*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 8 Issue 10 October 2020

IMCCC algorithm applied to multilabel classification

Yong-xianJin

(College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua Zhejiang 321004, China)

jyx@zjnu.cn

ABSTRACT

In CC algorithm, because the selection order of each tag is random, the tag chain relationship is not strong, as a result, the performance of the algorithm is not high.In order to improve the performance of classification algorithm, IMCCC algorithm is proposed in this paper. In the process of constructing tag sequence, the mutually exclusive relationship of tags is introduced to strengthen the connection between tags and improve the generalization ability of the algorithm. Experiments show that the performance of IMCCC algorithm is better than other commonly multi-label classification algorithms.

Keywords:Multilabel classification; Label space; Random; The mutex relationship

1.INTROUDUCTION

The problem of multi-tag learning is widespread in the real world. For example, in image classification [1], an image can often correspond to multiple themes, such as "beach" and "sunset". In document classification [2], a document can belong to more than one topic, such as "World Cup" and "football". As you can see, each sample corresponds to a set of tags. Multi-tag learning mainly studies how to build a classifier to accurately predict the tag set of unknown samples when samples have multiple category tags simultaneously.The traditional binary classification and multi-class classification problems can be regarded as special cases of multi-tag learning problems. At present, researchers have proposed a variety of methods to solve the problem of multi-tag learning, which are mainly divided into two types: algorithm adaptation and problem transformation [3]. Because of its simplicity and good application in most data sets, this paper mainly discusses the problem transformation method.

Binary Relevance(BR) method and Label Powset(LP) method are two of the most basic and commonly used methods in problem conversion method. Among them, BR method learns a number of two-category classifiers, each classifier only for a certain label classification. This method is simple and feasible, but it ignores the correlation between tags, and the prediction results are often not satisfactory. On the basis of BR, the Classifier in literature [4] is proposed in the Classifier Chain(CC), and multiple Chain structures are constructed. The so-called chain structure is to add the class attributes of the previous classifier to the feature attributes of the training set to build a new training set, while the latter classifier is built on the new training set. In this way, the dependency relationship between tags can be effectively utilized, but the order in which the classifier chain is built will affect the performance of the classifier.

The Tree-based Classifier Chain(TCC) algorithm proposed in literature [5] is improved on the basis of Classifier Chain algorithm, and it establishes Classifier Chain in a certain order. In the LP method, the multi-tag classification problem is decomposed into multi-class single-tag problem by treating each unique tag set in the multi-tag data set as a category. For a given test example, the multi-class LP classifier can predict the most likely category and then be converted into a tag set. Compared with the simple BR method, the LP method considers the tag correlation to some extent. However, with the increase of the label numbers and training sample instances, the possible categories increase proportionally, which increases the computational cost. On the other hand, there are too few training samples for individual categories, which makes learning difficult. And LP could only predict the tag combinations that appeared in the training set.

Literature [6] proposed the Pruned Problem Transformation(PPT) algorithm, which retained the tag set with occurrence times greater than the threshold, divided the tag set with occurrence times less, and established the LP classifier for the subsets after division. However, only the tag set that had appeared in the training set could be obtained when the instance was predicted.

Literature [7] proposed the Random K-Labelsets (RAKEL) algorithm, which randomly selects a part of the tag subset from the original set of tags each time, trains the corresponding classifier using LP method, and finally integrates multiple LP classifier to predict by voting. This method solves the shortage of data skew generated by LP through integration, and also takes into account the correlation between tags by randomly constructing tag subset. Due to the characteristics of random selection, the correlation between tags is not fully utilized, resulting in low classification accuracy and certain influence on generalization performance.

Aiming at the shortcomings of CC algorithm [4], this paper proposes IMCCC algorithm. In the process of constructing tag sequence, the mutually exclusive relationship of tags is introduced to strengthen the connection between tags and improve the generalization ability of the algorithm. Experiments show that the performance of IMCCC algorithm is better than other commonly used multi-label classification algorithms.

2.IMCCC ALGORITHMS

To understand the idea of mutual exclusion, take a simple example.

Label space: $L = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\}$,

Training sample: $D = \{(x_1, Y_1), (x_2, Y_2), (x_3, Y_3), (x_4, Y_4)(x_5, Y_5), (x_6, Y_6), (x_7, Y_7), (x_8, Y_8)(x_9, Y_9), (x_{10}, Y_{10})\}$,

Here, $Y_1 = \{\lambda_1, \lambda_2, \lambda_3\}$, $Y_2 = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$, $Y_3 = \{\lambda_3, \lambda_5\}$, $Y_4 = \{\lambda_2, \lambda_3, \lambda_5\}$, $Y_5 = \{\lambda_1, \lambda_2, \lambda_3\}$, $Y_6 = \{\lambda_2, \lambda_3, \lambda_4, \lambda_6\}, Y_7 = \{\lambda_3, \lambda_4, \lambda_5\}, Y_8 = \{\lambda_6\}, Y_9 = \{\lambda_3\}, Y_{10} = \{\lambda_6\}$.

Table *1* Multi-label data sets

The instance	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
x1	1	1	1	0	0	0

x2	1	1	1	1	0	0
x3	0	0	1	0	1	0
x4	0	1	1	0	1	0
x5	1	1	1	0	0	0
x6	0	1	1	1	0	1
x7	0	0	1	1	1	0
x8	0	0	0	0	0	1
x9	0	0	1	0	0	0
x10	0	0	0	0	0	1

From the table, we can get some implication relationships of tags: $\lambda_1 \to \lambda_2$, $\lambda_1 \to \lambda_3$, $\lambda_2 \to \lambda_3$, $\lambda_3 \to \lambda_4$,and also find mutually exclusive relationships between tags, such as tags: λ_1, λ_5, λ_6 cannot coexist.

In order to mine the exclusive relationship between tags in multi-tag dataset, the following multidimensional Euclidean metric formula is used:

$$d(\lambda_i, \lambda_j) = \sqrt{\sum_{k=1}^{n} (\lambda_{ik} - \lambda_{ik})^2}$$

Where, $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{in})$ is the *ith* label vector in the data set, and similarly, $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jn})$ is the *j*thlabel vector in the data set. When the distance between the labels is larger, the difference between the labels is greater, and the labels cannot coexist in the description of an instance. The smaller the distance, the smaller the difference between the tags, the more likely the tags will coexist in one instance. By this method, the mutually exclusive relationship between the tags can be mined, more tag information can be obtained from the training set, and the utilization rate of tag combinations with fewer occurrences can be improved, so as to improve the classification performance of the classification algorithm. And this idea is introduced into CC to design a new IMCCC.

Different from the random label selection method of CC algorithm, the IMCCC algorithm mainly selects the next label to be added to the training set based on the mutex exclusive relationship of the current label, so as to obtain a mutex exclusive label sequence chain. Specifically, you randomly select a tag λ_1' from the tag set, then find the mutex λ_2' based on the Euclidean distance, and then find the most distant mutex λ_3' based on the mutex λ_2', and so on until you find a mutex chain $\lambda_1', \lambda_2', \ldots \lambda_q'$. Then, when constructing a new training data set D_j from the original eigenvector, the previouslabel $\lambda_1', \lambda_2', ..\lambda_{j-1}'$ is added to the original eigenvector as an additional eigenvector to form a new eigenvector.In the classifier training stage, it is the same as the original CC, and it is only necessary to train according to the label sequence already found. In the prediction stage, it is still the result of synthesizing various classifiers. The following is the realization process of the mutex tag chain algorithm in IMCCCalgorithm.

Input: All label set L

Output: Mutex tag sequence chain No. $W_1, W_2..., Wq$

$$
\begin{aligned}
& for\, i \rightarrow 1\, to\, q \\
& \ for\ j \rightarrow 1\, to\, q-1 \\
& \{\, d_{ij} \leftarrow distance(\lambda_i, \lambda_j);\} \\
& W \leftarrow \varnothing; \\
& for\, k \rightarrow 1\, to\, q \\
& \{ \\
& w_k \leftarrow \arg\max(d_{ij}); \\
& remove\, d_{w_k} \\
& \}
\end{aligned}
$$

3. EXPERIMENT AND RESULT ANALYSIS

3.1 Experimental data set

All experimental data in this paper are emotions[8], Scene [9], Birds [10], and Medical [11]. Table 2 shows the detailed statistics.

Table *2*multi-label data sets and statistics

name	domain	instances	label	nominal	numeric	cardinality
emotions	sound	593	6	0	72	1.869
scene	image	2407	6	0	294	1.074
birds	audio	645	19	2	258	1.014
medical	text	978	45	1449	0	1.245

The four multi-label training sets of the experiment come from different fields. Table 1 shows the detailed label statistics of these data sets. Among them, name, domain, instances and label respectively represent the name, domain, numbers of samples and numbers of label. Nominal and Numeric are attributes of the tag, of which Nominal refers to the discrete characteristic attributes of the tag and the continuous characteristic attributes of the numeric tag. Cardinality is the average number of tags per sample instance, and that's a statistic. Emotions is a musical data set, consisting of more than 100 songs from seven different genres: Classical, reggae, Rock, Pop, hip-hop, electronic, and jazz. Three songs from each of the 233 albums were selected as samples. The image data set scene focuses on semantic retrieval objects of still scene images. For the scene image data set, its feature vector consists of the dimension of 294 LUV space color moments. This image is divided into 49 blocks in 7 squares. The mean and variance of the first and second moments are used to calculate each band. The end result is the eigenvector of the dimension. Medical is a text data set that describes Medical aspects. Birds is data that describes audio aspects.

3.2 Algorithm experimental results and analysis

In order to verify the effectiveness of the algorithm, **IMCCC** algorithm is compared with the **BR** algorithm mentioned above, **CC**[4] algorithm and KNN-based algorithm **ML-KNN**[12]. This experiment uses the ***emotions***, ***yeast,birds***, and ***medical*** four data sets. The***Accuracy***,***SubsetAccuracy***,***Ranking Loss***, ***Recall*** and ***F-measure*** are compared in order. Among them, the base classifier algorithms of BR and CC use the support vector machine classification algorithm in Weka, while the base classifier algorithm of RAKEL is LP, LP also uses the support vector machine classification algorithm. The label subset size K is set as 3, the number of models M is set as 2 times the number of labels, and the threshold value T is set as 0.5. In ML-KNN algorithm, K =10, smoothing=1. In the experiment, 5-fold cross validation method was used to verify each test data set, and all experiments were implemented in Java on Mulan [13] open source library.

Table 3***Accuracy***of each algorithm on 5 data sets

	BR	ML-KNN	CC	IMCCC
emotions	0.5156	**0.5322**	0.5281	0.5302
yeast	0.5013	**0.5199**	0.4877	0.5123
birds	0.5735	0.4925	0.5642	**0.5648**
medical	0.7578	0.5604	0.7743	**0.7773**

Table4***Subset Accuracy***of each algorithm on 5 data sets

	BR	ML-KNN	CC	IMCCC
emotions	0.2782	0.2899	0.2882	**0.3321**
yeast	0.1489	0.1862	0.1924	**0.1998**
birds	**0.4813**	0.4625	0.4720	0.4720

medical	0.6697	0.6922	0.6922	**0.6983**

Table5***Ranking Loss***of each algorithm on 5 data sets

	BR	ML-KNN	CC	IMCCC
emotions	0.2770	**0.1627**	0.2831	0.2842
yeast	0.3175	**0.1660**	0.3335	0.2951
birds	0.4813	**0.1285**	0.1928	0.1931
medical	0.0983	**0.0395**	0.0948	0.0924

Table6***F-measure***of each algorithm on 5 data sets

	BR	ML-KNN	CC	IMCCC
emotions	0.5922	**0.6131**	0.6087	0.6120
yeast	0.6114	0.6246	0.5890	**0.6265**
birds	**0.6070**	0.5051	0.5978	0.5983
medical	0.7876	0.5859	0.7876	**0.8038**

Table7***Recall***of each algorithm on 5 data sets

	BR	ML-KNN	CC	IMCCC
emotions	0.5929	0.6087	0.6321	**0.6430**
yeast	0.5789	0.5992	0.5719	**0.6328**
birds	**0.6201**	0.4940	0.6115	0.6079
medical	0.8025	0.5803	0.8137	**0.8147**

The experimental results of the IMCCC algorithm and other algorithms are shown in Table*3* to Table *7* respectively. Among them, the algorithm corresponding to the data in bold and black (the bigger the value, the better the performance) in the table is the best algorithm of the four algorithms in this data set. It is found from the experimental results that the performance of IMCCC algorithm is significantly improved. Compared with the original CC algorithm, most of the indexes are improved by more than 1%, and some indexes are even higher. At the same time, it can be seen from the above results that IMCCC algorithm is superior to other algorithms in multiple indexes and achieves good results in multiple data sets.

4. CONCLUSION

In this paper, IMCCC, a multi-label classification algorithm based on tag correlationis proposed, and corresponding experiments are carried out on multiple data sets.By comparing with many popular algorithms, the experimental results show that the proposed IMCCC algorithm achieves good classification effect on all data sets and different evaluation methods.

5. REFERENCES

[1]Liu Peng, YE Zhipeng, ZHAO Wei, et al. A Multi-level abstract Semantic decision image classification method [J]. Journal of Automation, 2015, 41(5):960-969.

[2] Zhang Jing, LI Deyu, Wang Suge, et al. Multi-marker text Classification based on robust Fuzzy Rough Set Model [J]. Computer Science, 2015, 42(7): 270-275.

[3] Li Sinan, Li Ning, Li Zhanhuai. Multi-label Data Mining Technology: Research Review [J]. Computer Science, 2013, 40(4):14-21.

[4] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification [J]. Machine Learning, 2011, 85(3):254-269.

[5]Fu Bin, WANG Zhihai. Multi-marker classification algorithm based on tree Dependent Structure [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(4):573-580.

[6]Read J. A pruned problem transformation method for multi-label classification[C]//Proceeding of New Zealand Computer Science Research Student Conference. Christchurch: Canterbury University,2008, 143-150.

[7] Tsoumakas G, Katakis I, Vlahavas I. Random k-labelsets for multilabel classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 1079-1089.

[8] Trohidis K, Tsoumakas G, Kalliris G, et al. Multi-label classification of music into emotions[J]. Eurasip Journal on Audio Speech & Music Processing, 2008, 11(1):325-330.

[9]Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(4):1757-1771.

[10]Briggs F, Lakshminarayanan B, Neal L, et al. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach [J]. Journal of the Acoustical Society of America, 2012, 131(6):4640-4650.

[11]Kajdanowicz T, Kazienko P. Multi-label classification using error correcting output codes [J]. International Journal of Applied Mathematics & Computer Science, 2012, 22(4):829-840.

[12] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7):2038 - 2048.

[13]Tsoumakas G, Spyromitros-Xioufis E, Vlahavas I P, et al. MULAN: A Java library for multi-label learning [J]. Journal of Machine Learning Research, 2011, 12(7):2411-2414.

