Multi-label learning with label-specific feature selection via local positive and negative correlation

Xia Shuxin¹

¹Affiliation not available

May 03, 2024
Multi-label learning with label-specific feature selection via local positive and negative correlation

Xia Shuxin: banditxsx@163.com

School of Information Engineering Zhe Jiang Ocean University, Zhoushan, 316022, China

Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province School of Information Engineering, Zhejiang Ocean University, Zhoushan, 316022, China

ABSTRACT

High dimension data increasingly emerged in many application areas. Feature selection methods are widely studied to solve the issues, playing a crucial role in eliminating redundant or irrelevant feature. Previous feature selection algorithms seldom consider the fact that each label might be determined by its specific features, which has its own characteristics. Meanwhile, they often neglect mutual exclusion among labels, named “negative label correlations” and assume labels share global correlation in exploring label correlation. To address the above problem, we propose a novel algorithm called “Multi-label learning with label-specific feature selection via local positive and negative correlation” (LLCPN-MIFS). Multiple datasets manifest that the effectiveness of the proposed method.

Keywords: label-specific feature selection; label correlation; positive label correlation; negative label correlation

1. INTRODUCTION

In the Big Data era, a huge quantity of data appeared and the dimension of data increased rapidly, which results in “dimensionality curse”\(^1\). Different from single label learning, dimension reduction of data is one of important problems of classification in multi-label learning. For multi-label data, a great many approaches have been proposed to address it. Those methods roughly classified into two categories: Feature extraction\(^2\) and Feature selection\(^3\). Compared with feature extraction, feature selection selects a representative features subset from the original dataset, which not only reduces the dimensions of data with high-dimension, but also preserves the original physical structure of the dataset that are interpretable. We concentrate on principally feature selection in this paper. As for feature selection algorithms, it is mainly divided into Filter\(^4\), Wrapper\(^5\) and Embedded methods\(^6\). But for algorithm performance, embedded methods typically are better than them so that numerous researchers have proposed common feature selection and label-specific feature selection. Among, the classic common feature selection algorithms include MDFS\(^7\), DMMFS\(^8\) and SCMFS\(^9\). Nevertheless, label-specific features can extract pivotal characteristics corresponding to each label, which means if the correlation coefficient between labels is large, the significant characteristics extracted by label-specific features have a high similarity learning from labels. Consequently, the combination of feature selection with learning label-specific feature can be more effective to improve the model performance, which is also that extant embedded multi-label feature
selection algorithm has not taken into account it. Some reflective methods embody in LSF-CI\textsuperscript{[10]}, GLFS\textsuperscript{[11]}, RFSFS\textsuperscript{[12]} and LRLSF\textsuperscript{[13]}.

For multi-label data, sufficiently exploring correlation among labels might be beneficial to enhance the performance of multi-label classification. But, most of multi-label algorithms centre on global correlation which signifies label is shared by all examples. Actually, different samples share different label correlations that local correlation exists in label information. There is not only positive correlation among labels, but also mutually exclusive one, namely negative label correlation. Probing negative label correlation between labels also be conducive to caputure classification accuracy. Some of proposed papers\textsuperscript{[14-15]} intergrate negative label correlation into multi-label learning.

As discussed above, the paper proposes a new algorithm “Multi-label learning with label-specific feature selection via locally positive and negative correlation ”(LLCPN-MIFS). The following experiments verify the effectiveness of method.

2. MATERIALS AND METHODS

2.1 Definition

Given a multi-label dataset $D = \{(x_i, y_i)\}_{i=1}^n$ with n instances. $X = [x_1, x_2, \ldots, x_n] \in R^{n \times m}$ represents instance matrix, where $x_i$ denotes i-th d dimension feature vector. $Y = [y_1, y_2, \ldots, y_m] \in R^{m \times n}$ is observed label matrix and $y_j (1 \leq i \leq m) \in R^n$ denotes correlation. For $\forall y_{ij}$, $y_{ij} = 1$ represents the instances $x_i$ belongs to $y_j$ but $y_{ij} = 0$ represents $x_i$ not belongs class $y_j$.

2.2 Learning label specific feature

In view of that each class label is correlated with some specific features from original feature space, which those features are recognitive to the corresponding label. For the construction of label-specific feature, inspired by the concept and operation\textsuperscript{[16-17]}, we use line regression to build the learning model of label-specific features and sparse $W$ whose some elements are changed into 0 by adding $l_2$ norm. So, the objection function is formulated as:

$$\min_W \frac{1}{2} \|X^T W - Y\|_F^2 + \lambda_1 \|W\|_1$$

(1)

Where, $W_i = [w_{i1}, w_{i2}, \ldots, w_{id}]^T$ shows the feature of coefficient vector of label $y_i$ which is used to measure identifiability $j$-th feature to $i$-th label.

2.3 Local label correlation

The intrinsic label correlation plays an important role in multi-label learning and reflects the relationship between feature recognition and label correlation strongly or weakly. However, most algorithms focus on exploring global correlation that assumes labels correlation are shared by all examples, which is not realistic in practice. Therefore, it shows that exploring local correlation among labels is the first choice. Enlightend by the ideal\textsuperscript{[18]}, we are still on the basis of the approach that: the smoothness is usually introduced in exploring local label correlation. It is perceived that the distance in the feature space can measure the similarity of their corresponding labels. Equivalently saying it can be equated that if label vectors $y_i$ and $y_j$ are very similar in inherent geometric space, and the ground-truth labels $\hat{Y}_i$ and
\( \hat{Y}_j \) should have similar structural features in the course of prediction. The paper introduces the manifold regularization term based on local smooth structure to achieve the construction of local label correlation followed by\(^{[19]}\). Detail steps are the following that: we denote \( C \) as label correlation matrix, where \( C_{ij} \) manifests the similarity between \( y_i \) and \( y_j \). Then, it can be expressed as follows:

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} \| YC_i - YC_j \|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} (\| YC_i \|_2^2 + \| YC_j \|_2^2) - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} \langle YC_i, YC_j \rangle \\
= tr(C^T Y^T L_x Y C)
\]

(2)

Where, \( L_x \) is Laplacian matrix of instance weight \( Q \) which is signified for the similarity between \( x_i \) and \( x_j \) and obtained by \( K \) means:

\[
Q_{ij} = \begin{cases} \\
\exp\left(\frac{\| x_i - x_j \|_2^2}{2\sigma^2}\right) x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\
0 \text{ otherwise}
\end{cases}
\]

(3)

2.4 Positive label correlation

In multi-label classification, sufficiently excavating complex relationship among labels can improve precision. Analogy to instance manifold regularization\(^{[20]}\), we apply to label manifold regularization to depict it on the hypothesis of that if there is a positive correlation between label, the smaller the Euclidean distance of regression coefficients, namely \( C_{ij} \) or \( C_{ji} \) larger. It can be written as follows:

\[
\sum_{i=1}^{l} \sum_{j=1}^{l} C_{ij} \| W_i - W_j \|_2^2 = tr(WL_p W^T)
\]

(4)

Where, \( L_p = D_p - C \) indicates the positive of \( C \) diagonal matrix and \( D_p \in R^{[l]} \) represents matrix whose elements both are one.

2.5 Negative label correlation

Not only exists in positive correlation among labels, but also is there a mutual exclusion, called “negative correlation” \(^{[21]}\). We can concretely describe if there is negative correlation among labels, then the Euclidean distance between \( W_i \) and \( -W_j \) the corresponding regression coefficients is greater, that is to say the distance between \( W_i \) and \( -W_j \) is smaller. Thus, it is formulated as:
\[
\sum_{i=1}^{l} \sum_{j=1}^{l} C_{ij} \| W_i - (-W_j) \|^2_2 = tr(WL_nW^T)
\]  \hspace{1cm} (5)

Where, \( D_s = L_s + C \) denotes Laplacian matrix of \( C \) whose elements both are -1 as diagonal matrix. Additionally, inspired by the paper\(^{[22]}\), we note that one class label may be correlated with only a subset of class labels. Thus, \( l_1 \)-norm is added on \( C \) to learn sparse label dependencies.

### 2.6 Objective function

Based on the above understanding, we can get final the model:

\[
\min_{w,c} \frac{1}{2} \| X^T W - Y \|_F^2 + \lambda_1 \| W \|_1 + \lambda_2 \| C \|_1 \\
+ \lambda_3 \text{tr}(C^T Y^T L_s YC) + \lambda_4 \text{tr}(WL_pW^T) + \lambda_5 \text{tr}(WL_nW^T)
\]  \hspace{1cm} (6)

Where, \( \lambda_i (i = 1, 2, 3, 4, 5) \) is trade-off parameter, used to control the complexity of the model.

### 2.7 Optimization process

The objective function is a convex optimization problem because both \( f(W) \) and \( g(W) \) are convex. We take a method called “Accelerated Proximal Gradient Method”\(^{[23]}\) to tackle it. For parameter with dual model, \( \theta \) is introduced to express it as a united form. Hence, the objective function is written as follows:

\[
\min_{\theta} \quad f(\theta) + g(\theta)
\]  \hspace{1cm} (7)

Where, they can be delivered respectively as follows:

\[
f(\theta) = \frac{1}{2} \| X^T W - Y \|_F^2 + \lambda_3 \text{tr}(C^T Y^T L_s YC) + \lambda_4 \text{tr}(WL_pW^T) + \lambda_5 \text{tr}(WL_nW^T)
\]  \hspace{1cm} (8)

\[
g(\theta) = \lambda_1 \| W \|_1 + \lambda_2 \| C \|_2
\]  \hspace{1cm} (9)

Furthermore, we can’t directly minimize \( F(W) \) and introduce the quadratic approximation of it, which can be expressed as:

\[
Q_L(\theta, \theta^T) = f(\theta) + \langle \nabla f(\theta^T), \theta - \theta^T \rangle + \frac{1}{2} \| \theta - \theta^T \|_F^2 + g(\theta)
\]  \hspace{1cm} (10)

And homologous to previous method from paper\(^{[24]}\), we solve \( \theta \) by minimizing the following formula:

\[
\theta' = \text{arg} \min_{\theta} \frac{1}{2} \| \theta - \theta^T \|_F^2
\]  \hspace{1cm} (11)

\( \forall L \geq l_f \) , where \( l_f \) denotes positive Lipschitz constant. Simultaneously, \( b_i \) that satisfies \( b_i^2 - b_i^2 \leq b_i \), can help
convergence rate improve $O(t^{-2})$, and $\theta_t$ represents $t$-th iteration of $\theta$.

Provided that there are consisting of two variables about the target function, we optimize it through alternating iteration.

### 2.7.1 Update $W$

Firstly, the derivative of $f(\theta)$ by fixing $C$ is written as:

$$\nabla f(W) = XX^T W - XY + \lambda_4 WL_p + \lambda_5 WL_n$$  \hspace{1cm} (12)

The iteration of $W$ represents:

$$W^t = W_t + \frac{b_{l-1} - 1}{b_t} (W_t - W_{t-1})$$  \hspace{1cm} (13)

$$W^{t+1} = \text{prox}_\varepsilon \left( W^t - \frac{1}{L} \nabla f(W^t) \right)$$  \hspace{1cm} (14)

g(\theta) uses $l_1$ norm in regard to $W$, which can be calculated as by the following definition of soft threshold operation:

$$\text{prox}_\varepsilon (w_{ij}) = (|w_{ij}| - \varepsilon)^+ \text{sign}(w_{ij})$$  \hspace{1cm} (15)

### 2.7.2 Update $C$

The gradient of $f(C)$:

$$\nabla f(C) = \lambda_3 Y^T L_x YC$$  \hspace{1cm} (16)

The iteration of $C$ is as follows:

$$C^t = C_t + \frac{b_{l-1} - 1}{b_t} (C_t - C_{t-1})$$  \hspace{1cm} (17)

$$C^{t+1} = \text{prox}_\varepsilon \left( C^t - \frac{1}{L} \nabla f(C) \right)$$  \hspace{1cm} (18)

g(\theta) uses $l_1$ norm in regard to $C$, which can be calculated as:

$$\text{prox}_\varepsilon (C_{ij}) = (|C_{ij}| - \varepsilon)^+ \text{sign}(C_{ij}) \cdot (a)_+ = \max(a, 0)$$  \hspace{1cm} (19)

we demonstrate the estimation of $l_f$ in order to verify its continuity by the following equation:
Algorithm 1: Multi-label learning with label-specific feature selection via locally positive correlation and negative correlation

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Initialization: ( W, C, b );</td>
</tr>
<tr>
<td>2</td>
<td>According to (4) and (5) calculate positive label matrix ( L_p ) and negative label ( L_n );</td>
</tr>
<tr>
<td>3</td>
<td>According to (21) calculate ( l_f );</td>
</tr>
<tr>
<td>4</td>
<td>According to (12) and (16) calculate ( \nabla f(W) ) and ( \nabla f(C) );</td>
</tr>
<tr>
<td>5</td>
<td>According to (13) update ( W' );</td>
</tr>
<tr>
<td>6</td>
<td>( G_\alpha^t \leftarrow W' - \frac{1}{l_f} \nabla f(W', C) );</td>
</tr>
<tr>
<td>7</td>
<td>( W_{t+1} \leftarrow \text{prox}<em>{\frac{\lambda}{l_f}}(G</em>\alpha^t) );</td>
</tr>
<tr>
<td>8</td>
<td>According to (17) update ( C' );</td>
</tr>
<tr>
<td>9</td>
<td>( G_\alpha^t \leftarrow C' - \frac{1}{l_f} \nabla f(W_t, C') );</td>
</tr>
<tr>
<td>10</td>
<td>( C_{t+1} \leftarrow \text{prox}<em>{\frac{\lambda}{l_f}}(G</em>\alpha^t) );</td>
</tr>
<tr>
<td>11</td>
<td>( b_{t+1} \leftarrow \frac{1 + 4\sqrt{b_t^2 + 1}}{2} );</td>
</tr>
<tr>
<td>12</td>
<td>( t + 1 \leftarrow t );</td>
</tr>
<tr>
<td>13</td>
<td>until convergence, return ( W^* \leftarrow W_t ) and ( C^* \leftarrow C_t );</td>
</tr>
</tbody>
</table>

3. Experiment

we choose ten common multi-label datasets on the six different feature algorithms so as to prove the effectiveness of algorithm.

3.1 Dataset

In order to conduct a comprehensive performance evaluation, we used 10 real-world multi-label datasets in the experiment: Birds, Business, Computers, Education, Emotions, Entertainment, Creation, Science, Society, Yeast.
detailed information about datasets is in the following table.

Table 1 Description of multi label dataset

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Dataset</th>
<th>Instance</th>
<th>Training</th>
<th>Test</th>
<th>Feature</th>
<th>Label</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Birds</td>
<td>5000</td>
<td>2000</td>
<td>3000</td>
<td>462</td>
<td>26</td>
<td>Text</td>
</tr>
<tr>
<td>2</td>
<td>Business</td>
<td>645</td>
<td>322</td>
<td>323</td>
<td>260</td>
<td>19</td>
<td>Audio</td>
</tr>
<tr>
<td>3</td>
<td>Computers</td>
<td>5000</td>
<td>2000</td>
<td>3000</td>
<td>681</td>
<td>33</td>
<td>Text</td>
</tr>
<tr>
<td>4</td>
<td>Education</td>
<td>5000</td>
<td>2000</td>
<td>3000</td>
<td>550</td>
<td>33</td>
<td>Text</td>
</tr>
<tr>
<td>5</td>
<td>Emotions</td>
<td>593</td>
<td>391</td>
<td>202</td>
<td>72</td>
<td>6</td>
<td>Music</td>
</tr>
<tr>
<td>6</td>
<td>Entertainment</td>
<td>5000</td>
<td>2000</td>
<td>3000</td>
<td>640</td>
<td>21</td>
<td>Text</td>
</tr>
<tr>
<td>7</td>
<td>Recreation</td>
<td>5000</td>
<td>2000</td>
<td>3000</td>
<td>606</td>
<td>22</td>
<td>Text</td>
</tr>
<tr>
<td>8</td>
<td>Science</td>
<td>5000</td>
<td>2000</td>
<td>3000</td>
<td>743</td>
<td>40</td>
<td>Text</td>
</tr>
<tr>
<td>9</td>
<td>Society</td>
<td>3782</td>
<td>2546</td>
<td>1236</td>
<td>1079</td>
<td>22</td>
<td>Text</td>
</tr>
<tr>
<td>10</td>
<td>Yeast</td>
<td>2417</td>
<td>1499</td>
<td>918</td>
<td>103</td>
<td>14</td>
<td>Biology</td>
</tr>
</tbody>
</table>

3.2 Experimental result

The paper utilizes four evaluations which contains Hamming-Loss; Ranking Loss; Average Precision and One Error (Arrows upwards indicate good, but arrows downwards indicate poor) to conduct the comparability of the proposed algorithm LLPCN-MIFS through in comparison with six different feature selection algorithms which are respectively MDDM_proj; MDDM_spc; MDFS; CMFS; LSF-CI and SCLS. The detailed information is analyzed as follow as:

Figure 3.2.1 Comparative experimental results of six algorithms on ten datasets in terms of Hamming-Loss (↓)

![Hamming-Loss](image1)

Figure 3.2.2 Comparative experimental results of six algorithms on ten datasets in terms of AP (↑)

![Average Precision](image2)
From the results of the above figure, we further analyze that our proposed algorithm LLCPN-MIFS has the better performance compared with traditional algorithms on most of datasets. We still need to respectively describe results on four evaluation metrics. In terms of Hamming-Loss, the proposed method relatively better performance on six datasets except Computers, Yeast, Birds and Education which shows CMFS on Birds, Computers and Yeast as well as SCLS on Education perform better than it. In terms of Average-Precision, LLCPN-MIFS exceeds comparable methods on most datasets broadly except Education. That shows our proposed method significantly performs well on certain datasets.
especially. In terms of AUC, the proposed method have a good performance but MDDM_proj, MDDM_spc performs well on dataset Emotions and SCLS have the goodness on datasets Education. In terms of Ranking-Loss, our proposed algorithm reflects barely satisfactory on the rest of seven datasets in comparison with CMFS, SCLS that both perform well on datasets Birds, Computers and Education. Based on observed results, it can seen that our proposed method actually behaves well than other algorithms which only focus on common feature selection and global correlation on ten datasets and achieve the effectiveness of dimensionality reduction through combining label-specific feature with local positive and negative correlation. Though MDFS and CMFS also explore local correlation, they fail to be in-depth analysis of label relevance in label correlation, which denotes the correlation among labels is strong or weak. Therefore, experimental results prove that the proposed algorithm is actually superior to other six typical feature selection algorithms when considering label-specific feature and local positive and negative correlation and enhance the accuracy of classification in multi-label learning.

4. Conclusion

In multi-label data, feature selection is so important in dimensionality reduction that further improves algorithm performance in particularly exploring the characteristics of each label. Some features are only related to a part of labels, which means it is not rational to use all labels in each iteration. Additionally, label correlation plays an indispensable role in reflecting the complex relationship between labels. Next, how to effectively analyze feature redundancy in selected discriminative label-specific features and integrate it with the reconstruction from original feature space to label space and correlation between labels into a unified framework is worthy considering.

5. Reference


