An Improved ML-kNN Multi-label Classification Model Based on Feature Dimensionality Reduction

Zhi-qiang Li¹, Shuai-yi Cao¹ and Hong-chen Guo²,

¹School of Software, Beijing Institute of Technology, Beijing, China
²Network Information Technology Center, Beijing Institute of Technology, Beijing, China

*Corresponding author

Keywords: Multi-label classification algorithm, Feature reduction, LDA, Label relevance.

Abstract. ML-kNN cannot be used in real-time classification because of the huge computational cost, time and space resources. Therefore, this paper proposes a LDA-ML-kNN multi-label classification model based on feature dimensionality reduction. Firstly, we use LDA to extract the features of text data, and reduce the dimension of the extracted data matrix, then use data to train the classifiers and add the correlation between the labels. Finally, get the multi-label classification results.

In this paper, the experiment results show that the improved model has a significant improvement in the reduction of computational complexity. When the model has a certain improvement in the average accuracy rate, the compression time and space complexity are greatly reduced. And this model has great practical significance for real-time classification and processing of massive data.

Introduction

For the text data classification in the real scene, it mainly deals with multi-label text data. Multi-label text data refers to a text data containing multiple labels which are different from each other.

The multi-label classification problem needs to solve the problem of the relevance and co-occurrence between labels. In reality, the labels are not completely independent of each other, which have certain relevancies. So sorting out and adding the correlation between the labels for multi-label classification algorithm makes sense.

The solution based on algorithm transformation proposed by Zhou¹ in 2005 is ML-kNN. ML-kNN is derived from the K-nearest neighbor algorithm. Firstly, for each test instance, we use KNNs to identify the training set. Then, based on the statistical information obtained from the label sets of these neighboring instances, we calculate the number of neighbor instances belonging to each possible class. Finally, maximum a posterior principle is used to determine the label set for the test instance. The ML-kNN algorithm also has many improvements, mainly in adding label relevance. However, the ML-kNN algorithm is computationally intensive and does not have the dependency of label, and cannot be applied in real time.

The main content of this thesis is to classify the real text data by using the improved learning algorithm. Due to the huge amount of real text data, the topic model and dimension feature are extracted by LDA clustering method. And then use multi-label classification algorithm for text classification. In order to improve the classification efficiency and accuracy, this paper introduces the relevancies between the labels into the multi-label classification algorithm².

Related Works

There are several ways to improve the ML-kNN algorithm, relevance is added through different forms between the labels.
In 2014, Wang Xiao et al. proposed an improved ML-kNN multi-label classification algorithm based on label relevance. It determines whether the instance contains the label by calculating the relevance between any two labels in the label set.

In 2012, Wang Chunyan put forward a weighted improved ML-kNN multi-label classification algorithm. Each data label is clustered, and the distance between the sample points in the same category is calculated as the density of the class.

The iterative ML-kNN multi-label classification algorithm is proposed by Xu Zhaoyang in 2012. By changing the classification function of the original ML-kNN, it increases the probability of the original subset of the collection to classify. And finally after several iterations, the labels set will be constantly.

LDA\[^3\] is a three-layer Bayesian model structure. In LDA, Dirichlet first gives the topic distribution by using dirichlet prior, and then randomly picks up the topic from the distribution of the given topic. Finally, the words are randomly sampled from the corresponding word distribution of the determined topic, and the topic is extracted\[^4\]. The semantic relation between words and words is calculated, and the correlation of label matrix is calculated\[^5\]. By increasing the correlation between labels, the classification efficiency of ML-kNN is improved.

Methods

The model proposed in this paper is divided into two parts. Firstly, LDA is used to extract the feature and reduce the dimension of the feature. Then put the data into the ML-kNN to classify. Secondly, in the process of classification, the relevance between the labels is added, and finally the desired multi-label classifier is obtained.

Classification by ML-kNN

For each test instance $t$, ML-kNN firstly identifies its KNNs $N(t)$ in the training set. Let $H^l_b$ be the event that $t$ has label $l$, while $H^0_b$ be the event that $t$ has not label $l$. Furthermore, let $E^j_l$ ($j \in \{0, 1, \ldots, K\}$) denote the event that, among the KNNs of $t$, there are exactly $j$ instances which have label $l$. Therefore, based on the membership counting vector $\vec{C}_t$, the category vector $\vec{y}_t$ is determined using the following MAP principle:

$$\vec{y}_t(l) = \arg\max P(H^1_b) \left( E^1_l(\vec{C}_t(l)) \big| H^1_b \right) (b \in \{0, 1\})$$

Add Relevance between Labels

By analyzing the co-occurrence probability between words in text, we can find the relevance of the semantics in the text. Therefore, we can analyze the relevance between labels, calculating the probability of relevance between labels.

The calculated probability matrix combines with the parameters $\alpha$ ($0 \leq \alpha \leq 1$). $Y1(i), Y0(i)$ are outcomes of ML-kNN, $YY1(i), YY0(i)$ are final outcomes.

$$y1(i) = \alpha \times Y1(i) + (1 - \alpha) \times YY1(i)$$

$$y0(i) = \alpha \times Y0(i) + (1 - \alpha) \times YY0(i)$$

Experiment

Dataset

To evaluate the performance of our architecture, we have used the BIT Web Crop. It has 4223 instances. In this experiment, it has 6 categories, which have 10 or 25 labels and 50, 100 or 150 features.
Chose Weight Parameter

In this paper, we use the relevance matrix and the weight parameter to add the labels relevance into the ML-kNN algorithm. Experiments will be iterated 20 times to observe the efficiency of the classifier. When $\alpha = 0.75$, the classification of the model is the best. So in the following experiment weight parameter is 0.75.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Label</th>
<th>Feature</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
<td>10</td>
<td>50</td>
<td>4223</td>
</tr>
<tr>
<td>Category 2</td>
<td>10</td>
<td>100</td>
<td>4223</td>
</tr>
<tr>
<td>Category 3</td>
<td>10</td>
<td>150</td>
<td>4223</td>
</tr>
<tr>
<td>Category 4</td>
<td>25</td>
<td>50</td>
<td>4223</td>
</tr>
<tr>
<td>Category 5</td>
<td>25</td>
<td>100</td>
<td>4223</td>
</tr>
<tr>
<td>Category 6</td>
<td>25</td>
<td>150</td>
<td>4223</td>
</tr>
</tbody>
</table>

Experiment Steps

In this paper, the feature dimension of text data is extracted by LDA, and the feature dimension reduction of data text is carried out at the same time. The aim is to reduce the matrix computation and running time of ML-kNN algorithm. The goal of real-time computation is achieved by reducing the execution time without reducing the average accuracy of the algorithm itself and other evaluation indexes. Below are two groups of different data sets for comparative experiments, the number of tags were 10, 25, the number of features were 50, 100, 150.

Results

As we can see from the Figure below, the average accuracy does not decrease in the process of decreasing the number of features after the change of the number of features, which is noticeably improved when the number of features is 100. While the feature accuracy is equal to 50, the average accuracy decreases slightly but does not decrease to the starting size. So the number of tags decreased at the same time, accuracy has improved, indicating that the data matrix in the same conditions, the number of labels will reduce the classifier has more sufficient data classification, which means that you can choose a relatively small number of labels to improve the accuracy of classification.

![Figure 1. Hamming Loss.](image)

As we can see in Figure 1, when we use LDA-ML-kNN model the hamming loss is lower than ML-kNN. It can make hamming loss 2% lower and we can get a balance classifier in this way.

As we can see above Figure 2 and Figure 3, in both models the average precision and time are almost same, so we can get a similar classifier but a more balance one.
Conclusion

The experimental part of this paper validates the efficiency of LDA-ML-KNN model. On the one hand, the contrastive experiments were performed by comparing the classifier with the label relevance and the label correlation. The results show that the classification accuracy of the multi-label classifier is more balanced and the classification result is more reasonable under the condition that the mean accuracy of the classifier is not decreased obviously. On the other hand, after reducing the characteristic dimension of the matrix, the running time of the model is reduced obviously, which can save more than 10% of the time, meanwhile the average accuracy is almost the same and other evaluation criteria become more reasonable. This further shows that the model saves computation time and saves the computational space. In this experiment, we can reduce the dimensionality of the high dimensional data and add the correlation between the labels during the training process, and finally save the model running time.

References


