Robust Multi-View Feature Selection Method

Yufang Liu¹, Sibin Xu², He Huang¹, Fei Zhao²*, Linyu Wang³, Xin Liang², Ruixin Tang²

¹State Grid Jiangsu Electric Power Company Limited
²Guangdong University of Technology
³State Grid (Suzhou) City & Energy Research Institute Co., LTD
*Corresponding author’s e-mail: 503188839@qq.com

Abstract. Recently, Multi-view learning has attracted widely attention. Inspired by the idea of RFS [6], we propose to utilize the Multi-View features to develop a Multi-View feature selection method. In our work, the weight of each view can be adaptively adjusted. An efficient algorithm is designed to solve the proposed method. Numerical study confirms the effectiveness of our proposed method.

1. Introduction

Feature dimension explosion is one of the critical issues in efficiency improvement of machine learning model. Too many features significantly decrease the accuracy, training speed and nearly all other efficiencies in model adjustment [1][2]. Thus feature selection becomes a necessary steps before model learning with large feature dimension [3][4].

The feature selection plays a key role in machine learning, effective selection is the basis of accuracy of prediction. Feature Selection refers to an art of extracting a subset of features from an original data space. Feature selection process will try to find the smallest input set for a predictor that produces the best results concerning accuracy and model size [5]. The most recent approaches have been reported in [6]-[8]. L1 regularization can produce sparse weight matrix used for feature selection, in [6], the author proposed a new robust feature selection method with emphasizing joint $L_{2,1}$-norm minimization on both loss function and regularization, which makes the feature selection process more efficient. PCA(Principle Component Analysis) is also a powerful tool for analysis data, in [7], the article shows that features selection can be done effectively using combination of thresholding-based ROI template and PCA (Principle Component Analysis) methods. Pearson product-moment correlation coefficient can measure the correlation between two variables, in [8], the author proposed sentiment analysis on movie review using ensemble features and Bag of Words and selection Features Pearson's Correlation to reduce the dimension of the feature and get the optimal feature combinations.

The development of data collection technology allows people to describe things through more complex and diverse means, which also leads to more data features. Among the features have different attributes, so it is not suitable to use the same model for learning, Multi-View method can solve this problem effectively. The most recent approaches have been reported in [9]-[11]. In [9], the author proposed a Multi-View semi-supervised sparse feature selection method based on graph Laplacian, the results show that the proposed Multi-View method outperforms the single-view methods. In [10], the paper present a new semi-supervised sparse feature selection framework based on Multi-View Hessian regularization to obtain better performance, experimental results show that the proposed method can...
realize feature selection well. In [11], the author applies robust Multi-View K-means to obtain the robust and high quality pseudo labels for sparse feature selection in an efficient way.

This paper is an attempt on combination between Multi-View structure and model in [6]. The new model selects the $L_{2,1}$-norm from [6] across all samples with joint sparsity for feature significance presentation. Different from model in [6], the proposed model is established on Multi-View structure with new parameters. With this structure, feature selection could be not only unitedly processed but also evaluated under different views.

Content of this paper is introduced below. Section 2 introduced the detail of the proposed model. Section 3 introduces the training process of the proposed model. Section 4 selects a numerical study to verify performance of the proposed model.

2. Feature Selection Modelling

This paper selects Least Square Regression as based model. Equation (1) introduces the objective function in [6] and equation (2) introduces the objective function of proposed model.

$$
\min_w \text{Obj}_{NPS10} = \frac{1}{\gamma} \|X^TW - Y^T\|_{2,1} + \|W\|_{2,1}
$$

$$
\begin{cases}
\min_{\alpha, W} \text{Obj}_{NPS} = \frac{1}{\gamma} \sum_{i=1}^{N} \alpha_i \|X_i^TW_n - Y^T\|_{2,1} + \|W\|_{2,1} \\
s.t.: \sum_{i=1}^{N} \alpha_i = 1
\end{cases}
$$

$$
\|W\|_{2,1} = \left( \sum_{i=1}^{N} u_{i,j} \right)^{\frac{1}{2}}
$$

Difference between equation (1) and (2) is the term representing accuracy insurance. In equation (1), the accuracy insuring term is a typical least square regression. In equation (2), it becomes a least square regression under Multi-View data. In equation (2), feature space is separated into N views. Each view is marked as $X_n$. The corresponding weights matrix to each $X_n$ is $W_n$. Since different views capture the different levels of information about the object, we introduce an independent parameter $\alpha_n$ to reflect the weight of each view. Moreover, the sparsity constraint is added in equation (2) to select the reduced features of each view. Equation (3) is the expression of $L_{2,1}$-norm.

From equation (3), $L_{2,1}$-norm is the summary of $L_2$-norm of each row in W. When separates W under Multi-View data, W can be directly separated into several $W_n$ by telling apart rows only. So equation (2) can be rewritten as equation (4).

$$
\begin{cases}
\min_{\alpha, W} \text{Obj}_{NPS} = \frac{1}{\gamma} \sum_{i=1}^{N} \alpha_i \|X_i^TW_n - Y^T\|_{2,1} + \gamma \|W_n\|_{2,1} \\
s.t.: \sum_{i=1}^{N} \alpha_i = 1
\end{cases}
$$

3. Model Learning Algorithm

Model learning is the optimization process on equation (4). $\alpha$ and W are two sets of parameters requiring optimization. This section chooses to optimize them sequentially. First to update $\alpha$ and then W.

3.1. Optimization Step for $\alpha_n$ Only

In the step for $\alpha_n$, W is recognized as constant value. The Lagrangian function of equation (4) can be rewritten into equation (5).
Consider the KKT condition, equation (6) can be obtained.

\[
\frac{\partial L}{\partial \alpha_n} = \frac{r \alpha_n^{r-1}}{\gamma} \| X_n^T W_n - Y^T \|_{2,1} - \lambda = 0
\]

\[
\sum_{n=1}^{N} \alpha_n = 1
\]

Solution of the function in equation (6) is shown in equation (7).

\[
\alpha_n = \left( \frac{1}{\| X_n^T W_n - Y^T \|_{2,1}} \right)^{\frac{1}{r-1}} \left( \sum_{n=1}^{N} \frac{1}{\| X_n^T W_n - Y^T \|_{2,1}} \right)^{\frac{1}{r-1}}
\]

With equation (7), the update of \( \alpha_n \) can be obtained in equation (8). In equation (8), \( W_{n,0} \) means \( W_n \) in the ith step. \( \alpha_{n,(i+1)} \) means \( \alpha_n \) in the i+1 step.

\[
\alpha_{n,(i+1)} = \frac{\sum_{n=1}^{N} \left( \frac{1}{\| X_n^T W_{n,0} - Y^T \|_{2,1}} \right)^{\frac{1}{r-1}}}{\sum_{n=1}^{N} \left( \frac{1}{\| X_n^T W_{n,0} - Y^T \|_{2,1}} \right)^{\frac{1}{r-1}}}
\]

### 3.2. Optimization Step for W Only

In the step for \( W \), \( \alpha_n \) is recognized as constant. As the constraint in equation (4) is independent to \( W \), this step can ignore the constraint in equation (4). Following similar spirit in [6], objective function can be rewritten into equation (9).

\[
\text{Obj}_{\text{NSW}} = \sum_{n=1}^{N} \{ \| E_n \|_{2,1} + \gamma \| W_n \|_{2,1} \}
\]

\[
Y^T = \alpha_n^T X_n^T W_n + \gamma E_n
\]

To further simplification, equation (9) can be rewritten into equation (10).

\[
U_n = \begin{bmatrix} W_n \\ E_n \end{bmatrix} \Rightarrow \text{Obj}_{\text{NSW}} = \sum_{n=1}^{N} (\| U_n \|_{2,1})
\]

s.t. \( A_n U_n = Y^T, A_n = [\alpha_n^T X_n^T, \gamma I_n] \)

Based on equation (10), the Lagrangian function can be obtain in equation (11).

\[
L_n = \| U_n \|_{2,1} - \text{Tr}(A_n^T (A_n U_n - Y^T))
\]

\[
\frac{\partial L_n}{\partial U_n} = 2D_n U_n - A_n^T A_n
\]

\[
D_n \text{ is diagonal matrix, } d_u = \frac{1}{2\| u_n \|_2}
\]

Consider that the gradient is equal to 0, equation (12) can be obtained.

\[
U_n = D_n^{-1} A_n^T (A_n D_n^{-1} A_n^T)^{-1} Y^T
\]

Due to matrix \( D_n \) is dependent to \( U_n \), the optimization can firstly updates matrix \( D_n \) and then updates the matrix \( U_n \). Equation (13) finally reveals the updates of \( W_n \).
3.3. Learning Algorithm

The sequential learning algorithm is constructed by expression in equation (8) and equation (13). Table I reveals the optimization process of model learning.

Table 1. The Entire Optimization Process

| Data Import: | Xn, Y  |
|-------------|--------|
| Result:     | Wn, αn |
| Set t = 0. Initialize | Wn, αn, Dn |
| for i = 0 to k | calculate αn with equation (8) |
| calculate U with equation (13) |
| calculate the diagonal matrix D_{i+1}, where the i-th diagonal element is | \frac{1}{2\|u_{i+1}'\|_2} |
| Obtain Wn from Un |

In Table 1, the necessary data Xn and Y, Wn, αn, Dn are initialized. A limited maximum step is set to end the algorithm. With in each iteration, the algorithm firstly fix Wn and Dn, but update αn with equation (8). Secondly, the algorithm fix αn and update U, Dn with equation (13). Wn can be obtained from the upper half of Un.

4. Numerical Study

4.1. Background

A numerical study is implemented to verify the capability of the proposed method. MSRC-v1 and Caltech 101-7 are the two sets of data selected for model training. In MSRC-v1, there are 6 views and the total feature dimension is 2428 [12]. In Caltech 101-7, there are 6 views and the total feature dimension is 3766 [13]. The proposed method will be compared with the model in [RFS]. The selected features from both models will be sent to a typical KNN model to examine the effect of feature selection. Matlab 2016 is used as the platform for model learning.

4.2. Experimental Result of MSRC-v1

Table II reveals the result on MSRC-v1. It compares the results from RFS and the proposed model under two dimensions, which includes different value of r and different quantity of selected features.
Table 2. Experimental Result of RFS and The Proposed Model on MSRC-v1

|          | F5    | F5    | F10    | F10    |
|----------|-------|-------|--------|--------|
| r = 0.1  | 0.9   | 0.890476 | 0.933333 | 0.933333 |
| r = 0.2  | 0.92381 | 0.890476 | 0.961905 | 0.933333 |
| r = 0.4  | 0.919048 | 0.890476 | 0.971429 | 0.933333 |
| r = 0.8  | 0.928571 | 0.890476 | 0.942857 | 0.933333 |
| r = 1.2  | 0.857143 | 0.890476 | 0.947619 | 0.933333 |
| r = 1.4  | 0.885714 | 0.890476 | 0.928571 | 0.933333 |
| r = 1.6  | 0.866667 | 0.890476 | 0.933333 | 0.933333 |
| r = 1.8  | 0.857143 | 0.890476 | 0.92381  | 0.933333 |
| r = 2    | 0.890524 | 0.890476 | 0.895238 | 0.933333 |
| r = 4    | 0.733333 | 0.890476 | 0.871429 | 0.933333 |
| r = 8    | 0.733333 | 0.890476 | 0.871429 | 0.933333 |
| r = 10   | 0.733333 | 0.890476 | 0.871429 | 0.933333 |

Table 3. Experimental Result of RFS and The Proposed Model on Caltech 101-7

|          | F5    | F5    | F10    | F10    |
|----------|-------|-------|--------|--------|
| r = 0.1  | 0.901934 | 0.892964 | 0.916786 | 0.918726 |
| r = 0.2  | 0.905295 | 0.892964 | 0.924266 | 0.918726 |
| r = 0.4  | 0.908665 | 0.892964 | 0.930354 | 0.918726 |
| r = 0.8  | 0.926089 | 0.892964 | 0.932893 | 0.918726 |
| r = 1.2  | 0.913909 | 0.892964 | 0.92088  | 0.918726 |
| r = 1.4  | 0.919942 | 0.892964 | 0.936325 | 0.918726 |
| r = 1.6  | 0.915964 | 0.892964 | 0.927509 | 0.918726 |
| r = 1.8  | 0.909853 | 0.892964 | 0.920105 | 0.918726 |
| r = 2    | 0.903839 | 0.892964 | 0.920105 | 0.918726 |
| r = 4    | 0.879577 | 0.892964 | 0.895976 | 0.918726 |
| r = 8    | 0.879577 | 0.892964 | 0.895976 | 0.918726 |
| r = 10   | 0.879577 | 0.892964 | 0.895976 | 0.918726 |

From angle of dimension r, it is obvious that new model performs better than RFS when r is smaller than 1. Results from Table 2 shows that the new model obtains 1% accuracy improvement than RFS averagely. Fig.1 introduces the average improvements at different optimal feature quantity with r < 1. As the following, we explain why setting r < 1 can obtain the better performance. In Eq.(8), we can find that the weight $\alpha_n$ is proportional to the regression residual of each view when r < 1. The large (small) residual regression can lead to a large (small) weight. In the next iteration, this large
weight can enhance (reduce) the residual regression. Hence, this updating rule can help to fully utilize the complementary information of Multi-View data.

4.3. Experimental Result of Caltech 101-7

Table III reveals the result on Caltech 101-7. It compares the results from RFS and the proposed model under two dimensions, which includes different value of r and different quantity of selected features.

From angle of dimension r, new model performs better than RFS in most of the cases, especially for case with r = 0.4. Results from Table 3 show that the new model obtains about 1% accuracy improvement than RFS averagely. Fig.2 introduces the average improvements at different optimal feature quantity with r = 0.4.

5. Conclusions

This paper proposed a feature selection method for classification under Multi-View data. The idea is derived from RFS and is combined with Multi-View data. It evaluates the effect of all features together under several views with $\alpha_n$. A two stages Lagrangian optimization is proposed for this model. In each iteration, one stage is to optimize W and fix $\alpha_n$ and the other is to optimize $\alpha_n$ and fix W. Simulation result shows that the proposed model performs better when r is less than 1.

Acknowledgment

This research work was supported in part by the National Science Foundation of China: A Time-Based-Demand-Response Program of Compensated Multiple-Shape Pricing Scheme, Grant No. 51707041; in part by the State Grid Energy Research Institute Multiple Capability Research of Distribution Grid and The algorithm of Relevant Analyse Optimization Model; in part by the State Grid (Suzhou) City & Energy Research Institute project (Critical Data Acquisition and Obtaining Method research on City Energy Research); in part by the science and technology project of State Grid Corporation of China (The key technologies and applications of State Grid’s social image improvement based on social media big data, Grant no. 1400-201957293A-0-0-00); in part by the young talents project of State Grid Energy Research Institute Co. Ltd. (Research on recognition and processing technology of news hotspots in energy industry based on the natural language processing, Grant no.526700180001).

References

[1] J. Shawe-Taylor, N. Cristianini. (2004) Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge.
[2] Y. Sun, D. Wu. (2008) A relief based feature extraction algorithm. Proc. SIAM Int. Conf. Data
[3] R. Kohavi, G. H. John. (1997) Wrappers for feature subset selection. Artif. Intell., vol. 97, no. 1/2, pp. 273-324.

[4] P. A. Mundra, J. C. Rajapakse. (2010) SVM-RFE with MRMR filter for gene selection. IEEE Trans. NanoBiosci., vol. 9, no. 1, pp. 31-37.

[5] S. Visalakshi, V. Radha. (2014) A literature review of feature selection techniques and applications: Review of feature selection in data mining. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-6.

[6] Nie, Feiping, Huang, Heng, Cai, Xiao , Ding, Chris. (2010) Efficient and robust feature selection via joint 2;1-norms minimization. 24th Annual Conference on Neural Information Processing Systems, pp. 1-6.

[7] N.R. Ratnasari, Adhi Susanto, Indah Soesanti, Maesadji. (2013) Thoracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes. 2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME), pp. 65-69.

[8] Fachrul Rozy Saputra Rangkuti, M. Ali Fauzi, Yuita Arum Sari, Eka Dewi Lukmana Sari. (2018) Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection. 2018 International Conference on Sustainable Information Engineering and Technology (SIET), pp. 88-91.

[9] Caijuan Shi, Qiuci Ruan, Gaoyun An. (2014) Multi-view Laplacian sparse feature selection for web image annotation. 2014 12th International Conference on Signal Processing (ICSP), pp. 1026-1029.

[10] Caijuan Shi, Jian Liu, Liping Liu, Xiaodong Yan, (2015) Semi-supervised sparse feature selection based on multi-view Hessian regularization. 6th International Conference on Wireless, Mobile and Multi-Media (ICWMMN 2015), pp. 158-161.

[11] Hongfu Liu, Haiyi Mao, Yun Fu, (2016) Robust Multi-View Feature Selection. 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 281-290.

[12] Winn J M, Jojic N. (2005) LOCUS: Learning Object Classes with Unsupervised Segmentation[C]. 10th IEEE International Conference on Computer Vision (ICCV 2005), Beijing, China.

[13] Fei-Fei L , Fergus R , Perona P. (2007) Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories[J]. Computer Vision and Image Understanding, 106(1):59-70.