

$$D_{i,j} = \frac{(\boldsymbol{P}^T\boldsymbol{H}\boldsymbol{Q})_{i,j}}{2(\boldsymbol{\Phi})_{i,i}(\boldsymbol{\Psi})_{j,j} + (\rho_1+\rho_2)} \tag{17}$$

where $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, d$, $(\boldsymbol{P}^T\boldsymbol{H}\boldsymbol{Q})_{i,j}$ is the i-th row and j-th column element of the matrix $\boldsymbol{P}^T\boldsymbol{H}\boldsymbol{Q}$, $(\boldsymbol{\Phi})_{i,i}$ is the i-th element of the diagonal matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$, and $(\boldsymbol{\Psi})_{j,j}$ is the j-th element of the diagonal matrix $\boldsymbol{\Psi} \in \mathbb{R}^{d\times d}$.

As we know, $(\boldsymbol{\Phi})_{i,i} \geq 0$ and $(\boldsymbol{\Psi})_{j,j} \geq 0$. In the meantime, ρ_1 and ρ_2 are greater than zero in practice, so the denominator in the equation above is greater than zero. After obtaining $\boldsymbol{D}$, we can easily calculate $\boldsymbol{X}$ as

$$\begin{aligned} \boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q} &= \boldsymbol{D} \in \mathbb{R}^{n\times d} \\ \boldsymbol{P}\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Q}\boldsymbol{Q}^T &= \boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{Q}^T \in \mathbb{R}^{n\times d} \end{aligned} \tag{18}$$

Therefore, $\boldsymbol{X}^{(t+1)} = \boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{Q}^T$.

(2) Compute the subproblem about $\boldsymbol{Y}^{(t+1)}$:

Further we calculate the subproblem about $\boldsymbol{Y}^{(t+1)}$. When the other variables are fixed with the former iteration result $(\boldsymbol{X}^{(t+1)}, \boldsymbol{Z}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)})$, the subproblem about $\boldsymbol{Y}^{(t+1)}$ is as

$$\begin{aligned} \boldsymbol{Y}^{(t+1)} &= \mathrm{argmin}_{\boldsymbol{Y}\in\mathbb{R}^{n\times d}} \mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X}^{(t+1)}, \boldsymbol{Y}, \boldsymbol{Z}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)}) \\ &= \mathrm{argmin}_{\boldsymbol{Y}\in\mathbb{R}^{n\times d}} \{\alpha\|\boldsymbol{Y}\|_{2,1} + \frac{\rho_1}{2}\|\boldsymbol{X}^{(t+1)} - \boldsymbol{Y} + \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)}\|_F^2\} \end{aligned} \tag{19}$$

In order to solve the subproblem (19), we first decouple it as

$$\begin{aligned} \boldsymbol{Y}^{(t+1)} = \mathrm{argmin}_{\boldsymbol{Y}_i\in\mathbb{R}^d, i=1,2,\cdots,n} \{&\alpha\sum_{i=1}^{n}\|\boldsymbol{Y}_i\|_2 \\ &+ \frac{\rho_1}{2}\sum_{i=1}^{n}\|\boldsymbol{Y}_i - (\boldsymbol{X}^{(t+1)} + \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)})_i\|_2^2\} \end{aligned} \tag{20}$$

where $\boldsymbol{Y}_i \in \mathbb{R}^d$ and $(\boldsymbol{X}^{(t+1)} + \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)})_i \in \mathbb{R}^d$ $(i = 1, 2, \cdots, n)$ are the i-row of the matrices $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{X}^{(t+1)} + \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)} \in \mathbb{R}^{n\times d}$, respectively.

Define $\boldsymbol{Y}_i^{(t+1)} \in \mathbb{R}^d$ as the i-row of the matrix $\boldsymbol{Y}^{(t+1)} \in \mathbb{R}^{n\times d}$ $(i = 1, 2, \cdots, n)$, then, we have

$$\begin{aligned} \boldsymbol{Y}_i^{(t+1)} = \mathrm{argmin}_{\boldsymbol{Y}_i\in\mathbb{R}^d} \{&\alpha\|\boldsymbol{Y}_i\|_2 \\ &+ \frac{\rho_1}{2}\|\boldsymbol{Y}_i - (\boldsymbol{X}^{(t+1)} + \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)})_i\|_2^2\} \end{aligned} \tag{21}$$

The problem (21) can be solved by the following lemma:

Lemma 1 *For any $\lambda > 0$ and $\boldsymbol{u} \in \mathbb{R}^d$, the minimizer of*

$$\boldsymbol{w}^\star = \text{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\{\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{w}\|_2\} \tag{22}$$

is given by

$$\boldsymbol{w}^\star = \begin{cases} \mathbf{0} & \text{if } \|\boldsymbol{u}\|_2 \le \lambda \\ (1-\frac{\lambda}{\|\boldsymbol{u}\|_2})\boldsymbol{u} & \text{if } \|\boldsymbol{u}\|_2 > \lambda \end{cases} \tag{23}$$

Based on this lemma, we can obtain the optimal $\boldsymbol{Y}_i^{(t+1)} \in \mathbb{R}^d$ as

$$\boldsymbol{Y}_i^{(t+1)} = \begin{cases} \mathbf{0} & \text{if } \|\boldsymbol{S}_i^{(t)}\|_2 \le \frac{\alpha}{\rho_1} \\ (1-\frac{\alpha}{\rho_1\|\boldsymbol{S}_i^{(t)}\|_2})\boldsymbol{S}_i^{(t)} & \text{if } \|\boldsymbol{S}_i^{(t)}\|_2 > \frac{\alpha}{\rho_1} \end{cases} \tag{24}$$

where $\boldsymbol{S}_i^{(t)}$ is the i-th row of matrix $\boldsymbol{S}^{(t)}$ ($i = 1, 2, \cdots, n$), and $\boldsymbol{S}^{(t)} = \boldsymbol{X}^{(t+1)} + \frac{1}{\rho_1}\boldsymbol{\Lambda}_1^{(t)} \in \mathbb{R}^{n\times d}$.

(3) Compute the subproblem about $\boldsymbol{Z}^{(t+1)}$:

Further we calculate the subproblem about $\boldsymbol{Z}^{(t+1)}$. When the other variables are fixed with the former iteration result $(\boldsymbol{X}^{(t+1)}, \boldsymbol{Y}^{(t+1)}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)})$, the subproblem about $\boldsymbol{Z}^{(t+1)}$ is as

$$\begin{aligned} \boldsymbol{Z}^{(t+1)} &= \text{argmin}_{\boldsymbol{Z}\in\mathbb{R}^{d\times n}} \mathcal{L}_{\rho_1,\rho_2}(\boldsymbol{X}^{(t+1)}, \boldsymbol{Y}^{(t+1)}, \boldsymbol{Z}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)}) \\ &= \text{argmin}_{\boldsymbol{Z}\in\mathbb{R}^{d\times n}}\{\beta\|\boldsymbol{Z}\|_{2,1} + \frac{\rho_2}{2}\|(\boldsymbol{X}^{(t+1)})^T - \boldsymbol{Z} + \frac{1}{\rho_2}\boldsymbol{\Lambda}_2^{(t)}\|_F^2\} \end{aligned} \tag{25}$$

Similar to solve (19), the optimal $\boldsymbol{Z}^{(t+1)}$ can be easily obtained by

$$\boldsymbol{Z}_j^{(t+1)} = \begin{cases} \mathbf{0} & \text{if } \|\boldsymbol{T}_j^{(t)}\|_2 \le \frac{\beta}{\rho_2} \\ (1-\frac{\beta}{\rho_2\|\boldsymbol{T}_j^{(t)}\|_2})\boldsymbol{T}_j^{(t)} & \text{if } \|\boldsymbol{T}_j^{(t)}\|_2 > \frac{\beta}{\rho_2} \end{cases} \tag{26}$$

where $\boldsymbol{T}_j^{(t)}$ is the j-th row of matrix $\boldsymbol{T}^{(t)}$ ($j = 1, 2, \cdots, d$), and $\boldsymbol{T}^{(t)} = (\boldsymbol{X}^{(t+1)})^T + \frac{1}{\rho_2}\boldsymbol{\Lambda}_2^{(t)} \in \mathbb{R}^{d\times n}$.

5. Complexity Analysis

The main computation in each iteration comes from updating $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and dual variables $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$. The update steps of the dual variables refer to one matrix multiplication and several matrix additions whose computational complexity is $\mathcal{O}(n^2d)$. For updating $\boldsymbol{X}$, it refers to several matrix multiplications and two eigenvalue decompositions, which costs $\mathcal{O}(n^3 + n^2d + d^2n + d^3)$. For updating $\boldsymbol{Y}$, the complexity is of order $\mathcal{O}(nd)$. Updating $\boldsymbol{Z}$ needs $\mathcal{O}(n^2d)$. Therefore, the total computational complexity in each iteration is $\mathcal{O}(n^3 + n^2d + d^2n + d^3)$.

6. Conclusion

This work provides a theoretical foundation for a unified framework that uses CUR matrix decomposition to perform active learning (selecting informative instances) and feature selection simultaneously. The analysis proves that selecting the most influential rows and columns via their optimization processing offers strong performance guarantees. The method is interpretable as it uses actual data points and features, offering a great potential for high-dimensional data analysis [7, 8, 20, 29], group fairness analysis [19, 27], and outlier or anomaly detection tasks [10, 15, 18].

References

[1] Hervé Abdi. Singular value decomposition (svd) and generalized singular value decomposition. *Encyclopedia of Measurement and Statistics*, 907(912):44, 2007.

[2] Afshin Ahmadi, Felice Manganiello, Amin Khademi, and Melissa C Smith. A parallel jacobi-embedded gauss-seidel method. *IEEE Transactions on Parallel and Distributed Systems*, 32(6): 1452–1464, 2021.

[3] Christos Boutsidis and David P Woodruff. Optimal cur matrix decompositions. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 353–362, 2014.

[4] Davide Cacciarelli and Murat Kulahci. Active learning for data streams: a survey. *Machine Learning*, 113(1):185–239, 2024.

[5] Quézia Cavalcante and Milton J Porsani. Low-rank seismic data reconstruction and denoising by cur matrix decompositions. *Geophysical Prospecting*, 70(2):362–376, 2022.

[6] Zhong Chen, Zhide Fang, Wei Fan, Andrea Edwards, and Kun Zhang. Cstg: An effective framework for cost-sensitive sparse online learning. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 759–767. SIAM, 2017.

[7] Zhong Chen, Zhide Fang, Victor Sheng, Jiabin Zhao, Wei Fan, Andrea Edwards, and Kun Zhang. Adaptive robust local online density estimation for streaming data. *International Journal of Machine Learning and Cybernetics*, 12(6):1803–1824, 2021.

[8] Zhong Chen, Huixin Zhan, Victor Sheng, Andrea Edwards, and Kun Zhang. Projection dual averaging based second-order online learning. In *2022 IEEE International Conference on Data Mining*, pages 51–60. IEEE, 2022.

[9] Zhong Chen, Huixin Zhan, Victor Sheng, Andrea Edwards, and Kun Zhang. Proximal cost-sensitive sparse group online learning. In *2022 IEEE International Conference on Big Data*, pages 495–504. IEEE, 2022.

[10] Zhong Chen, Victor Sheng, Andrea Edwards, and Kun Zhang. An effective cost-sensitive sparse online learning framework for imbalanced streaming data classification and its application to online anomaly detection. *Knowledge and Information Systems*, 65(1):59–87, 2023.

[11] Zhong Chen, Yi He, Di Wu, Huixin Zhan, Victor Sheng, and Kun Zhang. Robust sparse online learning for data streams with streaming features. In *Proceedings of the 2024 SIAM International Conference on Data Mining*, pages 181–189. SIAM, 2024.

[12] Zhong Chen, Yi He, Di Wu, Liudong Zuo, Keren Li, Wenbin Zhang, and Zhiqiang Deng. $\ell_{1,2}$-norm and cur decomposition based sparse online active learning for data streams with streaming features. In *2024 IEEE International Conference on Big Data*, pages 384–393. IEEE, 2024.

[13] Zhong Chen, Victor Sheng, Andrea Edwards, and Kun Zhang. Cost-sensitive sparse group online learning for imbalanced data streams. *Machine Learning*, 113(7):4407–4444, 2024.

[14] Zhong Chen, Yi He, Di Wu, Wenbin Zhang, and Zhiqiang Deng. $\ell_{1,\infty}$ mixed norm promoted row sparsity for fast online cur decomposition learning in varying feature spaces. In *Proceedings of the 2025 SIAM International Conference on Data Mining*, pages 124–133. SIAM, 2025.

[15] Zhong Chen, Yi He, Di Wu, Chen Zhao, and Meikang Qiu. A novel sparse active online learning framework for fast and accurate streaming anomaly detection over data streams. In *International Joint Conferences on Artificial Intelligence*, pages 2740–2748, 2025.

[16] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block admm with o (1/k) convergence. *Journal of Scientific Computing*, 71(2):712–736, 2017.

[17] Claudio Gambella and Andrea Simonetto. Multiblock admm heuristics for mixed-binary optimization on classical and quantum computers. *IEEE Transactions on Quantum Engineering*, 1:1–22, 2020.

[18] Heng Lian, Yi He, Di Wu, Zhong Chen, Xingquan Zhu, and Xindong Wu. Online outlier detection in open feature spaces. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[19] Heng Lian, Chen Zhao, Zhong Chen, Xingquan Zhu, My T Thai, and Yi He. Metric-agnostic continual learning for sustainable group fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18648–18657, 2025.

[20] Cheng Liang, Di Wu, Yi He, Teng Huang, Zhong Chen, and Xin Luo. Mma: Multi-metric-autoencoder for analyzing high-dimensional and incomplete data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 3–19. Springer, 2023.

[21] Sanmin Liu, Shan Xue, Jia Wu, Chuan Zhou, Jian Yang, Zhao Li, and Jie Cao. Online active learning for drifting data streams. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):186–200, 2021.

[22] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[23] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[24] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*, 2021.

[25] Jialei Wang, Peilin Zhao, Steven CH Hoi, and Rong Jin. Online feature selection and its applications. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):698–710, 2013.

[26] Jing Wang, Meng Wang, Peipei Li, Luoqi Liu, Zhongqiu Zhao, Xuegang Hu, and Xindong Wu. Online feature selection with group structure analysis. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3029–3041, 2015.

[27] Zichong Wang, Jocelyn Dzuong, Xiaoyong Yuan, Zhong Chen, Yanzhao Wu, Xin Yao, and Wenbin Zhang. Individual fairness with group awareness under uncertainty. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 89–106. Springer, 2024.

[28] Di Wu, Yi He, Xin Luo, and MengChu Zhou. A latent factor analysis-based approach to online sparse streaming feature selection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(11):6744–6758, 2021.

[29] Di Wu, Shengda Zhuo, Yu Wang, Zhong Chen, and Yi He. Online semi-supervised learning with mix-typed streaming features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4720–4728, 2023.

[30] Xindong Wu, Kui Yu, Wei Ding, Hao Wang, and Xingquan Zhu. Online feature selection with streaming features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5): 1178–1192, 2012.

Appendix

The key steps of the CUR algorithm are summarized in Algorithm 1. We can extend our method to the kernel version by defining a new data representation to incorporate the kernel information.

Algorithm 1 The CUR Decomposition Algorithm

Online input: the data matrix $\boldsymbol{W} \in \mathbb{R}^{d\times n}$, parameters, $\alpha > 0$ and $\beta > 0$.
Online output: the matrix, $\boldsymbol{X}^{(t)} \in \mathbb{R}^{n\times d}$.

1: **Initialization**: $\boldsymbol{Y}^{(0)} = \mathbf{0} \in \mathbb{R}^{n\times d}$, $\boldsymbol{Z}^{(0)} = \mathbf{0} \in \mathbb{R}^{d\times n}$, $\boldsymbol{\Lambda}_1^{(0)} = \mathbf{0} \in \mathbb{R}^{n\times d}$, $\boldsymbol{\Lambda}_2^{(0)} = \mathbf{0} \in \mathbb{R}^{d\times n}$, $\rho_1 > 0$, $\rho_2 > 0$, $\rho = 10^{10}$, $\tau > 0$, $\varepsilon > 0$, $t = 0$;
2: **While** *not converged* **do**
3: fix the other variables and update $\boldsymbol{X}^{(t+1)}$ by $\boldsymbol{X}^{(t+1)} = \boldsymbol{X} = \boldsymbol{PDQ}^T$ in Eq. (18);
4: fix the other variables and update $\boldsymbol{Y}^{(t+1)}$ by Eq. (24);
5: fix the other variables and update $\boldsymbol{Z}^{(t+1)}$ by Eq. (25);
6: update the multipliers $\boldsymbol{\Lambda}_1^{(t+1)} = \boldsymbol{\Lambda}_1^{(t)} + \rho_1(\boldsymbol{X}^{(t+1)} - \boldsymbol{Y}^{(t+1)})$ and $\boldsymbol{\Lambda}_2^{(t+1)} = \boldsymbol{\Lambda}_2^{(t)} + \rho_2((\boldsymbol{W}^{(t+1)})^T - \boldsymbol{Z}^{(t+1)})$;
7: update the parameters ρ_1 and ρ_2 by $\rho_1 = \min(\tau\rho_1, \rho)$ and $\rho_2 = \min(\tau\rho_2, \rho)$;
8: $t \leftarrow t + 1$;
9: check the convergence conditions (1) $\|\boldsymbol{X}^{(t)} - \boldsymbol{Y}^{(t)}\|_\infty < \varepsilon$; (2) $\|(\boldsymbol{X}^{(t)})^T - \boldsymbol{Z}^{(t)}\|_\infty < \varepsilon$; and (3) $|\frac{f(\boldsymbol{X}^{(t)}) - f(\boldsymbol{X}^{(t-1)})}{f(\boldsymbol{X}^{(t-1)})}| < \varepsilon$ are all satisfied, where $f(\boldsymbol{X}) = \|\boldsymbol{W} - \boldsymbol{WXW}\|_F^2 + \alpha\|\boldsymbol{X}\|_{2,1} + \beta\|\boldsymbol{X}^T\|_{2,1}$ is the objective function value of Eq. (4) at the point $\boldsymbol{X}$.
10: **end**