Research on Text and Image Classification Based on Multi-Label Algorithm

Shengxuan Xu[1], Yuanyuan Tan[2] and Kaiyu Zhang[3]

[1]College of Information and Electrical Engineering, China Agriculture University, Beijing, 100083, China

[2]School of Information Technology and Management, University of International Business and Economics, Beijing, 100029, China

[3]College of Mathematics, Chongqing University of Posts and Telecommunications, Chongqing, 400000, China

Keywords: label probability matrix, multi example multi label learning, weak label, multi label learning

Abstract: Due to the excellent performance of multi label learning in the process of solving the problem that a single object may have multiple associated categories, the researchers have attracted extensive attention. In this paper, according to the information and needs of different users, we use data mining to analyze the correlation between many label features, and use the sample clustering information to adjust the similarity matrix of weak label samples. We propose an evaluation standard classifier reliability to measure the classifier, and construct the similarity matrix based on k-means. In addition, we introduce different levels of multi instance learning algorithm as our classifier model, and propose two kinds of classification distance: the minimum classification distance and the average classification distance. Finally, we apply the model to the natural scene image classification and text classification, and compare the method proposed in this paper with the common samples. It is found that under the same training samples, the method in this paper can achieve better classification performance in various evaluation indexes.

1. Introduction

Multi label learning can be called a paradigm of supervised learning. Different from the two classification problem, multi label learning allows one object to belong to multiple categories, but also different from the multi classification problem, multi label learning allows one object to belong to multiple categories at the same time. In most classification problems in the past, each data object can be represented by a data sample, and each sample can have multiple attributes, including a Category attribute. The training data is made up of samples. For data classification, first of all, we need to fully train and learn the training samples to get a model that can predict the attribute value of the test samples; finally, we use this model to predict the tag of unlabeled data.

Generally, this kind of learning without polysemy is also called single label learning. In this problem, each object only corresponds to a single category label, that is, each sample only belongs to one category. In this context, as a learning framework, multi label learning is proposed to describe the ambiguous objects in the real world. At first, multi label learning is used to describe the ambiguity problem in document classification [2, 3]. For example, in document classification, the same article may correspond to multiple keywords, such as "fruit" and "fitness". In the traditional two classification and multi classification problems, only one label can be selected as the attribute value of the category when training the sample data. All the labels here belong to the same label set. In multi tag learning, the sample data can select several tags in the tag set as their own category attribute values. After training, the Category attribute of each sample is a subset of tags. At present, the application fields of multi label learning are increasing with the continuous emergence of multi label problems, such as text classification, gene function analysis [4], emotion analysis [5], image analysis [6] and other fields.

2. Weak label learning

Tags with wrong or incomplete tags are called weak tags [7]. As the name implies, each sample of weak label data may only have a part of the corresponding category label, or even may not have the corresponding category label. In the real world, the problem of missing tags in training samples can be seen everywhere, and there are more and more weak tags in data sets, which brings great challenges to the classification performance of multi tag data. If the existing methods are directly applied to the weak tag data, the tags of the training samples will be correct and complete by default, and the tags that are not in the training samples must not belong to this sample, which will also bring noise data. In order to ensure the accuracy of the algorithm, it is necessary to de noise the data first, and the simplest and direct way is to delete the noise data. In this way, the number of training samples will be greatly reduced, and the performance of learning model depending on training samples will be correspondingly reduced, which wastes a lot of labeled data. On the contrary, if we can effectively use the weak label data, we will improve the performance of multi label learning to a certain extent.

From different perspectives, weak label data can be either marked data or unmarked data. The key is whether the selected reference label has been marked. Therefore, weak label data can be processed with reference to the solution of unlabeled data. In recent years, with the rapid increase of data volume, the weak tagging of data has gradually increased, and the research on weak tagging data has also attracted many researchers' attention.

3. Similarity matrix construction method based on Clustering

Graph construction is very important for graph based methods, and it also has a great influence on the final effect of multi label learning. Based on the semi supervised learning of graph, the constructed graph is usually a complete graph, that is, there are edges between any two samples in the graph. Therefore, how to set the edge weight of a graph is the key to the problem. The edge weight of a graph is used to represent the similarity between two samples. The commonly used measurement method is to use the Euclidean distance of features to calculate.

However, the similarity calculation method only based on the similarity between the two samples, ignoring the structural information of the whole sample data. As shown in Figure 1, the sample points in Figure 1 (a) can be divided into four categories; however, if only the distance between the sample points is considered, the sample points in the figure are likely to be classified as shown in Figure 1 (b). If the labeled samples are distributed near the boundary, it is not enough to only consider the distance between the samples in the process of edge weight calculation, otherwise the classification will develop in the wrong direction. In the tag propagation algorithm, the tags of the labeled samples are basically unchanged, and the tags are propagated to the unlabeled samples. In the case of clustering into several classes obviously, the impact of the same kind of labeled samples may not be as great as that of the nearby other classes of labeled samples, which may affect the final classification results. Theoretically, in some cases of sample distribution, without considering the global edge weight calculation method, the results may be relatively poor.

4. Graph construction method based on Clustering

4.1 Similarity matrix construction based on K-means clustering

The tag propagation algorithm is based on the sample graph, so we need to construct the corresponding sample graph before the tag propagation. In general, the weight of edge is determined by the similarity of features between two connected samples, and the weight decreases with the increase of distance. At present, the commonly used similarity measurement methods are mainly Euclidean distance, Mahalanobis distance, Makowski distance, correlation distance, etc. the Euclidean distance is chosen in this paper.

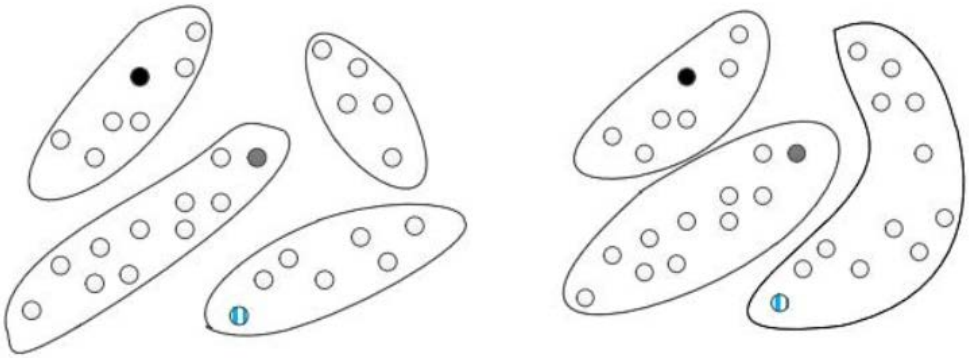

Figure 1. Impact of overall distribution of samples (a) Consider the overall distribution of samples (b) don't consider the overall distribution of samples

Through the analysis of the overall distribution of weak label data, on the basis of the original Euclidean distance weight calculation method, we use the clustering results of all samples to adjust the weight for the edge of the sample relationship graph. First, a sample graph is constructed on all training samples. Each sample is represented by a vertex in the graph. It can be a labeled sample or an unlabeled sample. The edge between vertex X_i and vertex $X_{j.}$ represents the similarity between sample X_i and sample $X_{j.}$ There are many ways to construct a graph, which can be completely connected graph or partially connected graph. This paper assumes that the constructed sample graph is KNN graph, each vertex has only k connected edges, that is to say, each sample only has edges with k neighbor samples, that is, there is similarity, and the similarity with other samples is ignored. The graph constructed in this way is a sparse graph, and the matrix used to represent the graph is also a sparse similarity matrix. The steps of building similarity matrix are as follows:

4.1.1 Calculate the weight of the edge

The label propagation algorithm propagates labels through the edges between vertices. The greater the weight of the edge, the more similar the two sample points are, the easier the label will pass. The weight of edge is calculated by the distance between two sample points. Generally, the formula of edge weight between any two samples Xi and sample XJ can be expressed as formula (1):

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) = \exp\left(-\frac{\sum_{d=1}^{D}\left(x_i^d - x_j^d\right)}{\sigma^2}\right) \tag{1}$$

Among them, d_{ij} is the Euclidean distance between sample x_i and sample x_j, and σ is used to control the weight of edge. The value of σ in this paper is the average value of Euclidean distance between samples. Accordingly, the smaller the Euclidean distance between two vertices in the graph is, the larger the weight w_{ij} is.

4.1.2 Generate probability transfer matrix T

The probability transfer matrix is used to represent the propagation probability of the tag between any two samples. From the weight matrix between samples, row normalization is obtained. For the convenience of subsequent calculation, the probability transfer matrix of the sample is represented by a $(l+u)\times(l+u)$ matrix T, and the generation formula of matrix t is shown in formula (2):

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}} \tag{2}$$

Where, Tij represents the transfer probability from vertex i to vertex j, that is, the probability of sample X_i to sample X_j tag propagation.

4.1.3 Adjusting probability transfer matrix by clustering

Based on K-means clustering, the input is training samples, including labeled samples and unlabeled samples, and the output is probability transfer matrix T. the specific algorithm flow is as follows:

(1) Data preprocessing, attribute reduction for all samples, high-dimensional samples into low-dimensional samples, convenient for the follow-up algorithm, the attribute reduction method used in this paper is mddm;

(2) K-means clustering is applied to the total sample space, and an n * n-dimensional clustering matrix C is constructed according to the clustering situation. Each matrix element indicates whether two samples belong to the same category, as shown in formula (3):

$$C_{ij} = \begin{cases} 1+\alpha, & if \quad i, j \in cluster \\ 1-\alpha, & if \quad i, j \in cluster \end{cases} \tag{3}$$

Among them, as a parameter to control the influence of clustering on graph construction. The smaller the value is, the smaller the influence is; on the contrary, the greater the influence is;

(3) A new similarity matrix W is constructed by combining the clustering matrix C with the original similarity matrix W. For each two samples x_i and x_j, there is $W_{ij}^{'} = W_{ij} \times C_{ij}$.

Normalize the row of the new similarity matrix, and construct a new probability transfer matrix. The formula is as follows (4):

$$T_{ij}^{'} = P\left(j \to i\right) = \frac{w_{ij}^{'}}{\sum_{k=1}^{l+u} w_{kj}^{'}} \tag{4}$$

4.1.4 Label propagation algorithm

(1) Construct the label matrix Y (where the initial value of labeled samples is 1, and the initial value of unlabeled samples is 0;

(2) Label communication Y' $\leftarrow$TY;

(3) Reset label with label sample in $Y_L^{'} = Y_L$:;

(4) Repeat until Y' converges.

5. Multi example and multi label learning

5.1 Description of multi example and multi label learning algorithm

In the multi example multi label learning framework, each sample data is represented in the following forms: input space for multiple examples, output space for multiple tags. The advantage of this representation is that it not only considers the ambiguity of the sample in the input space, but also reflects the diversity of the output space, so that the model can better match the actual scene, and then get a higher accuracy. The learning framework is shown in Figure 4. Currently, the model has been applied to scene classification, text classification and other practical applications.

In the multi instance and multi tag learning framework, an object is often represented as a package composed of multiple samples and associated with a set of tag categories. Suppose s represents the input space of the example, Y represents the collection space of the label category, and the data set $S = \left\{\left(X_1, Y_1\right), \left(X_2, Y_2\right), \cdots, \left(X_m, Y_m\right)\right\}$ is given.The goal of multi example and multi label learning is to learn the objective function $f_{MLML} : 2^X \to 2^Y$. Obviously, there is a strong correlation between multi example multi tag learning and multi example learning and multi tag learning framework.

5.2 Multi example and multi label learning algorithm

Because a picture has complex semantics, it can be given multiple tags, as shown in the figure below.

Figure 2. Schematic diagram of multi example and multi label learning

In addition, different region blocks of the image correspond to different subject category labels, and one image data sample in the natural scene image classification can be represented as a combination of multiple examples and labels. Similarly, in text classification, each document has many different topic labels, so it can be assigned with many different category labels; at the same time, a document can be divided into many different paragraphs, each paragraph corresponds to different topic category labels,

Therefore, in text classification, each document data sample can also be represented as a combination of multiple samples and multiple tags. In multi instance and multi tag learning, considering that the samples in the input and output space may have the characteristics of ambiguity, the samples that can better express the natural world can achieve better learning effect than single example and single example and multi tag learning.

6. Conclusion

Weak label learning problem is widely used in practical applications, including gene function analysis, text analysis, multimedia information analysis, medical diagnosis analysis and many other fields. It is one of the important research in academia, medical industry and government departments. However, the traditional multi label learning method can not be effectively applied to the problem of weak label data classification, because multi label learning will take the label of weak label samples as a whole, as long as the label is not marked, it does not exist. It will lead to deviation of final classification. In this paper, the graph based semi supervised multi label learning is improved to make it suitable for weak label data sets. The effectiveness of the improved algorithm is verified by the classification experiments on multiple open multi label datasets. The main research work is as follows:

1. In weak label data, each sample corresponds to multiple labels. The similarity measurement between samples is very important, which is related to the degree of label completion of weak label samples. The semi supervised learning method based on graph needs to construct graph in advance, using the weight of edge in graph to determine the spread degree of label, and using the distance between sample features to measure weight calculation, which is too localized. Therefore, combining with clustering algorithm, this paper constructs a graph model based on K-means for weak label data. When constructing a graph, the method uses the overall distribution of samples and the feature distance between samples to jointly determine the similarity matrix of samples, so as to improve the performance of tag completion and classification.

2. Combined with the characteristics of multi example and multi label learning, the multi example and multi label learning is transformed into multi example and single label learning. For multiple classifiers, we use the information of labeled samples and unlabeled samples to propose a criterion to measure the evaluation of classifiers - classifier reliability. According to the characteristics of multi instance single label learning, we apply active learning to it, and introduce two different levels of multi instance learning algorithm as our classifier model. In the learning process of classifier, we can actively and iteratively select the unmarked multi instance package which can improve the performance of classifier and put it into the training set for learning, effectively reducing the cost of training multi instance package finally, the performance of the classifier is improved.

References

[1] Tsoumakas G, Katakis I.Multi-Label Classification: An Overview [J]. International Journal of Data Warehousing and Mining, 2007, 3 (3): 1-10.

[2] Zhang Minling, Zhou Zhihua. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern recognition, 2007, 40 (7): 2038-2048.

[3] Mc Callum A. Multi-label text classification with a mixture model trained by EM [C].Working Notes of the AAAI'99 Workshop on Text Learning. Orlando: AAAI Press, 1999: 1-7.

[4] Barutcuoglu Z, Schapire R E, Troyanskaya O G. Hierarchical multi-label prediction of gene function [J]. Bioinformatics, 2006, 22 (7): 830-836.

[5] Trohidis K, Tsoumakas G, Kalliris G, et al. Multi-Label Classification of Music into Emotions [C] //Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008). Philadelphia: Springer Press, 2008: 325-330.

[6] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification [J]. Pattern recognition, 2004, 37 (9): 1757-1771.