**ABSTRACT**

Feature selection is an important component of many machine learning applications. In this paper, we propose a new robust feature selection method with emphasizing joint $\ell_{2,1}$-norm minimization on both loss function and regularization. The $\ell_{2,1}$-norm based loss function is robust to outliers in data points and the $\ell_{2,1}$-norm regularization selects features across all data points with joint sparsity. An efficient and general algorithm is introduced with proved convergence.

**NOTATIONS AND DEFINITIONS**

In the paper, the matrices are written as boldface uppercase letters, and the vectors are written as boldface lowercase letters. For matrix $M = (m_{ij})$, its $i$-th row, $j$-th column are denoted by $m_i$, $m_j$ respectively. The $\ell_{2,1}$-norm of a matrix $M$ is defined as

$$\|M\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} m_{ij}^2} = \sum_{i=1}^{n} \|m_i\|_2$$

The problem becomes:

$$\min_{U} \|X^TW - Y\|_{2,1} + \gamma \|W\|_{2,1}$$

The regularization term penalizes all $c$ regression coefficients corresponding to a single feature as a whole. This has the effects of feature selection. Denote data matrix $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$ and label matrix $Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{n \times c}$. Then the problem (3) can be written as

$$\min_{W} \|X^TW - Y\|_{2,1} + \gamma \|W\|_{2,1}$$

It is a convex problem. The first $\ell_{2,1}$-norm induces sparsity along the direction of data points, and the second $\ell_{2,1}$-norm induces sparsity along the direction of features. Thus this joint $\ell_{2,1}$-norm minimization produces data point selection and feature selection simultaneously.

**Algorithm Convergence Analysis**

**Lemma 1** For any nonzero vectors $u, v \in \mathbb{R}^n$, the following inequality holds:

$$\|u\|_2 - \frac{1}{2}\|u\|_2^2 \leq \|u - v\|_2 - \frac{1}{2}\|u - v\|_2^2$$

**Theorem 1** The Algorithm will monotonically decrease the objective in each iteration, and converge to the global optimum of the problem.

**Proof:** In the $t$ iteration, according to the step 1 of the loop in the algorithm, we have $U_{t+1} = \arg \min_{U} Tr(U^TDAU)$, which indicates that $Tr(U_{t+1}^TDAU_0) \leq Tr(U_{t}^TDAU_0)$. That is to say,

$$\sum_{i=1}^{m} \|u_{i+1}^t\|_2^2 \leq \sum_{i=1}^{m} \|u_{i}^t\|_2^2$$

On the other hand, according to Lemma 1, we have:

$$\sum_{i=1}^{m} \left(\|u_{i+1}^t\|_2 - \frac{1}{2}\|u_{i+1}^t\|_2^2\right) \leq \sum_{i=1}^{m} \left(\|u_{i}^t\|_2 - \frac{1}{2}\|u_{i}^t\|_2^2\right)$$

Combining Eq. (7) and Eq. (8), we arrive at $\sum_{i=1}^{m} \|u_{i+1}^t\|_2 \leq \sum_{i=1}^{m} \|u_{i}^t\|_2$. That is to say, $\|U_{t+1}\|_{2,1} \leq \|U_t\|_{2,1}$. Thus the algorithm will monotonically decrease the objective in each iteration $t$. As the problem is convex, the algorithm will converge to the global optimum of the problem.

**EXPERIMENTAL RESULTS**