Computer Science and Application 计算机科学与应用, 2022, 12(2), 465-472
Published Online February 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.122047

Hans汉斯

基于双重注意力机制的多标签司法文本分类

郭绮雯，王　勇，王　瑛

广东工业大学计算机学院，广东　广州

收稿日期：2022年1月21日；录用日期：2022年2月17日；发布日期：2022年2月24日

摘　要

多标签文本分类问题是自然语言处理领域中的一项重要子任务。考虑到传统的多标签文本分类问题往往没有对标签的信息进行充分利用，本文针对司法领域文本处理过程中遇到的多标签分类问题，提出了一种基于双重注意力机制的网络模型，对标签的固有信息进行充分挖掘，并从标签语义注意力机制以及标签结构注意力机制这两个角度为文本的特征向量进行权重的分配，捕获标签与文本之间的潜在关系。为验证模型的有效性，本文设计了对比实验，结果表明，本模型在宏平均F1值、微平均F1值、综合F1值上均有明显的性能提升。

关键词

多标签文本分类，注意力机制，标签相关性

Multi-Label Classification Based on Dual Attention Mechanism for Judicial Documents

Qiwen Guo, Yong Wang, Ying Wang

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Jan. 21st, 2022; accepted: Feb. 17th, 2022; published: Feb. 24th, 2022

Abstract

The multi-label text classification problem is an important subtask of natural language processing. Considering that the traditional multi-label text classification problems often do not make full use of the information of the labels, this paper proposes a model based on dual attention mechanism for the multi-label text classification problem in the judicial field. The inherent information of the text is fully mined, and the weights are assigned to the feature vectors of the text from the two aspects of the label semantic attention layer and the label structure attention layer to capture the potential relationship between the label and the text. In order to verify the validity of the model, a

文章引用: 郭绮雯, 王勇, 王瑛. 基于双重注意力机制的多标签司法文本分类[J]. 计算机科学与应用, 2022, 12(2): 465-472. DOI: 10.12677/csa.2022.122047

comparative experiment is designed in this paper. The results show that the model has obvious performance improvement in macro-F1, micro-F1, and union-F1.

Keywords

Multi-Label Text Classification, Attention Mechanism, Label Correlation

1. 引言

随着互联网的蓬勃发展，网络为司法公开拓宽了新的思路，利用信息化技术的优势，不断拉近法院与群众之间的距离，让群众近距离感受司法，并提高群众的法律意识。截至 2022 年 1 月，在中国裁判文书网上公开的裁判文书总量已有 1 亿 2 千多万篇，单日新增文书高达 7 万多篇。裁判文书是案件判决结果的一个文字描述，其中包括了对案情的描述、被告人违反的法律条文、罪名、判处的刑期等信息。这些海量的裁判文书对于从事司法行业的人来说，是宝贵的参考资料，倘若能对它们进行有效归类，有助于为相似的案情提供指导意见，对被告人所犯的罪名进行辅助预测，降低失误风险，缓解“案多人少”的现状。

然而，在现实场景中，案情往往是错综复杂的，案件和罪名并不是一对一的关系，而是一对多的关系，涉及到数罪并罚的情况，即犯罪主体同时触犯了两个或以上的罪行。这属于多标签分类问题，要求模型需要具备预测多个标签的能力，并对标签之间存在的关联性进行充分利用。

2. 相关工作

现阶段，已经出现了多种方法解决多标签分类问题。前期解决问题的思路主要有两种，一种是基于问题转换的方法，一种是基于算法扩展的方法。

基于问题转换的方法是将多标签分类问题拆解成多个单标签分类问题，思想简单直接。Boutell 等[1]人提出了 BR (Binary Relevance)算法，为每一个标签构造一个二分类器，但 BR 算法基于标签独立性假设，忽略了标签之间的相互关系，导致标签之间的关联信息丢失。Read 等[2]针对上述 BR 算法的不足，提出了 CC (Classifier Chains)算法，采用了链式结构，分类器链上的后一个分类器都建立在前一个二分类器的预测结果上，对标签之间的关联性进行了利用，但可能会导致错误的累积和传递。Tsoumakas 等[3]将标签的子集视为一个原子标签，并通过学习一个单标签分类器来进行预测。

基于算法扩展的方法则是修改机器学习算法，使之适用于多标签分类任务，典型的算法扩展方法包括：基于支持向量机的算法 Rank-SVM [4]、基于决策树的算法 ML-DT [5]、基于 k 最近邻的算法 ML-kNN [6]等。

近年来，随着神经网络的发展，研究者们纷纷提出了各种基于神经网络模型运用在多标签文本分类问题上。Kurata 等[7]利用 CNN 来进行多标签分类，并捕捉标签之间的共现关系，来对输出层参数进行初始化。Chen 等[8]将 CNN 和 RNN 融合，将 CNN 的输出作为 RNN 的输入，同时对文本的局部语义信息和全局语义信息进行捕获。Yang 等[9]提出了 SGM (Sequence Generation Model)，首次将多标签分类任务视为序列生成任务，利用 Seq2Seq 架构，能在一定程度上对标签之间的关联性进行建模。You 等[10]

使用了 BiLSTM 去捕获单词之间的长距离依赖关系，并使用注意力机制得到与每个标签最相关的文档内容。刘等[11]提出了联合模型，使用多头注意力机制去学习单词的权重分布，并利用胶囊网络和 BiLSTM 分别提取文本的局部特征和全局特征。

考虑到在多标签分类任务中，标签也是文本，具有特殊的语义信息，同时，标签之间存在复杂的依赖性和相关性。针对上述问题，本文同时引入了标签的语义信息、结构信息，充分利用标签的固有信息，从而提高多标签文本分类性能。

3. 基于双重注意力机制的多标签文本分类模型

本文提出了一种基于双重注意力机制的多标签文本分类模型，该模型主要由词嵌入层、特征提取层层、图嵌入层、注意力层、特征融合层和输出层组成，模型的整体结构如图 1 所示。

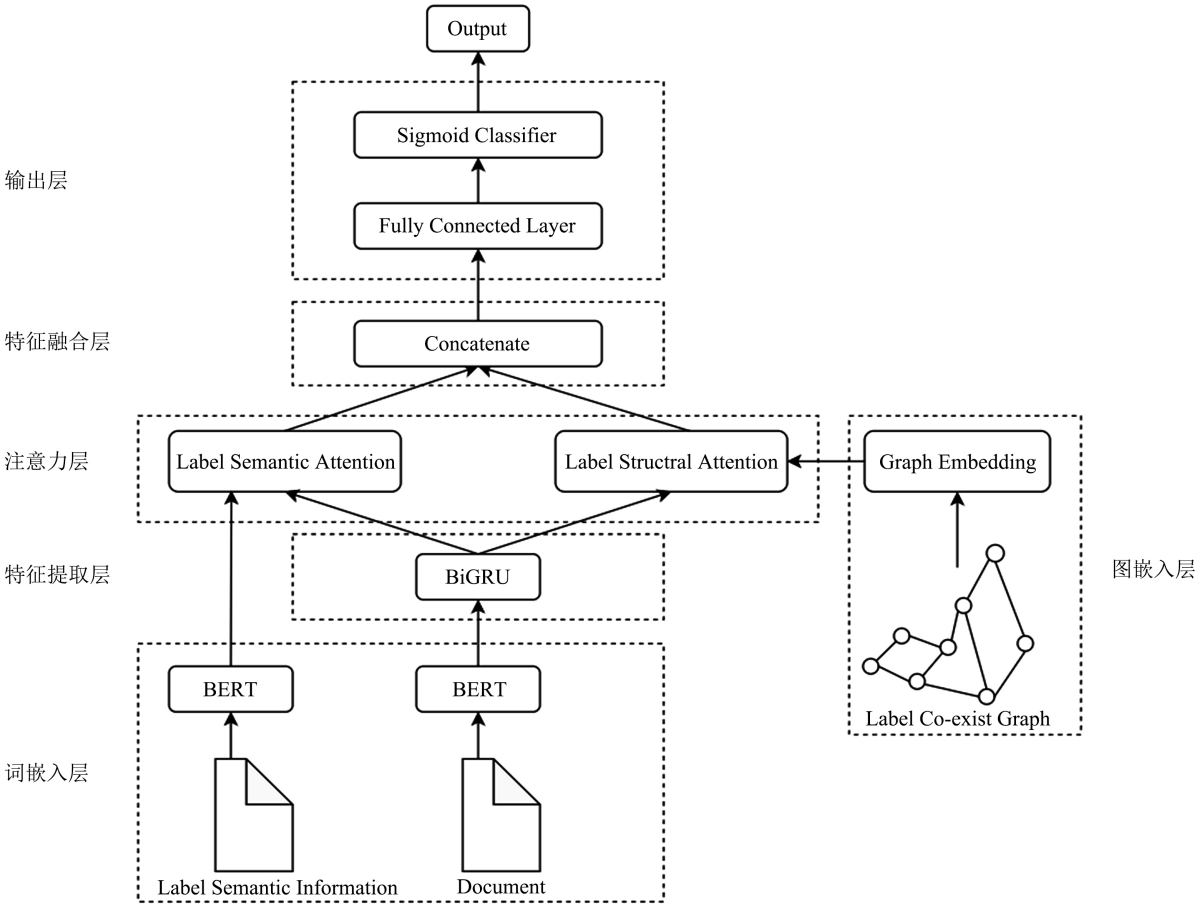

Figure 1. Overall structure of our model
图 1. 模型框架图

3.1. 任务定义

给定一个数据集共包括 N 个文档，标签集合为 $L=\{l_1,l_2,\cdots,l_k\}$，k 为标签的总数。第 i 个文档 (x^i,y^i) 由一段文本 x^i 及对应的标签子集 y^i 组成，其中 $x^i=\{w_1^i,w_2^i,\cdots,w_n^i\}$，$y^i=\{y_1^i,y_2^i,\cdots,y_m^i\}(y_j^i\in L,1\le j\le m)$，$n$ 为文档 x^i 的长度，m 为文档所属的标签子集大小。多标签文本分类的任务就是训练一个预测模型，从而将最相关的一些标签分配给一个新的未标记的样本。

3.2. 词嵌入层

BERT 预训练语言模型是由 Google 研究人员于 2018 年发布[12]，全称为 Bidirectional Encoder Representation from Transformers，即基于 Transformers 的双向编码表示。BERT 模型在发布之时就在多个 NLP 任务取得了卓越的效果。BERT 模型在大规模纯文本的语料库上通过无监督学习进行预训练，得到通用的语言表示，在下游 NLP 任务中再通过微调从而完成相应任务。BERT 模型中包含多层的 Encoder 结构，模型结构如图 2 所示。

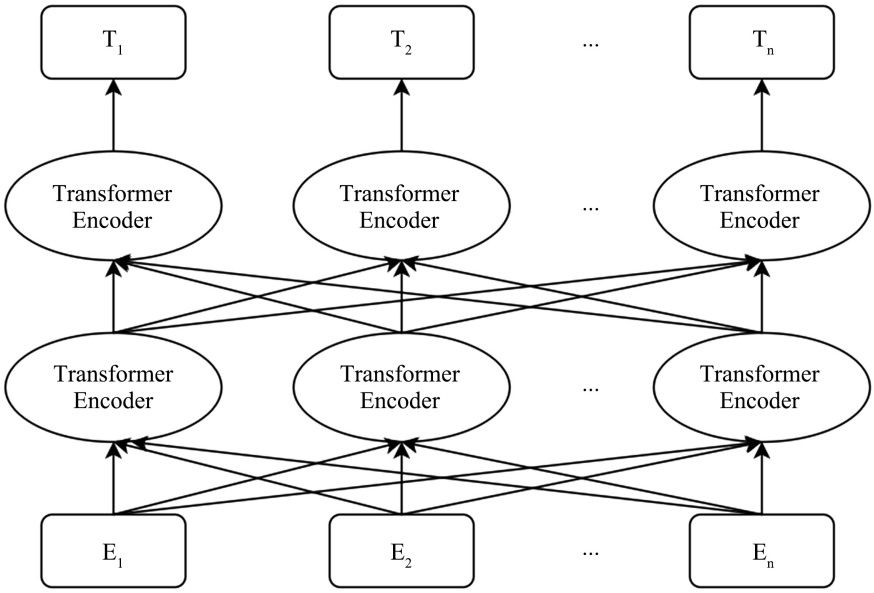

Figure 2. BERT model structure
图 2. BERT 模型结构

本模型采用 BERT 作为编码器。对输入的司法文本进行编码，将输入文本转化为具有上下文信息的字符级别特征向量表示。另外，考虑到标签也是含有语义信息的文本，本模型将标签的语义信息也输入到 BERT 模型中，取特殊标记[CLS]对应的隐层状态作为标签语义信息的句子级别特征向量表示。

3.3. 特征提取层

GRU 是长短时记忆网络的一种变体，和 LSTM 一样同是 RNN 的变体，能有效克服 RNN 面临的梯度消失问题。GRU 将遗忘门和输入门结合为更新门，输出门更名为重置门，和 LSTM 相比较少了一个门，因此参数量更少，训练速度更快。由于 GRU 网络的单向性，只能从前向捕获上文的语义信息，具有一定的局限性，可能导致重要信息的丢失。因此，本模型增添反向 GRU 对文本的逆序语义信息也进行学习，即利用 BiGRU 从前、后两个方向同时对文本序列的全局上下文特征进行提取，BiGRU 的输入为 BERT 预训练语言模型得到的司法文本特征向量。

假设 x_t 为 t 时刻的输入向量，$\vec{h}_t$ 和 $\overleftarrow{h}_t$ 分别代表正向和反向 GRU 在 t 时刻的输出向量，并将全局正向隐向量 $\vec{h}_t$ 和全局反向隐向量 $\overleftarrow{h}_t$ 组合成全局隐向量 h_t，数学表达式如下：

$$\vec{h}_t = \mathrm{GRU}\left(x_t, \vec{h}_{t-1}\right) \tag{1}$$

$$\overleftarrow{h}_t = \mathrm{GRU}\left(x_t, \overleftarrow{h}_{t-1}\right) \tag{2}$$

$$h_t = \left[\vec{h}_t; \overleftarrow{h}_t\right] \tag{3}$$

3.4. 图嵌入层

在多标签文本分类任务中，一个样本可能会对应多个标签，标签之间相互关联、相互影响，因此，要对标签之间的相关性进行充分利用。本模型首先会利用训练集中的标签共现关系构建出标签共存图 $G=(V,E)$，其中 V 表示节点的集合，E 表示边的集合。如果两个标签同时作为一个文档的标签出现，则它们之间连有一条边。对标签共存图，使用 SDNE (Structural Deep Network Embedding) [13]来进行图嵌入，对图中的结构信息进行保存，将标签共存图中的节点映射成向量表示。SDNE 是一个半监督的模型，利用深度自编码器来对高度非线性的网络结构进行捕获。SDNE 包括了两个组件，分别是无监督组件和有监督组件。无监督组件利用二阶相似性来捕获网络的全局结构，有监督组件使用一阶相似性来捕获网络的局部结构。通过在半监督模型中对一阶、二阶相似度进行联合优化，能对网络结构信息进行有效保留，优化函数如下：

$$L_{\text{mix}} = vL_{\text{reg}} + \alpha L_{\text{1st}} + L_{\text{2nd}} \tag{4}$$

$$L_{\text{1st}} = 2tr\left(Y^T LY\right) \tag{5}$$

$$L_{\text{2nd}} = \left(\hat{S} - S\right) \odot B_F^2 \tag{6}$$

其中，Y 是顶点的嵌入向量，L 是拉普拉斯矩阵。S 为邻接矩阵，$\hat{S}$ 为自编码器重构后的邻接矩阵，$\odot$ 是哈达玛积，B 是一个形状与邻接矩阵 S 相同的矩阵，当 $s_{i,j}=0$ 时，$b_{i,j}=1$；当 $s_{i,j}=1$ 时，$b_{i,j}=\beta>1$。其中，α、β 和 v 都是参数，α 控制一阶相似性，β 控制邻接矩阵中非零元素的重构程度，v 控制 L2 正则化项。

3.5. 注意力层

近几年来，注意力机制(Attention Mechanism)在文本分类任务中被广泛应用，如图 3 所示。注意力机制模仿了人类大脑在对大量信息处理的过程中，聚焦于重要信息，忽略不必要信息的思维方式。在网络中添加注意力机制，可以让网络对输入序列中的某些特殊信息进行重点关注，从而提高网络对特征的捕捉能力。

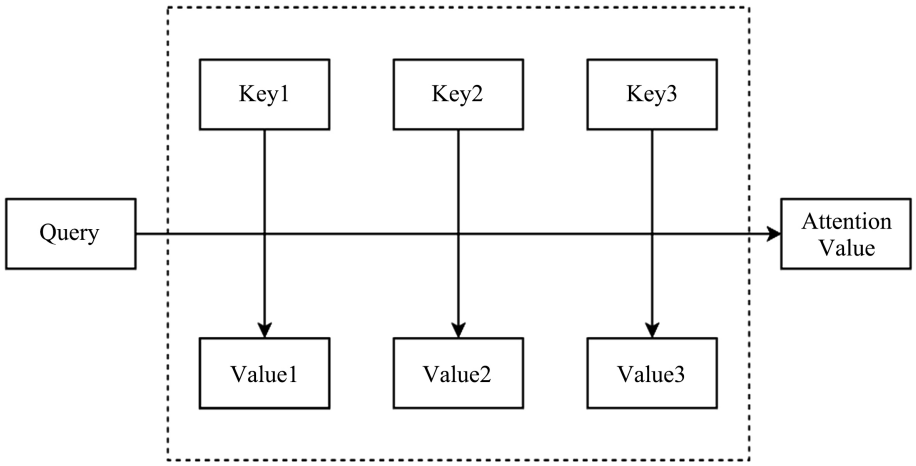

Figure 3. Attention mechanism
图 3. 注意力机制

注意力机制的具体的计算过程可分为三步：

1) 根据 Query 和对应的 Key 值进行相似度计算得到二者的注意力得分，如下所示：

$$f\left(Q,K\right)=QK^{\mathrm{T}} \tag{7}$$

2) 使用 Softmax 函数对上述得到的注意力得分进行归一化处理得到权重系数，如下所示：

$$a_i=\mathrm{Softmax}\left(f\left(Q,K\right)\right) \tag{8}$$

3) 将权重系数与对应的 Value 值进行加权求和得到注意力输出，如下所示：

$$\mathrm{Attention}\left(Q,K,V\right)=\sum a_iV \tag{9}$$

将通过特征提取层获得的输入文本序列上下文信息分别传入语义注意力层、结构注意力层，不同的查询会给源文本中的内容赋予不同的权重，从而对潜在信息进行捕获。本模型引入注意力机制对标签信息进行充分的利用，突出文本对每个标签分类的贡献，语义注意力层关注标签固有的语义信息，结构注意力层倾向于关注标签之间的关联性，通过这两个注意力层来对文本特征进行进一步的提取。

3.6. 特征融合层和输出层

将语义注意力层和结构注意力层的输出进行拼接，融合后的特征作为全连接层的输入 z。全连接层以 Sigmoid 为激活函数，将各个标签所对应的输出值压缩至[0, 1]区间，计算公式如下所示，并将输出值大于等于设定阈值的标签作为预测结果。

$$\mathrm{Output}=\mathrm{Sigmoid}\left(Wz+b\right) \tag{10}$$

其中，W 是参数矩阵，b 是偏置项。

4. 实验与分析

4.1. 数据集及预处理

数据来源于 CAIL2018 罪名预测任务[14]，原始数据集包括单标签数据和多标签数据，提取出数据集中的多标签数据作为本实验数据集，并将出现次数少于 100 的罪名删除，因为其样本过少，难以支撑模型的训练。本实验的标签为罪名，因此从东方法律网上搜集罪名的法律释义作为标签的语义信息。在数据集预处理的过程中，去除文档中的特殊符号。由于裁判文书存在特定的写作规范，因此对无用的子句，比如“经审理查明”、“公诉机关指控”等进行去除。

4.2. 评价指标

本实验的评价指标使用宏平均 F1 值 $\mathrm{F1}_{\mathrm{macro}}$、微平均 F1 值 $\mathrm{F1}_{\mathrm{micro}}$ 及综合 F1 值 $\mathrm{F1}_{\mathrm{union}}$，其计算公式如下所示：

$$\mathrm{F1}_{\mathrm{macro}}=\frac{\sum_{i=1}^{n}\mathrm{F1}_i}{n} \tag{11}$$

$$\mathrm{F1}_{\mathrm{micro}}=\frac{2\times\mathrm{Precision}_{\mathrm{micro}}\times\mathrm{Recall}_{\mathrm{micro}}}{\mathrm{Precision}_{\mathrm{micro}}+\mathrm{Recall}_{\mathrm{micro}}} \tag{12}$$

$$\mathrm{F1}_{\mathrm{union}}=\frac{\mathrm{F1}_{\mathrm{macro}}+\mathrm{F1}_{\mathrm{micro}}}{2} \tag{13}$$

4.3. 参数设置

本文实验参数如下：BERT 模型的隐藏单元数为 1024，输入文本句子长度为 500。LSTM 隐藏层维度为 1024，图嵌入维度为 1024，学习率为 1e−4。为了降低模型过拟合的风险，设置 Dropout 为 0.2，并使用早停(Earlystopping)策略，若模型的损失在验证集上的效果没有明显的提升，则提前结束训练。

4.4. 结果与分析

为了验证本文提出的模型有效性，本实验与以下一些常用的多标签模型进行对比实验，其中比较的模型有：

TextCNN：利用多种不同尺寸的卷积核对文本中不同大小的信息量提取，能更好地捕获局部相关性。

TextRCNN：通过双向 RNN 和一层最大池化层进行特征提取。

BiGRU-Attention：通过双向 GRU 进行特征提取，GRU 是 LSTM 的变体，结构较 LSTM 简单，同时利用 Attention 机制对部分文本加强特征提取。

Bert：利用预训练模型，直接获取文本的句子级向量，再输入到分类器进行分类。

Table 1. Experiment Results
表 1. 实验结果

模型	$F1_{macro}$ (%)	$F1_{micro}$ (%)	$F1_{union}$ (%)
TextCNN	24.96	88.04	56.50
TextRCNN	24.38	91.12	57.75
BiGRU-Attention	26.37	89.41	57.89
Bert	25.10	88.51	56.81
Our model	31.13	94.32	62.73

表 1 显示了 TextCNN、TextRCNN、BiGRU-Attention、Bert 和本文提出的模型在数据集上的实验结果。从表中的数据可以看出，宏平均 F1 值、微平均 F1 值以及综合 F1 值，都是本文提出的模型最高，证明本模型有明显的性能提升。另外，由实验结果可以观察到宏平均 F1 值均小于微平均 F1 值，这是标记样本量少以及类别不平衡所导致的。宏平均 F1 值的提升能反映出引入标签语义信息和标签的结构信息能建立标签固有信息于文档内容之间的潜在关联，有效改善模型在“小类”上的预测性能。

5. 总结

本文针对多标签分类问题提出了基于双重注意力机制的网络模型。模型首先使用了 BERT 作为词嵌入层，得到源本文和标签语义信息的向量表示；使用 BiGRU 对文本向量进行双向特征提取，获取上下文信息；并利用图嵌入得到标签之间的依赖关系；然后利用标签语义注意力层、标签结构注意力层来对文档中的关键信息进行提取；并对特征进行融合，最后输入到全连接层获取分类结果。经过与多个模型进行比对，本文提出的模型能有效提高多标签文本分类的性能。

参考文献

[1] Boutell, M.R., Luo, J., Shen, X., *et al*. (2004) Learning Multi-Label Scene Classification. *Pattern Recognition*, **37**, 1757-1771. https://doi.org/10.1016/j.patcog.2004.03.009

[2] Read, J., Pfahringer, B., Holmes, G., *et al*. (2009) Classifier Chains for Multi-label Classification. *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*: *Part II*. Berlin, 2009, 254-269. https://doi.org/10.1007/978-3-642-04174-7_17

[3] Tsoumakas, G. and Vlahavas, I. (2007) Random *k*-Labelsets: An Ensemble Method for Multilabel Classification. In: Kok, J.N., Koronacki, J., Mantaras, R.L., Matwin, S., Mladenič, D. and Skowron, A., Eds., *European Conference on Machine Learning*, Springer, Berlin, Heidelberg, 406-417.

[4] Elisseeff, A. and Weston, J. (2001) A Kernel Method for Multi-Labelled Classification. Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS, Vancouver, 3-8 December 2001, 681-687.

[5] Clare, A. and King, R.D. (2001) Knowledge Discovery in Multi-Label Phenotype Data. In: De Raedt, L. and Siebes, A. Eds., *Principles of Data Mining and Knowledge Discovery*, Springer, Berlin, Heidelberg, 42-53. https://doi.org/10.1007/3-540-44794-6_4

[6] Zhang, M.L. and Zhou, Z.H. (2007) ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recognition*, **40**, 2038-2048. https://doi.org/10.1016/j.patcog.2006.12.019

[7] Kurata, G., Bing, X. and Zhou, B. (2016) Improved Neural Network-based Multi-label Classification with Better Initialization Leveraging Label Co-occurrence. *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*. San Diego, June 2016, 521-526. https://doi.org/10.18653/v1/N16-1063

[8] Chen, G., Ye, D., Xing, Z., *et al*. (2017) Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-Label Text Categorization. 2017 *International Joint Conference on Neural Networks* (*IJCNN*). Anchorage, 14-19 May 2017, 2377-2383. https://doi.org/10.1109/IJCNN.2017.7966144

[9] Yang, P., Sun, X., Li, W., *et al*. (2018) SGM: Sequence Generation Model for Multi-Label Classification. *Proceedings of the* 27*th International Conference on Computational Linguistics*, Santa Fe, 20-26 August 2018, 3915-3926.

[10] You, R., Dai, S., Zhang, Z., *et al*. (2018) Attention XML: Extreme Multi-Label Text Classification with Multi-Label Attention Based Recurrent Neural Networks. *Computing Research Repository*, 18, 17-27.

[11] 刘心惠, 陈文实, 周爱, 等. 基于联合模型的多标签文本分类研究[J]. 计算机工程与应用, 2020, 56(14): 111-117.

[12] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, 2-7.

[13] Wang, D., Cui, P. and Zhu, W. (2016) Structural Deep Network Embedding. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, August 2016, 1225-1234. https://doi.org/10.1145/2939672.2939753

[14] Xiao, C., Zhong, H., Guo, Z., *et al*. (2018) CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction.