Multi-Instance Multi-Label Learning for Automatic Classification of Chinese Patents: A Postprint

Authors: Bao Xiang, Liu Guifeng, Cui Jinghua

Date: 2023-04-01T16:02:50+00:00

Abstract

[Purpose/Significance] This study aims to achieve rapid classification of large volumes of Chinese patents to meet the requirements of patent examination and intelligence analysis. [Method/Process] Considering the inherent format of patent texts and the practical scenario where multiple IPC classification numbers exist, multi-instance multi-label learning is applied to automatic patent classification. After introducing the fundamental principles of several classical multi-instance multi-label models, these models are employed for determining IPC classification numbers of Chinese patents. [Results/Conclusion] Experiments demonstrate that multi-instance multi-label models are suitable for automatic patent classification. Furthermore, analysis based on metrics including Average precision, Hamming Loss, Ranking Loss, One Error, Coverage, and Training time reveals that the MIMLRBF model can be rapidly and accurately applied to determine IPC classification numbers for Chinese patents, providing a reference for large-scale automatic patent classification.

Full Text

Research on the Application of Multi-Instance Multi-Label Learning in Chinese Patent Automatic Classification

Bao Xiang[1], Liu Guifeng[1], Cui Jinghua[2]
[1]Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013
[2]School of Information Management, Nanjing University, Nanjing 210093

Abstract: [Purpose/Significance] This study aims to achieve rapid classification of large volumes of Chinese patents to meet the requirements of patent examination and intelligence analysis. [Method/Process] Considering the inherent format of patent texts and the practical situation where multiple IPC

classification numbers exist, multi-instance multi-label learning was applied to patent automatic classification. After introducing the basic principles of several classical multi-instance multi-label models, these models were employed to determine IPC classification numbers for Chinese patents. [Result/Conclusion] Experiments demonstrate that multi-instance multi-label models are suitable for patent automatic classification. Analysis based on metrics including Average Precision, Hamming Loss, Ranking Loss, One Error, Coverage, and Training Time reveals that the MIMLRBF model can be applied quickly and accurately to determine IPC classification numbers for Chinese patents, providing a reference for large-scale automatic patent classification.

Keywords: patent; classification; IPC classification number; multi-instance multi-label

Classification Number: G251
DOI: 10.13266/j.issn.0252-3116.2021.08.011

Patents serve as crucial property rights incentive tools that safeguard the marketization of technological innovation. In recent years, China's invention patent applications have consistently ranked first globally. While this vast number of patent applications reflects the strengthening of China's scientific and technological capabilities, it also imposes higher demands on patent management, analysis, and examination.

Patent classification represents an effective means for organizing, retrieving, analyzing, and managing massive patent literature collections. Currently, widely adopted international patent classification systems include the International Patent Classification (IPC), United States Patent Classification (USPC), European Classification (ECLA), Japanese Patent Classification (FI/F-term), and Cooperative Patent Classification (CPC). Classification based on these systems enables rapid patent retrieval and localization. However, existing patent classification number determination primarily relies on manual judgment, which suffers from drawbacks such as being influenced by the annotator's knowledge structure. Therefore, introducing intelligent technologies to solve patent classification problems holds significant importance for improving classification efficiency and accuracy.

Patent text automatic classification systems primarily involve two research aspects: patent text feature extraction algorithms and patent text classification algorithms. The most commonly used feature extraction algorithms for patents are the Bag-of-Words (BOW) method and Term Frequency-Inverse Document Frequency (TF-IDF), though both models discard substantial information from the text. Consequently, word embeddings have gained attention, with classical models including Continuous Bag-of-Words (CBOW) and Skip-Gram. Wen Chaodong et al. enhanced word vector representation capabilities by using ALBERT pre-trained dynamic word vectors to replace static word vectors trained through traditional Word2vec methods. Yu Bengong et al. mapped patent

texts into both Word2vec word vector sequences and POS part-of-speech sequences, employing two feature channels for model training. In the field of patent text classification methods, traditional machine learning approaches are frequently applied, including Naive Bayesian (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and collaborative filtering techniques. In recent years, deep learning technologies have also been widely used in patent classification, such as Convolutional Neural Networks (CNN), Bidirectional Gated Recurrent Unit (BiGRU), and Gated Recurrent Unit (GRU). Some scholars have combined these with other methods; for example, Zhou Cheng et al. used a Self-Organizing Feature Map (SOM) and SVM-based patent classification model employing self-organizing mapping for patent category determination and Random Forest (RF) for importance ranking and feature selection. Y. Lu et al. utilized Long Short-Term Memory (LSTM) networks combined with attention-based bidirectional LSTM for model training on patent corpora, classifying through a Softmax model. Lü Lucheng et al. designed seven deep learning models based on Word2Vec, CNN, Recurrent Neural Network (RNN), and Attention mechanisms, considering patent text's sequential features, contextual features, and key classification characteristics.

While these methods have studied patent classification from various perspectives and achieved satisfactory results, they fail to consider the unique structural characteristics of patent texts, including their hierarchical structure, standardized topic description, and inclusion of multiple classification numbers. In recent years, Multi-Instance Multi-Label (MIML) learning has emerged as a rapidly developing machine learning model that has demonstrated excellent performance in text and image classification. Therefore, this study investigates MIML-based patent automatic classification methods and evaluates them by combining the inherent format of patent texts with the practical situation of multiple classification numbers.

2 MIML Introduction

2.1 MIML Concept and Patent Text Structure

MIML represents a novel machine learning model that differs from other methods in that its training set consists not of individual instances but of labeled bags. Multiple instances constitute a bag, and one bag can correspond to one or multiple labels. If a bag contains at least one positive example for a particular label, the bag possesses that corresponding label; if no positive example exists for a label, the bag lacks that label.

This description naturally suggests applying MIML machine learning models to determine IPC classification numbers for patent texts. To classify patents using MIML models, we first train on existing patent bags with IPC classification numbers, then apply the learned model to predict IPC classification numbers for unlabeled patent data, thereby classifying patents with unknown IPC numbers.

Patent texts typically have fixed formats, including titles, abstracts, claims,

specifications, and specification drawings. These sections can serve as multiple instances, while the entire patent document constitutes a bag. This bag corresponds to one or multiple labels, giving patent texts their multi-instance multi-label characteristics. As shown in Table 1 , a patent can be viewed as a bag, with the patent title, abstract, and other contents serving as multiple instances. Through IPC number retrieval on the SooPAT website, this patent's IPC classification numbers include D06 (treatment of textiles; laundering; flexible materials not otherwise specified), B32 (layered products), and C08 (organic macromolecular compounds; their preparation or chemical processing; compositions based thereon). Expert semantic analysis of each instance in the patent reveals that the bag contains at least one positive example for each of the D06, B32, and C08 IPC classification numbers.

2.2 Mathematical Description of MIML Learning Models

Multi-instance multi-label learning models are based on multi-label learning and multi-instance learning, representing a more general formulation that encompasses single-instance single-label learning, multi-instance single-label learning, and single-instance multi-label learning. These three learning models can be derived as special cases of multi-instance multi-label learning, making MIML universal and complete.

The mathematical formulation of MIML can be expressed as: let X denote the instance space and Y denote the label space. The function f: 2^X → 2^Y can be trained through the dataset $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X, Y)\}$. Here, X is a bag describing a real object, consisting of a set of instances $\{x_1, x_2, \ldots, x\}$ where $x \in X$ $(j = 1, 2, \ldots, n)$, while Y represents a set of labels $\{y_1, y_2, \ldots, y\}$ corresponding to the instances, where $y \in Y$ $(k = 1, 2, \ldots, l)$. The variable n represents the number of instances describing the i-th real object, and l represents the number of labels for the i-th real object.

2.3 Introduction to MIML Learning Models

MIML builds learning models from already categorized bags and then uses these models to predict labels for unknown bags. Many real-world problems suit this learning framework, such as image classification and text classification. In image classification, an image can be viewed as a bag, divided into multiple blocks that serve as instances, with the image corresponding to multiple semantic labels like beach, cloud, or ocean. In text classification, each document can be a bag, with paragraphs as instances and the article assigned multiple topics.

MIML learning models have achieved significant theoretical extensions over decades of development, with innovations primarily concentrated on classifier learning methods. MIML models mainly fall into three categories: first, regularization-based approaches that require establishing optimization models and constraint conditions for solution; second, degeneration strategy-based approaches that use multi-instance or multi-label learning as bridges to

transform MIML problems into traditional supervised learning problems; and third, approaches that employ other methods like neural networks or gradient descent algorithms to solve classification and optimization problems.

This paper introduces and experiments with several classical MIML learning models, as summarized in Table 2 :

Table 2. Brief Introduction to Various MIML Learning Models

Model	Strategy	Description	Advantages	Disadvantages
M3MIML	Regularization and maximum margin	Assumes a linear model for each category, formulating the learning task as a Quadratic Programming (QP) problem solved in dual form	Directly utilizes connections between instances and labels without information loss	Optimization process is too complex, resulting in low algorithm efficiency, especially with large training sets
MIML-BOOST	Degeneration strategy	Transforms multi-instance multi-label into multi-instance single-label, uses Boosting methods to solve transformed multi-instance samples	Simple algorithm	Information loss during transformation
MIMLSVM	Degeneration strategy	Transforms multi-instance multi-label into single-instance multi-label, uses SVM to analyze transformed multi-label problem	Simple algorithm, high time efficiency	Information loss during transformation

Model	Strategy	Description	Advantages	Disadvantages
MIMLRBF	Radial Basis Function (RBF) neural network	Input layer contains instance bags, hidden layer consists of cluster centers after bag clustering, optimizes hidden-output layer weight matrix via minimized error sum of squares	Directly establishes connections between instances and labels	Overall network error increases when training data contains noise or is unrecognizable
MIMLNN	Backpropagation (BP) neural network	Two-stage Multilayer Perceptron (MLP) trained via backpropagation algorithm	Directly connects instances and labels, considers label correlations	Requires predetermined multiple parameters

MIML machine learning models have been widely used in text classification. The most commonly used dataset in MIML research is the Reuters-21578 text collection, which serves as a standard benchmark for MIML model evaluation. This text classification dataset contains 2,000 documents associated with 7 labels, where each bag corresponds to a document. Using sliding window technology, documents are segmented into multiple instances totaling 7,119, with 243-dimensional feature vectors extracted via bag-of-words representation. Y. Yang et al. established WKG Game-Hub, applying MIML to character classification in online games, with text corpora collected from the "Honor of Kings" game center consisting of 13,750 articles and 1,744 concept labels. Overall, no existing literature has applied MIML learning models to patent text classification.

2.4 Model Effectiveness Evaluation Metrics

MIML model learning effectiveness is typically evaluated through six metrics for determining unknown parameters and subsequent performance evaluation: Average Precision, Hamming Loss, Ranking Loss, One Error, Coverage, and Training Time. Average Precision measures classification accuracy, while Training Time measures the time consumed in training MIML models.

The Hamming Loss metric reflects misclassification degree of samples on a particular label, including two scenarios: relevant labels not appearing in the predicted label set, and irrelevant labels appearing in the predicted label set. Therefore, smaller values indicate better model performance. The formula is:

$$\text{Hamming Loss} = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{M}$$

where m represents the number of samples, M represents the total number of labels, Y represents the actual label set for sample i, Z represents the predicted label set for sample i, and Δ denotes the XOR operation between two sets.

The Ranking Loss metric evaluates the number of ranking errors in a sample's label ranking sequence, where irrelevant labels are ranked before relevant labels. Similarly, smaller values indicate better performance:

$$\text{Ranking Loss} = \frac{1}{m}\sum_{i=1}^{m}\frac{|\{(y_1, y_2) \mid f(x_i, y_1) \leq f(x_i, y_2), (y_1, y_2) \in Y_i \times \bar{Y}_i\}|}{|Y_i| \cdot |\bar{Y}_i|}$$

where $\bar{Y}_i$ is the complement of Y_i relative to the complete label set, and $f(x, y)$ is a real-valued function returning the confidence of label y for x.

The One Error metric evaluates cases where the top-ranked label in a sample's label ranking set does not belong to the relevant label set. Smaller values indicate better performance:

$$\text{One Error} = \frac{1}{m}\sum_{i=1}^{m}\left[\arg\max_{y \in Y} f(x_i, y) \notin Y_i\right]$$

The Coverage metric evaluates the search depth required to traverse all relevant labels in a sample's label ranking set. Smaller values indicate better performance:

$$\text{Coverage} = \frac{1}{m}\sum_{i=1}^{m}\left(\max_{y \in Y_i} \text{rank}_f(x_i, y) - 1\right)$$

where $\text{rank}_f(x_i, y)$ returns the rank of label y sorted in descending order by $f(x, \cdot)$.

3 Experimental Results and Analysis

3.1 Experimental Data and Implementation Process

The experimental method flow is shown in Figure 2 [Figure 2: see original paper], comprising database construction, text preprocessing and vectorization, model

training and parameter tuning, and model classification effectiveness evaluation. Database construction involves selecting training and test sets from labeled patent data according to a certain ratio. Text preprocessing and vectorization includes word segmentation, stop word removal, part-of-speech tagging with deletion of certain part-of-speech words, and establishing a vector space model based on TF-IDF. Model training and parameter tuning involves selecting the classical MIML models introduced in Section 2 and adjusting parameters to achieve optimal performance. Model evaluation assesses performance using the metrics described above, including Average Precision, Ranking Loss, Hamming Loss, One Error, and Coverage.

Each patent is treated as a bag, with the patent title and abstract serving as two instances within the bag. Specifically, 200 patent texts were selected as experimental data, with approximately 30% being multi-label bags and an average of 1.29 labels per bag. For word segmentation, this experiment employed the .NET version of jieba Chinese word segmentation in precise mode. During feature selection, the top 1,000 features corresponding to TF-IDF values were selected as data index words. The experimental processor parameters were: Intel(R) Core(TM) i5-7500 CPU @ 3.40GHz, 4GB RAM, 64-bit operating system, x64-based processor. The experimental software was MATLAB R2018a.

The patent data were selected from the water treatment technology domain in the Shanghai Intellectual Property Public Service Platform's Chinese patent database, primarily containing classification numbers, titles, abstracts, and main claims. According to SooPAT website's SooPAT IPC retrieval results, the main classification numbers and corresponding themes were: B01D (separation), C02F (treatment of water, wastewater, sewage, or sludge), and D06 (treatment of textiles; laundering; flexible materials not otherwise specified). Each category contained 250 patent documents, with approximately 60 patents containing two or more classification numbers among B01D, C02F, and D06.

3.2 Classification Results and Analysis

To validate each model's classification effectiveness and applicability, this study employed 10-times N-fold cross-validation with N ranging from 2 to 10, calculating the average of 10 test runs as the metric parameters. To achieve better experimental results, based on comprehensive literature and experimental data, the experimental parameters for various MIML models were adjusted as follows: M3MIML used SVM linear judgment model with a loss function threshold of 0.01; MIMLBOOST also selected SVM linear judgment model with rounds set to 100; MIMLSVM selected RBF judgment model with ratio parameter set to 0.2; MIMLRBF set ratio parameter to 0.1; MIMLNN set network judgment threshold to 0.5.

Considering metric variations under different N values, evaluation results for N = 3, 4, 5 were selected for discussion. The classification results of various MIML models on Chinese water treatment patents are analyzed in Tables 3-5,

with evaluation metrics including Average Precision, Ranking Loss, Hamming Loss, One Error, Coverage, and Training Time.

Table 3 . Results of Various MIML Models via Three-Fold Cross-Validation

Model	Average Precision	Ranking Loss	Hamming Loss	One Error	Coverage	Training Time (s)
M3MIML	0.8175$\pm$0.004	0.2328$\pm$0.009	0.2644$\pm$0	0.3276$\pm$0.017	0.7241$\pm$0	10.21$\pm$1.110
$MIMLBOOST$	0.8006$\pm$0					

Table 4 . Results of Various MIML Models via Four-Fold Cross-Validation

Model	Average Precision	Ranking Loss	Hamming Loss	One Error	Coverage	Training Time (s)
M3MIML	0.8428$\pm$0.010	0.2637$\pm$0.006	0.2322$\pm$0.018	0.2565$\pm$0.029	0.8523$\pm$0.005	19.25$\pm$0.781
$MIMLBOOST$	0					

Table 5 . Results of Various MIML Models via Five-Fold Cross-Validation

Model	Average Precision	Ranking Loss	Hamming Loss	One Error	Coverage	Training Time (s)
M3MIML	0.8088$\pm$0.020	0.2647$\pm$0	0.2843$\pm$0.010	0.3530$\pm$0.059	0.8535$\pm$0	16.39$\pm$3.984
$MIMLBOOST$	0.7696$\pm$0					

The bolded data in Tables 3-5 represent optimal metrics under N-fold cross-validation. Analysis yields the following conclusions:

1. MIML models can mostly classify patents accurately, with all models achieving approximately 80% classification precision. This demonstrates that MIML learning models possess high accuracy and are therefore suitable for determining Chinese patent IPC numbers.
2. As N increases, training time for all models generally increases due to larger training samples. However, other performance metrics do not consistently improve with increasing N. Overall, models achieve better metrics at four-fold cross-validation, likely because when $N < 4$, insufficient training samples prevent adequate model training, while $N > 4$ may cause overfitting, reducing generalization capability.
3. Regarding model selection, MIMLRBF consistently achieves the highest Average Precision across different N-fold cross-validation methods. Although other metrics are not always optimal, they differ minimally from

the best values, and training efficiency is second only to MIMLNN, significantly outperforming M3MIML and MIMLBOOST models. The MIMLRBF model's advantage lies in its neural network structure, where connections between instances and labels are explicit in both the clustering process between input and hidden layers and the optimization process between hidden and output layers. Practice demonstrates that MIMLRBF can solve patent classification problems accurately and efficiently.

In summary, MIML models offer significant advantages over traditional supervised learning models. Unlike conventional classification models that only consider single-instance single-label scenarios, MIML models account for both the multi-instance structural attributes and multi-label characteristics of patent texts, enabling multi-angle data selection for patent classification and consequently more precise results. Similarly, claims and specifications can be used as instances for training, and this study also considers the impact of different N-fold cross-validation methods to obtain more scientific training-test data ratios. Experiments prove that four-fold cross-validation yields optimal classification performance, setting the training set ratio at 75% and test set ratio at 25%. The MIMLRBF model is recommended for Chinese patent classification, suggesting that models with explicit connections between instances and labels should be selected for patent classification tasks.

This study fully considers the structural characteristics and inherent multi-label attributes of patent texts, applying MIML machine learning models to Chinese patent classification. Experimental metrics indicate that MIML models can achieve relatively accurate and rapid Chinese patent classification, making large-scale automatic IPC number determination feasible and significantly reducing the inefficiency and knowledge-structure dependency of manual annotation. Using only a small sample of labeled data enables large-scale patent classification, representing a positive attempt at applying artificial intelligence technology in the library and information science field.

The advantages and limitations of MIML model application are summarized as follows: First, MIML models are suitable for actual patent classification scenarios, particularly with limited labeled data, as the proposed approach can expand the application scope of patent classification and assist in determining multiple labels for large numbers of unannotated patents. Second, experiments demonstrate that many MIML models have high training efficiency, providing insights for efficient and accurate patent classification. Third, MIML model parameters in the experiments were predetermined through experimental methods, yielding relatively accurate classification results. However, in practical applications where training and test set conditions are unknown, MIML model parameters must be determined in advance, potentially affecting classification accuracy and efficiency.

Future research should address: (1) This experiment used limited patent test samples with few corresponding labels, whereas real-world scenarios involve vast numbers of patents with more labels. How to select patent text features, word

segmentation methods, and MIML models remains a key research direction. (2) This experiment only used titles and abstracts as instances for training, though specifications and claims also contain substantial technical information. How to incorporate these contents into MIML models and identify which instance combinations achieve higher classification accuracy requires investigation. (3) Traditional MIML models have numerous parameters that are difficult to determine, and parameters may significantly impact classification accuracy. How to develop a fast and accurate parameter estimation method for widespread application in patent classification research is another problem requiring solution.

References

[1] Gao Li. Patent system response to the marketization of technological innovation[J]. Journal of Jiangsu University (Social Science Edition), 2017, 19(1): 63-69.

[2] Lü Lucheng, Han Tao, Zhou Jian, et al. Research on Chinese patent automatic classification method based on deep learning[J]. Library and Information Service, 2020, 64(10): 75-85.

[3] Hu Jie, Li Shaobo, Yu Liya, et al. Patent text classification model based on convolutional neural network and random forest algorithm[J]. Science Technology and Engineering, 2018, 18(6): 268-272.

[4] Zhang Qun, Wang Hongjun, Wang Lunwen. Short text classification method fusing word vector and LDA[J]. New Technology of Library and Information Service, 2016, 32(12): 27-35.

[5] Wen Chaodong, Zeng Cheng, Ren Junwei, et al. Patent text classification combining ALBERT and bidirectional gated recurrent unit[J]. Journal of Computer Applications, 2021, 41(2): 407-412.

[6] Yu Bengong, Zhang Peixing. WPOS-GRU patent classification method based on dual-channel feature fusion[J]. Application Research of Computers, 2020, 37(3): 655-658.

[7] GOMEZ J. Analysis of the effect of data properties in automated patent classification[J]. Scientometrics, 2019, 121(3): 1239-1268.

[8] Hu Xuegang, Yang Hengyu, Lin Yaojin, et al. Patent TRIZ classification method based on collaborative filtering[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(5): 512-518.

[9] LI S, HU J, CUI Y, et al. DeepPatent: patent classification with convolutional neural networks and word embedding[J]. Scientometrics, 2018, 117(2): 721-744.

[10] Zhou Cheng, Wei Hongqin. Research on patent value evaluation and classification: based on self-organizing mapping support vector machine[J]. Data Analysis and Knowledge Discovery, 2019, 3(5): 117-124.

[11] LU Y, XIONG X, ZHANG W, et al. Research on classification and similarity of patent citation based on deep learning[J]. Scientometrics, 2020, 123(2): 813-839.

[12] ZHANG M L, ZHOU Z H. M3MIML: A maximum margin method for multi-instance multi-label learning[C]//Eighth IEEE international conference on data mining. Los Alamitos: IEEE Computer Society, 2008: 688-697.

[13] ZHOU Z H. A framework for machine learning with ambiguous objects[C]//5th international conference on active media technology. Berlin: Springer-Verlag, 2009: 6.

[14] ZHOU Z H, ZHANG M L. Multi-instance multi-label learning with application to scene classification[C]//Advances in neural information processing systems. Cambridge: Neural information processing systems foundation, 2006: 1609-1616.

[15] ZHANG M L, WANG Z J. MIMLRBF: RBF neural networks for multi-instance multi-label learning[J]. Neurocomputing, 2009, 72(16-18): 3951-3956.

[16] CHEN Z, CHI Z, FU H, et al. Multi-instance multi-label image classification: a neural approach[J]. Neurocomputing, 2013, 99: 298-306.

[17] HUANG S J, GAO W, ZHOU Z H. Fast multi-instance multi-label learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 41(11): 1868-1874.

[18] Yan Kaobi, Li Zhixin, Zhang Canlong. Multi-instance multi-label learning method based on topic model[J]. Journal of Computer Applications, 2015, 35(8): 2233-2237.

[19] SEBASTIANI F. Machine learning in automated text categorization[M]. New York: ACM, 2002.

[20] YANG Y, WU Y F, ZHAN D C, et al. Complex object classification: a multi-instance multi-label deep network with optimal transport[C]//The 24th ACM SIGKDD international conference. New York: Assoc Computing Machinery, 2018: 2594-2603.

[21] Bao Xiang, Liu Guifeng, Yang Guoli. Research on patent text classification method based on multi-instance learning framework[J]. Information Studies: Theory & Application, 2018, 41(11): 144-148.

Author Contributions: - Bao Xiang: Conceptualization, writing, experimentation - Liu Guifeng: Conceptualization, paper revision - Cui Jinghua: Data analysis

Note: Figure translations are in progress. See original paper for figures.