Learning Label Correlations for Multi-Label Online Passive Aggressive Classification Algorithm

ZHANG Yongwei
Nanjing NARI Intelligent Transport Technology Co., Ltd., Nanjing 210061, Jiangsu, China

© Wuhan University 2024

Abstract: Label correlations are an essential technique for data mining that solves the possible correlation problem between different labels in multi-label classification. Although this technique is widely used in multi-label classification problems, batch learning deals with most issues, which consumes a lot of time and space resources. Unlike traditional batch learning methods, online learning represents a promising family of efficient and scalable machine learning algorithms for large-scale datasets. However, existing online learning research has done little to consider correlations between labels. On the basis of existing research, this paper proposes a multi-label online learning algorithm based on label correlations by maximizing the interval between related labels and unrelated labels in multi-label samples. We evaluate the performance of the proposed algorithm on several public datasets. Experiments show the effectiveness of our algorithm.

Key words: label correlations; passive aggressive; multi-label classification; online learning

CLC number: TP391

0 Introduction

The multi-label classification problem is an emerging machine learning paradigm[1]. A single sample could be related to several classes simultaneously, and the class labels are no longer mutually exclusive. Nowadays, there are lots of real-world applications, such as text categorization, automatic scene classification, gene function prediction, and music emotions annotation[2]. In recent years, with in-depth research on multi-label classification problems, many multi-label classification methods have been proposed. Most authors agree with the taxonomy that differentiates between two main approaches in solving multi-label classification problems: problem transformation methods and algorithm adaptation methods[3]. Despite being studied extensively, most existing studies of multi-label classification are restricted to batch learning, which requires that the whole training dataset is read into memory and processed for the decision model at once. However, these batch learning methods, especially for classifying large-scale datasets, will consume a significant amount of time and space resources.

Online learning algorithms represent a family of quick and simple machine learning techniques, which generally make few statistical assumptions about the
datasets and are often used for classification problems. As discussed in this thesis, it is to update the model on each round without retraining. In general, online learning aims to incrementally learn some prediction models to make correct predictions on a stream of examples that arrive sequentially. Online learning is advantageous for its high efficiency and scalability for large-scale applications. It has been applied to solve online classification tasks in various real-world data mining applications. In the machine learning community, online learning has been actively studied, and various online learning algorithms have been proposed, including first-order and second-order online learning algorithms\(^3\). For example, the classical and popular first-order algorithms include perceptron\(^4\) and passive-aggressive (PA)\(^5\). Examples of second-order online learning algorithms include adaptive regularization of weights (AROW)\(^6\) and soft confidence-weighted (SCW)\(^7\). These methods can be explored to guide and enhance algorithm classification performance. Most aforementioned online algorithms, however, are online single-label algorithms. Aiming at these issues, we use online learning for large-scale learning tasks in multi-label classification problems.

In traditional online learning classification, a learner is trained sequentially to predict the class labels of a sequence of samples as precisely as possible\(^8\)-\(^10\). In this paper, we propose a multi-label online learning algorithm by maximizing the margin between the relevant and irrelevant labels in multi-label samples: learning label correlations for multi-label online passive-aggressive classification algorithm (MLRPA), which builds a sorted error set by predicting the pairs of labels and then updates its classifier model according to the size of error set. The experiment compares the MLRPA algorithm with four other algorithms on four multi-label datasets, according to four popular and indicative performance measures. The experimental results show that the MLRPA algorithm has good performance.

1 Related Work

1.1 Online Learning

Online learning reduces computation costs by incrementally updating classifiers in which new training samples are taken into account instead of training the classifier again from the beginning on the combined training samples. Online learning has been actively studied in the machine learning community, in which various online learning algorithms have been presented, including several first-order and second-order algorithms. One of the most well-known first-order online approaches is the perceptron algorithm, which updates the learning function by adding the misclassified example with a previous weight to update the current weight vectors. Recently, a number of online learning algorithms have been developed based on the criterion of maximum margin. One representative example is the PA algorithm. It updates the classification function when a new sample is misclassified, or its classification score does not exceed the predefined margin. Empirical studies showed that the maximum margin-based online learning algorithms are generally more effective than the perceptron algorithm. Recent years have seen a surge of studies on second-order online learning algorithms, which have shown that parameter confidence information can be explored to guide and improve online learning performance\(^3\).

Multi-label classification problem has always been an emerging machine learning paradigm. Because of many real-world applications, a few online multi-label classification methods have recently been studied\(^11\). For example, Ref. [12] proposed an extreme learning machine-based online universal classifier that is independent of classification type and can perform all three types of classification. Moreover, cost-sensitive dynamic principal projection (CS-DPP)\(^13\) resolves three important real-world issues: online updating, label space dimension reduction (LSDR), and cost sensitivity. Based on binary relevance, Refs.[14,15] presented online multi-label algorithms in which the dataset is divided into many single-label datasets to solve the multi-label classification problem. The online multi-label semi-supervised (OMLSS)\(^16\) introduced non-local labeling functions taking into account the topology of the network in the prediction of a label and improving the influence of the labeling strategy on the topology of the network in the multi-label case, using the labels to improve the synaptic links.

1.2 Label Correlations

To cope with the challenge of exponential-sized output space, it is essential to facilitate the learning process by exploiting correlations among labels\(^17\). Therefore, effectively exploiting the label correlation information is deemed crucial for the success of multi-label learning techniques\(^18\). According to the different ways and degrees of label correlation, the existing multi-label
classification methods can be roughly divided into three categories. First-order strategy: the task of multi-label learning is tackled in a label-by-label style, thus ignoring the co-existence of the other labels. The prominent merit of the first-order approach lies in its conceptual simplicity and high efficiency. Typical first-order strategy algorithms include Binary Relevance (BR) \(^{(19)}\) and multi-label \(k\)-nearest neighbor (ML-KNN) \(^{(20)}\). However, the effectiveness of the resulting approaches has low generalization performance due to the ignorance of label correlations. Second-order strategy: the task of multi-label learning is tackled by considering pairwise relations between labels. As label correlations are exploited to some extent by the second-order approach, the resulting methods can obtain good generalization performance. The most representative algorithm is RankSVM (Rank Support Vector Machine) \(^{(21)}\). High-order strategy: the task of multi-label learning is tackled by considering high-order relations among labels, such as imposing all other label’s influences on each label \(^{(22)}\). However, dealing with large-scale learning problems is difficult because of the high computational complexity \(^{(23)}\). Traditional multi-label online methods often need to request all class labels of each incoming sample. The correlations among labels are not considered in the multi-label online classification problem. We propose a multi-label online classification algorithm based on label correlations to compensate for this shortcoming.

2 Proposed Method

In the multi-label classification problem, a single instance could be related to several classes simultaneously, and the class labels are no longer mutually exclusive. For concreteness, we assume that there are \(q\) different possible labels and denote the set of all possible labels by \(Q = \{1, 2, \cdots, q\}\). At each step \(i\), the set of relevant labels \(Y_i\) of instance \(x_i\) is therefore a subset of \(Y\). We can say that label \(y\) is relevant to the sample \(x_i\) if \(y \in Y_i\). Otherwise, the set of irrelevant labels \(\tilde{Y}_i\) is a subset of \(Y\). This setting is often discussed in text categorization applications. As we all know, the learning paradigm and the analysis that we use in this paper belong to the mistake-bound model for online learning. The interval between sample labels in multi-label sorting is defined as follows:

\[
\min_{r \neq y} \{ w_r \cdot x \} - \max_{s \neq y} \{ w_s \cdot x \} \tag{1}
\]

On round \(i\), an online learning algorithm receives an instance \(x_i\). Given the instance \(x_i\), the learning algorithm outputs a ranking \(R_i = (\text{rank}(x_i, 1), \cdots, \text{rank}(x_i, q))\), where \(\text{rank}(x_i, 1)\) is induced by \(w_1 x_i\). The algorithm then receives the (correct) set of relevant labels \(y_i\). Given the feedback \(y_i\) and the predicted topic ranking \(R_i\), the algorithm computes the associated loss \(\ell_i = \text{loss}(y_i, R_i)\). If \(\ell_i\) is zero, the algorithm does not modify the model. Otherwise, it updates its label-ranking rule by modifying the set of weights \(w_1, \cdots, w_q\). The goal of the online label-ranking algorithm is to suffer a cumulative loss \(\ell\), that is competitive with a zero cumulative loss. The MLRPA algorithm we describe employs a refined notion of a mistake by examining all pairs of labels. The goal of this method is that the margin of an instance is positive, which means that all the sets of relevant labels are ranked higher than the non-relevant labels. Therefore, we define the multi-label loss function:

\[
\ell(w; (x, y)) = \begin{cases} 0, & \text{otherwise} \\
1 - \langle w - w_s - x, x \rangle, \langle w - w_r - x, x \rangle \leq 1 \end{cases} \tag{2}
\]

where \(w\) represents the weight vector of the instance. \(r\) and \(s\) denote the number of prediction errors in the relevant and irrelevant labels of instance \(x_i\), respectively. \(\langle \cdot \rangle\) represents the interval function between all relevant and irrelevant labels. It can be seen from the above loss that the sum of each pair of labels is calculated for the case where the ranking of irrelevant labels in the sample is higher than the relevant label. The classification performance of the algorithm model can be improved by fully considering the correlation among the labels.

To improve the classification performance of the algorithm, we add slack variables to the objective function. The MLRPA algorithm optimization model is as follows:

\[
w_{r,s,i} = \arg \min_{w_{r,s,i} \in \mathbb{R}} \frac{1}{2} \sum_{j \in \{1, \cdots, q\}} \|w_{r,s,i} - w_j\|^2 + C \sum_{i=1}^{N} \frac{1}{|Y_i|} \|w_{r,s,i} - y_i\| + \sum_{i=1}^{N} \zeta_{rs} \tag{3}
\]

where \(\zeta_{rs}\) is the slack variables and \(C > 0\), \(1 \leq i \leq N\). When the loss of function \(\ell_i = 0\), the algorithm does not update the weight value. When \(\ell_i > 0\), similar to the multi-class PA algorithm, the above optimization model is divided into two parts to meet the Lagrange function:

\[
L(w', \tau, \zeta) = \frac{1}{2} \|w' - w'_{i}\|^2 + C \sum_{i=1}^{N} \frac{1}{|Y_i|} \|w'_{i} - y_i\| + \sum_{i=1}^{N} \zeta_{rs}(w'_{s,i} - x_i - w'_{r,i}) - 1 + \sum_{i=1}^{N} \tau_{rs}(w'_{s,i} - x_i - w'_{r,i}) - 1 + \sum_{i=1}^{N} \zeta_{rs}
\]
where \( \lambda \geq 0 \) and \( \tau \geq 0 \) are Lagrange multipliers. \( \eta(\eta \geq 0) \) is slack variables. At time \( t \), taking the derivative of the above formula (4) and (5) concerning \( \mathbf{w}' \) and \( \mathbf{w}' \), we get
\[
\frac{\partial L}{\partial \mathbf{w}'} = 0 \Rightarrow \mathbf{w}_{r,s} = \mathbf{w}'_{r,s} + \mathbf{m}_s \tau_{r,s} \mathbf{x}_r
\]
and
\[
\frac{\partial L}{\partial \mathbf{w}'} = 0 \Rightarrow \mathbf{w}_{r,s} = \mathbf{w}'_{r,s} + \mathbf{m}_s \tau_{r,s} \mathbf{x}_r
\]
where \( \mathbf{m}_s \geq 0 \) indicates the sum of squares of the number of pairs of labels predicted to be unordered.

Taking the derivative of \( L(\cdot) \) with respect to \( \xi \) and setting it to zero, we get
\[
\frac{\partial L}{\partial \xi_{r,s}} = 0 \Rightarrow -\lambda_{r,s} - \tau_{r,s} = 0 \Rightarrow \tau_{r,s} = \lambda_{r,s}
\]
and
\[
\lambda_{r,s} = \frac{C}{|Y_r||\overline{Y_r}|} - \xi_{r,s} = \frac{C}{|Y_r||\overline{Y_r}|} - \lambda_{r,s}
\]
the Karush-Kuhn-Tucker (KKT) conditions confine \( \lambda \) to be non-negative, so we conclude that
\[
\lambda_{r,s} = \min \left\{ \frac{C}{|Y_r||\overline{Y_r}|}, \lambda_{r,s} \right\}
\]
Finally, taking the derivative of \( L(\cdot) \) with respect to \( \tau \) and setting it to zero, we get
\[
\frac{\partial L}{\partial \tau_{r,s}} = 0
\]
Plugging the Eq. (9) back into Eq.(4),
\[
\tau_{r,s} = \min \left\{ \frac{C}{|Y_r||\overline{Y_r}|}, \lambda_{r,s} \right\}
\]
Simplifying the above formula, we get
\[
\tau_{r,s} = \min \left\{ \frac{C}{|Y_r||\overline{Y_r}|}, \lambda_{r,s} \right\}
\]
Finally, we summarize the proposed method in the following Algorithm 1. In Algorithm 1, the classifier weights matrix and parameters are initialized. At time \( t \), the classifier accepts a sample and then predicts the relevant and irrelevant labels of the sample. Then, the MLRPA algorithm calculates the loss function based on the obtained real and predicted labels. Finally, the MLRPA algorithm updates the weight of the classifier according to whether the loss function is greater than zero.

**Algorithm 1** MLRPA learning algorithm

1. Initialize weight matrix \( \mathbf{W} \).
2. for \( t = 1, \cdots, T \) do
3. Receives an incoming sample \( \mathbf{x}_t \).
4. Calculate the learner \( \mathbf{W}_t, \cdot \).
5. Get the real relevant labels \( y_t \) and irrelevant labels \( \hat{y}_t \).
6. The loss \( \ell_t \) is calculated according to Eq. (2).
7. if \( \ell_t > 0 \) then
8. for \( k = 1, \cdots, q \) do
9. According to Eq. (6) and Eq. (7), update weight vector
10. end for
11. end if
12. end for
13. end for

### 3 Experiments

#### 3.1 Four Datasets and Four Existing Methods

We use four benchmark datasets: Corel5K(image), Corel16k001(image), Delicious(text), and Tmc2007(text) to validate our proposed method. Some valuable statistics of these datasets are provided in Table 1. In this study, we select three different multi-label online classification algorithms: perceptron algorithm based on multi class and multi label feedback (MMP)(30), BR-PA, and BR-perceptron (BR-PE), and a multi-label offline classification algorithm: ML-KNN. The MMP algorithm is based on a perceptron-based multi-label online classification algorithm, which considers label correlations. The BR-PA algorithm and BR-PE algorithm are based on the

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Instances</th>
<th>Features</th>
<th>Classes</th>
<th>Cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel5k</td>
<td>4500</td>
<td>500</td>
<td>499</td>
<td>174</td>
</tr>
<tr>
<td>Corel16k001</td>
<td>5188</td>
<td>1744</td>
<td>500</td>
<td>153</td>
</tr>
<tr>
<td>Delicious</td>
<td>12920</td>
<td>3185</td>
<td>500</td>
<td>983</td>
</tr>
<tr>
<td>Tmc2007</td>
<td>21519</td>
<td>7077</td>
<td>49060</td>
<td>22</td>
</tr>
</tbody>
</table>
problem decomposition strategy of the multi-label online classification algorithm. The BR-PA algorithm transforms the multi-label classification problem into two types of online PA algorithm, and the BR-PE algorithm is similar to the BR-PA algorithm, which transforms the multi-label classification problem into a binary online perceptron algorithm. The ML-KNN is derived from the traditional k-Nearest Neighbor (KNN) algorithm. For each unseen instance, its k nearest neighbors in the training set are first identified. After that, based on statistical information gained from the label sets of these neighboring instances.

In BR-PA and our MLRPA, two parameters are set as $C = 1$. For the ML-KNN algorithm, the smoothing factor $s = 1$ and the nearest neighbor $k = 10$. In order to ensure that the datasets are realistic, we randomize the dataset. In the experiment, we use online learning in the training dataset and offline in the test dataset. All the experiments were executed with about 10 runs of different random permutations for each dataset. All the results were reported by averaging over these 10 runs.

3.2 Performance Comparison

First, we have selected four large-scale multi-label datasets on four online algorithms in training datasets, where Fig. 1 and Fig. 2 show the average precision and one error values. Figure 1 shows the average precision of the four different algorithms for multi-label online learning. We observe that with the increase of the training sample set, the average precision of the four different online classification algorithms also increases. The MLRPA algorithm proposed in this paper has an average precision value better than other algorithms. In addition, on the image datasets (Corel16k001), the performance of the online algorithm considering label correlation is better than the algorithm based on the decomposition strategy. Figure 2 summarizes the one error performance of the four diverse algorithms. With the increase of the training sample set, the proposed MLRPA algorithm has fewer errors than other algorithms. The performance is more prominent, especially in the image dataset (Corel16k001). Experimental results show that the MLRPA algorithm based on label correlation signifi-

![Fig. 1 The average precision analysis of four algorithms on four datasets](image-url)
cantly outperforms the online algorithm based on problem decomposition when dealing with multi-label online classification. MLRPA algorithm is also more suitable for dealing with large-scale dataset classification problems.

To further evaluate the online MLRPA algorithm, we compare the MLRPA algorithm with online and offline algorithms for classification performance in test datasets. Tables 2-5 show detailed metrics of four different algorithms for multi-label online learning on four testing datasets in an offline learning way. The experiment results are listed in Tables 2-5 according to four criteria mentioned in this paper, where the optimal value of each dataset among four methods is highlighted in boldface. To compare these methods, we sort four methods on a single metric with 1-5 and calculate the average rank of each method over all four datasets, as listed in the last rows of Tables 2-5.

The results in Table 2 to Table 5 show that our proposed MLRPA algorithm achieves good classification performance for all datasets under four different evaluation criteria. Table 2 shows one error value on four datasets. The MLRPA works the best on four datasets. From the ranking loss in Table 3, the MLRPA performs the best on four datasets and achieves the first rank. In Table 4 for average precision, the four best values are from our MLRPA. At the same time, the MLRPA obtains the best rank. According to Table 5, the MLRPA works the best on all datasets and receives the best rank. In addition, we observe that our MLRPA outperforms the ML-KNN and MMP on all datasets. The experimental results also show that the MLRPA algorithm based on label correlation is more suitable for large-scale multi-label classifi-

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MLRPA</th>
<th>MMP</th>
<th>BR-PA</th>
<th>BR-PE</th>
<th>ML-KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel5k</td>
<td>48.8(1)</td>
<td>67.2(4)</td>
<td>54.6(2)</td>
<td>57.5(3)</td>
<td>70.3(5)</td>
</tr>
<tr>
<td>Corel16k001</td>
<td>66.6(1)</td>
<td>72.9(2)</td>
<td>77.4(5)</td>
<td>76.8(4)</td>
<td>74.5(3)</td>
</tr>
<tr>
<td>Delicious</td>
<td>48.8(1)</td>
<td>67.3(5)</td>
<td>54.7(2)</td>
<td>57.6(3)</td>
<td>61.9(4)</td>
</tr>
<tr>
<td>Tmc2007</td>
<td>30.0(1)</td>
<td>31.8(2)</td>
<td>32.4(3)</td>
<td>34.7(4)</td>
<td>92.9(5)</td>
</tr>
<tr>
<td>Average rank</td>
<td>1.0</td>
<td>3.25</td>
<td>3.0</td>
<td>3.5</td>
<td>4.25</td>
</tr>
</tbody>
</table>

Note: The data represents the mean and rank
take online second-order algorithms into account in multi-label classification problems.

### References


