International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 04 – Issue 11 || November 2019 || PP. 69-76

Integrated Multi-label Classification Algorithm

Yong-xian Jin[1], Yi Ding[2]

[1]*College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua Zhejiang 321004, China*

[2]*Graphic Information Center of Zhejiang Normal University, Jinhua Zhejiang, 321004, China*

Abstract: Multi-label classification has become a research hotspot in data mining technology. Its research results are widely used in various fields, such as image and video semantic annotation, functional genome, music emotional classification and marketing guidance etc. So far, researchers have proposed a variety of multi-label classification algorithms. This paper aims to provide a review on integrated multi-label classification algorithm and emphatically introduces integrated multi-label classification algorithm IMLCA proposed by us. Finally, it summarizes the problems and challenges in the current research and looks forward to the development trend in this field.

Keywords: multi-label classification; integrated method; integrated multi-label classification algorithm

1. Introduction

Classification, as one of the important research topics in the field of data mining, is to analyze and study known category samples to predict unknown category samples. The traditional classification is mainly single-label classification, that is, each sample belongs to only one category. In practice, however, a sample can have multiple categories at the same time. For example, in document classification [1-4], each document may belong to multiple predefined topics. In image classification [5-8], each picture may have different semantics, and in bioinformatics [9-11], each gene may have multiple functions simultaneously. This led to the study of multi-label learning [12-14]. With the application of multi-label classification more and more widely, multi-label classification has gradually become a hot spot in the international machine learning field. After nearly ten years of development, multi-label classification technology has been widely applied in bioinformatics, medical diagnosis, scene classification, music emotion classification [15-18] and other fields. Therefore, the study of multi-label classification has important theoretical and practical significance. So far, researchers have proposed a variety of multi-label classification algorithms.

2. Types of Multi-Label Classification Algorithms

At present, a large number of multi-label classification algorithms have been proposed. The literature[19] divides them into three categories: problem conversion method, algorithm adaptive method and integration method.

Problem transformation method: the main idea of it is to transform the learning of multi-label data into one or more single-label data, which is not restricted by specific algorithms. Label powerset (LP) algorithm [20], and Binary relevance (BR) algorithm [21] are commonly used for problem transformation. Since there are many mature algorithms in single label data mining, such as support vector machine, Bayesian classifier, k-nearest neighbor method, etc., we can continue to use these traditional methods to solve multi-label problem by transforming multi-label problem into single label problem.

Algorithm adaptation method: it is to directly improve some existing single label data learning algorithm, so that it can adapt to the processing of multi-label data. For example, the method based on support vector, the method based on BP neural network, the method based on probability generation model, etc. These algorithms have been successfully applied in many fields such as document classification, bioinformatics and scene classification.

Integrated method: it is a multi-label classification method based on problem transformation or algorithm adaptive. Common integration methods based on problem transformation include RAKEL[22] and ECC[23]. RAKEL method is to randomly select appropriate samples from LP transformed data to train the corresponding classifier, and then integrate all classification results as multi-label classification results. The ECC method first integrates the classification results of all classifiers, and then uses threshold method to select relevant labels as the labels of unknown data. The ensemble method based on the algorithm adaptive method is usually based on the algorithm adaptive acquired multi-label classifier. For example, RF-PCT and RF-MLC4.5 are integrated methods that use PCTs and MLC4.5 as base classifiers, respectively. Due to the advantages of integrated methods in reducing over fitting and dealing with unbalanced data, it can be well applied to multi label classification problems and improve the overall performance. Therefore, the following focuses on the research status of integrated multi label classification algorithm.

3. Integrated Multi-Label Classification Algorithms

3.1 Rakel

The Random k-label-sets (RAKEL) algorithm is a combination of LP methods, which are trained as training sets for each LP classifier using a small subset of random labels in a collection of tags.First randomly select a subset of labels containing k labels, as the algorithm conversion method LP algorithm training set training, get m LP classifiers.Finally, by counting the number of votes for the corresponding labels for each sub class per, sort all labels according to the voting rate, and filter through thresholds to get the most relevant categories as the final predicted results for the sample to be predicted. The algorithmic process is as shown in Fig.1.

Input: Number of models m, size of labelset k, set of labels L, training set D
Output: An ensemble of LP classifiers h_i and corresponding k-labelsets Y_i
$R \leftarrow L^k$;
for $i \leftarrow 1$ **to** $\min(m, |L^k|)$ **do**
  $Y_i \leftarrow$ a k-labelset randomly selected from R;
  train an LP classifier $h_i : X \rightarrow P(Y_i)$ on D;
  $R \leftarrow R \setminus \{Y_i\}$;

Fig. 1. The ensemble production phase of RAKEL.

This method takes full account of the dependencies between labels and compensates for the lack of skew data that the LP method may produce. However, in order to achieve near optimal performance, the method must carry out internal cross-testing of input parameters such as subset size, model number, threshold, etc., and it is difficult to find the optimal parameters in the case of insufficient training samples.

3.2 ECC (Ensembles of Classifier Chains)

The ECC combined classifier chain algorithm is an improvement on the BR method, which uses the CC (Classifier Chains) [24] classifier chain to connect the n two classifiers produced by BR method into a chain, in response to the loss of information caused by BR method failure to take into account the link between labels. Each time the training sample passes through a wo classifier, the prediction results are added to the sample attribute vector and continue to be trained in the next two classifiers. However, because the different order of the two classifiers in CC has a great influence on the result, the ECC adopts the CC combination of different label sequences produced by several random series to reduce the adverse effects of the individual CC from the internal two classifier order ingestion problem. The following is the training and classification process (shown in Fig.2 and Fig.3)

```
TRAINING(D = {(x_1, S_1), ··· , (x_n, S_n)})
for j ∈ 1 ··· |L|
      do ▷ single-label transformation and training
         D' ← {}
         for (x, S) ∈ D
             do D' ← D' ∪ ((x, l_1, ··· , l_{j-1}), l_j)
         ▷ train C_j to predict binary relevance of l_j
         C_j : D' → l_j ∈ {0, 1}
```

Fig.2 The training process

```
CLASSIFY(x)
Y ← {}
for j ← 1 to |L|
      do Y ← Y ∪ (l_j ← C_j : (x, l_1, ··· , l_{j-1}))
return (x, Y) ▷ the classified example
```

Fig.3 The classification process

3.3 RF-MLC4.5 and RF-PCT

ML-C4.5 is an improved algorithm of C4.5 decision tree [25]. Clare et al. Modify the definition formula of entropy in the original algorithm to adapt to multi label classification in order to allow multiple labels in the leaf nodes of the tree. The modified entropy definition formula is as follows.

$$Entropy(D) = -\sum_{j=1}^{q}\left(p(\lambda_j)\log p(\lambda_j) + q(\lambda_j)\log q(\lambda_j)\right) \tag{1}$$

Among them, D represents the training data set and q represents the total number of tags. $p(\lambda_j)$ indicates the frequency associated with the label.

$$q(\lambda_j) = 1 - p(\lambda_j) \tag{2}$$

Predictive clustering trees (PCTs)[26] (PCTs) is the decision tree as a hierarchical clustering. The top node corresponds to the cluster containing all the data, and the lower tree node is to divide the data into smaller clusters. The PCTs algorithm uses the standard top-down decision tree algorithm. PCTs are instantiated by task variance and prototype functions. Therefore, PCTs can deal with a variety of structured output results: continuous meta output or discrete variables, and time series constitute clustering. For the prediction of discrete tags, the variance function is obtained by calculating the Gini index of the target variable. For a given target variable, its prototype function returns a possibility vector. For multi label learning, it returns a possibility variable for a given label

RF-PCT and RF-MLC4.5 algorithm are the integration methods of the above two decision tree algorithms (PCTs and ML-C4.5) as the base classifiers, using bagging integration method to obtain the diversity of the base classifiers, and changing the feature set in the learning process. Then the base classifier obtains the final result by voting.

3.4 IMLCA

In order to improve the prediction accuracy of multi-label classification algorithm and enhance the generalization ability of the algorithm, we propose an integrated IMLCA. In the process of label subset selection, this algorithm pays attention to finding the label combination with few times in training set, which makes the constructed sub model more representative.

First of all, based on the balanced K-means clustering method, *K* label are randomly selected from the label set *L* as the label clustering center, and other label closest to the Euclidean distance of each label center are added to the corresponding label set. After each clustering, the label clustering center is recalculated. The similar label is clustered into *k* label clusters, and the size of each clustering label cluster is balanced by controlling the upper limit of the size of each tag cluster.

Then, in the process of model training, one tag was randomly taken out from different tag clusters to form *k*-labelsets. Based on the data set iteration of training set, m LP classifier models are constructed. In the training set, there are fewer samples corresponding to the subset of *k*-labelsets tags composed in this way, which makes the trained subclassifier more likely to predict the output tag combination as a negative example (the possibility of such tag combination is small), thus obtaining the subclassifier with higher classification accuracy. Finally, when predicting the classification, each classifier will get the prediction result of the tag of the unknown instance, and predict the tag of the unknown instance by comprehensively calculating the average rating of each tag. The IMLCA algorithms are shown in Figures 4, 5 and 6.

Input: Number of clusters *k*, all label collection *L*, number of loops *p*

Output: *k* balance label cluster

```
for i ← 1 to k do
    C_i ← ∅;
    c_i ← random member of L;
while p > 0 do
    for each λ ∈ L do
        for i ← 1 to k do
            d_λi ← distance(λ, c_i)
        finished ← false;
        v ← λ;
        while not finished do
            j ← argmin_i d_vi;
            insert sort(υ, d_v) to sorted list C_j;
            if |C_j| > ⌈|L| / k⌉ then
                v ← remove last element of C_j;
                d_vj ← ∞;
            else
                finished ← true;
  recalculate centers;
p ← p − 1
return C_1, C_2, ..., C_k;
```

Fig.4 Label Clustering Process

Input: number of models *m*, size *k* of labelsets(tag subset), set *L* of all tags, training sample set *D*
Output: LP classifier combination and corresponding *k*-labelsets Y_i

```
R ← L^k;
for i = 1 to min(m, |L^k|) do
    Y_i ← ∅;
  for j ← 1 to k do
    Y_i ← Y_i + randomly member select from C_j
  end
  train an LP classifier h_i : X → P(Y_i) on D;
    R ← R \ {Y_i};
  end
```

Fig.5 Model training process

Input: unknown instance *x*, LP classifier h_i combination, corresponding k-labelsets Y_i, all tag sets *L*, threshold *t*
Output: multi-label classification result vector *T*

```
for j ← 1 to |L| do
    S_j ← 0;
    V_j ← 0;
for i ← 1 to m do
    for all labels λ_j ∈ Y_i do
        S_j ← S_j + h_i(x, λ_j);
        V_j ← V_j + 1;
    for j ← 1 to |L| do
        A_j ← S_j / V_j;
        if A_j > t then
            T_j ← 1;
        else T_j ← 0;
```

Fig.6 prediction classification process

4. Existing Problems and Challenges

Although multi-label data mining has made great progress in the past years, multi-label classification algorithm is faced with the following problems.

(1) Algorithm complexity and prediction accuracy. As the number of models or tags will increase after the problem transformation of the multi-label classification algorithm, the algorithm complexity will increase significantly when processing the data volume and multi-label data with large tag set size, and meanwhile, the prediction accuracy will decrease. Therefore, it is still necessary to find more universally applicable algorithms or methods to reduce the computational complexity and improve the prediction accuracy.

(2) Multi-label data set labeling bottleneck and algorithm generalization ability. A large number of annotated samples are needed to construct the model, but the information provided by annotated samples is limited. On the other hand, compared with the labeled samples, the unlabeled samples are more easily obtained and closer to the data distribution in the whole sample space. Providing as many labeled samples as possible requires a lot of time-consuming manual labeling labor, which leads to the bottleneck problem of labeling. At the same time, the learning system trained with only a small number of labeled samples often has the phenomenon of overfitting, which makes it difficult to make it have a strong generalization ability. Therefore, the research on how to train a more generalized model with only a small number of samples still needs to be further deepened.

(3) Data set skew. Through many studies in the field of machine learning, it is found that the distribution of data sets about categories is often biased or unbalanced, that is, there may be an order of magnitude difference in the number of samples between categories, which is an important factor leading to the unsatisfactory classification effect.

IMLCA uses label clustering to divide the initial label set into label clusters, and then constructs a label set by selecting a label from each label cluster to identify the important and infrequent projections in the label space. Then we use these new infrequent label sets to form new data to train the corresponding classifier. At the same time, in the IMLCA algorithm, in the process of selecting the subset of tags, the mutual exclusion between tags is introduced to reconstruct the training data set, so as to improve the classification accuracy of the training sub model, thus greatly improving the prediction accuracy and generalization ability of the algorithm.

References

[1] 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.

[2] LV Xiaoyong. Research on multi label text classification algorithm [D]. Shanxi University of Finance and economics, 2010.

[3] Schapire R E, Singer Y. BoosTexter: A boosting-based system for text categorization[J]. Machine learning, 2000, 39(2): 135-168.

[4] McDonald R, Crammer K, Pereira F. Flexible text segmentation with structured multilabel classification[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 987-994.

[5] Liu Peng, ye Zhipeng, Zhao Wei, et al. An image classification method of multi-level abstract semantic decision [J]. Acta automatica Sinica, 2015, 41 (5): 960-969.

[6] Tengzhou, Guo Yuefei. EM based multi label region calibration algorithm for unsupervised image [J].

Computer application and software, 2012, 29 (2): 5-8.

[7] Vezhnevets V, Konouchine V. GrowCut: Interactive multi-label ND image segmentation by cellular automata[C]//proc. of Graphicon. 2005: 150-156.

[8] Grady L, Funka-Lea G. Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials[M].Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis. Springer Berlin Heidelberg, 2004: 230-245.

[9] Cerri R, de Carvalho A C. Hierarchical multilabel protein function prediction using local neural networks[M].Advances in Bioinformatics and Computational Biology. Springer Berlin Heidelberg, 2011: 10-17.

[10] Berry A F H, Heal W P, Tarafder A K, et al. Rapid multilabel detection of geranylated proteins by using bioorthogonal ligation chemistry[J]. ChemBioChem, 2010, 11(6): 771-773.

[11] Otero F E B, Freitas A A, Johnson C G. A hierarchical multi-label classification ant colony algorithm for protein function prediction[J]. Memetic Computing, 2010, 2(3): 165-181.

[12] Li Sinan, Li Ning, Li zhanhuai. Multi label data mining technology: research review [J]. Computer science, 2013, 40(4): 14-21.

[13] Spyromitros E, Tsoumakas G, Vlahavas I. An empirical study of lazy multilabel classification algorithms[M].Artificial Intelligence: Theories, Models and Applications. Springer Berlin Heidelberg, 2008: 401-406.

[14] Madjarov G., Kocev D, Gjorgjevkjj D, Dzerosji S. An extensive experimental comparison of methods for multi-label learning [J]. Pattern Recognition, 2012, 45(9):3084-3104.

[15] Trohidis K, Tsoumakas G, Kalliris G, et al. Multi-Label Classification of Music into Emotions[C]//ISMIR. 2008, 8: 325-330.

[16] Zhen Chao, Zheng Tao, Xu Jieping. Research on music genre classification based on music semantic information [C] / / Proceedings of the fifth national information retrieval academic conference, 2009.

[17] Zhang Danpu, Wang Lili, Fu Zhongliang, et al. Integrated learning algorithm of label matching based on double label set [J]. Computer application, 2014, 34(9): 2577-2580.

[18] Ness S R, Theocharis A, Tzanetakis G, et al. Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs[C]//Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009: 705-708.

[19] Madjarov G., Kocev D, Gjorgjevkjj D, Dzerosji S. An extensive experimental comparison of methods for multi-label learning [J]. Pattern Recognition, 2012, 45(9):3084-3104.

[20] Read J, Pfahringer B, Holmes G. Multi-label classification using ensembles of pruned sets[C]// Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008: 995-1000.

[21] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification[J]. Pattern Recognition,2004,37(9): 1757–1771.

[22] 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.

[23] Huang Jun, Li Guorong, Wang Shuhui, et al. Group sensitive classifier chains for multi-label classification[C]// IEEE International Conference on Multimedia and Expo. Italy: IEEE Press, 2015: 1 -6.

[24] Read J, Pfahringer B, Holmes G, et al. "Classifier chains for multi-label classification[J].Machine Learning, 2011, 85(3):333–359.

[25] Clare A, King R D. Knowledge discovery in multi-label phenotype data[M]//Principles of data mining and knowledge discovery. Springer Berlin Heidelberg, 2001: 42-53.

[26] Schapire R E, Singer Y. BoosTexter: A boosting-based system for text categorization[J]. Machine learning, 2000, 39(2): 135-168.