南京大学

研 究 生 毕 业 论 文

（申请硕士学位）

论　文　题　目　基于层次多标签学习的法律适用自动识别

作　者　姓　名　王景峰

学科、专业方向　计算机技术

指　导　教　师　柏文阳 副教授

研　究　方　向　数据挖掘

2016 年 5 月

学　　号 ： MF1333044

论文答辩日期 ： 2016 年 6 月 1 日

指 导 教 师 ：　　（签字）

Automatic Legal Application Recognition Based on Hierarchical Multi-label Learning

by

Jingfeng Wang

Directed by

Associate Professor Wenyang Bai

Department of Computer Science and Technology
Nanjing University

May 2016

Submitted in partial fulfilment of the requirements
for the degree of Master in Computer Technology

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于层次多标签学习的法律适用自动识别

计算机技术 专业 2013 级硕士生姓名：王景峰

指导教师（姓名、职称）：柏文阳 副教授

摘 要

本文是南京大学学位论文的 LaTeX 模板。目前不支持本科生学位论文格式。

除了介绍 LaTeX 文档类 NJUthesis 的用法外，本文还是一个简要的学位论文写作指南。

关键词： 南京大学; 学位论文; LaTeX 模板

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Automatic Legal Application Recognition Based on Hierarchical Multi-label Learning

SPECIALIZATION: Computer Technology

POSTGRADUATE: Jingfeng Wang

MENTOR: Associate Professor Wenyang Bai

Abstract

This paper is a thesis template of Nanjing University. Besides that the usage of the LaTeX document class `NJUthesis`, a brief guideline for writing the thesis is also included.

Keywords: Nanjing University (NJU), Thesis, LaTeX Template

目　录

目录 ····· iii

第一章 绪论 ····· 1
1.1 研究背景及意义 ····· 1
1.2 研究内容及目标 ····· 2
1.3 论文组织 ····· 3

第二章 文本挖掘技术 ····· 5
2.1 文本表示 ····· 6
2.1.1 文本表示模型 ····· 6
2.1.2 特征权重计算 ····· 6
2.2 文本特征选择 ····· 7
2.2.1 文档频率 ····· 8
2.2.2 信息增益 ····· 8
2.2.3 互信息 ····· 9
2.2.4 卡方统计 ····· 10
2.3 本章小结 ····· 10

第三章 层次多标签学习 ····· 11
3.1 问题定义 ····· 11
3.1.1 多标签学习 ····· 11
3.1.2 层次分类 ····· 13
3.1.3 层次多标签分类 ····· 14
3.2 评价指标 ····· 14
3.3 学习算法 ····· 16
3.3.1 多标签学习算法 ····· 16
3.3.2 局部学习方法 ····· 22
3.3.3 全局学习方法 ····· 22
3.4 本章小结 ····· 22

第四章 LocalBalance算法 ····· 23
4.1 算法描述 ····· 23
4.2 本章小结 ····· 23

第五章 法律适用自动识别 ····· 24
5.1 实验数据 ····· 24
5.1.1 数据获取 ····· 24
5.1.2 数据预处理 ····· 24
5.1.3 标签处理 ····· 26
5.1.4 特征处理 ····· 28
5.2 实验平台 ····· 29
5.3 评价指标 ····· 29
5.4 实验结果 ····· 29

第六章 总结与展望 ····· 30
6.1 工作总结 ····· 30
6.2 改进方向 ····· 30

参考文献 ····· 31

致谢 ····· 33

表　　格

5.1　特征词中动词名词比例 ····· 29

插　　图

1.1　法律条文树形结构示例 …… 3

2.1　文本挖掘流程 …… 5

3.1　类别层次结构类型 …… 14

5.1　裁判文书样例 …… 25

5.2　标签空间树形结构 …… 27

第一章　绪论

1.1　研究背景及意义

随着我国法治建设的逐步推进，人民的法律意识日渐提高，人们在遇到争议事件时会更多地选择诉诸法律，以公平公正地解决争端。根据最高人民法院的数据，2015年全国各级法院审结一审民事案件达622.8万件。然而，由于法律的专业性和复杂性，普通民众自身在借助法律维护自身权益的时候往往无所适从，只能求助律师等专业人士；另一方面，法律条文浩如烟海，即便是专业律师也只能专注于某一领域，在面对不熟悉的法律条文或者案例时，也需要一些决策辅助。

信息技术，尤其是信息检索和数据挖掘技术的发展，为法律辅助系统的实现提供了可能。“北大法宝”、“找法网”等一批在线法律信息平台，提供了法规案例检索、律师推荐等功能，在一定程度上为人们诉诸法律解决争端提供了便利。然而，上述平台提供的服务并未直接解决人们的问题：法规案例的检索往往需要用户有明确的搜索目标，甚至需要一定的法律领域知识，而且即便搜索引擎能够给出相应的搜索结果，这些结果通常也无法直接解决用户的问题，需要用户自己的分析和理解；律师推荐能够方便用户找到合适的律师，实际上是连接用户和律师的桥梁，不仅无法提供问题的直接解决方案，还容易受商业化的影响，出现一些律师滥竽充数的情况。

随着我国司法公开改革的推进以及最高人民法院关于人民法院在互联网公布裁判文书的规定的实施，蕴藏了海量信息的裁判文书可以方便地被获取和分析。2014年以来，全国各级法院共在“中国裁判文书网”上传裁判文书上千万份，最高人民法院和部分省市区法院实现了能够上网的生效裁判文书全部上网的目标。裁判文书记载了人民法院审理案件的过程和结果，是诉讼活动结果的载体，也是人民法院确定和分配当事人实体权利义务的惟一凭证。一份结构完整、要素齐全、逻辑严谨的裁判文书，既是当事人享有权利和负担义务的凭证，也是上级人民法院监督下级人民法院民事审判活动的重要依据。因此，裁判文书中包含的当事人诉求、犯罪行为、行政执法、司法裁判行为和过程、法律的适用等信息，作为重要的历史数据，通过数据挖掘手段进行分析，可以为司法人员、律师和普通民众提供必要的决策支持。[1]实现了一个裁判文书推荐系统，为法官提供与当前裁判文书相似的文书，作为裁判的参考。文中基于自然语言处理技术提取文书的语义信息，在裁判文书的相似度计算上取得了不错的

效果。裁判文书推荐可以提供决策辅助，但是逐条查阅相似文书需要耗费大量精力，同时由于与当前裁判文书相似度不同，用户需要自行确定各个文书的权重进行综合评判，使得决策辅助功能弱化。

从本质上讲，裁判是法院依照法律，对案件做出决定的过程。“以事实为根据，以法律为准绳”是我国社会主义法律适用遵循的基本原则，司法机关处理一切案件，都是根据客观事实，以国家法律为标准和尺度。因此，根据案件的描述确定适用的法律，是法院判决过程的核心部分，也是律师和普通民众在法律活动中需要解决的首要问题。运用信息技术，根据案件事实描述实现适用法律的自动识别，将在很大程度上为人们的法律活动提供更加直接和明确的帮助。现已公开的裁判文书中包含的案件事实描述以及法律适用信息，为我们提供了大量带类别标签的数据集，采取合适的数据挖掘手段，可以从中学习得到有效的预测模型，实现对未判案件适用法律的自动识别。

1.2　研究内容及目标

本文希望通过运用数据挖掘方法，从海量的裁判文书中，学习出由案件事实描述到适用法律的预测模型，从而为用户提供直接的法律决策辅助。

一份结构完整的裁判文书包括首部、事实、理由、裁判结果和尾部五个部分，其中首部包括裁判文书的类型、编号、裁判法院，案件当事人、委托代理人等信息，事实部分包含了对案件事实的文字描述，理由部分阐述了法院对于案件的分析以及做出相应裁判结果的理由，裁判结果部分给出了法院对于此次诉讼的判决或裁定结果，尾部则包含了案件的裁判人员、时间等信息。

运用数据挖掘手段进行案件适用法律的自动识别，首先需要提取能够充分描述案件事实的特征项。由于裁判文书及其中的事实描述部分主要是以文本形式存在，为非结构化数据，因此需要运用到文本挖掘技术[2] 对裁判文书进行处理，对其进行结构化，结构化的过程包括中文分词，文本表示，特征权重计算和特征选择等。

本文通过监督式学习方法来构建预测模型，样本的类别标签即为案件适用的法律条文，包含在裁判文书中的裁判理由部分。由于裁判文书格式的不规范性，案件适用法律条文的书写没有统一格式，因此需要对提取的案件适用法律条文做进一步处理，形成案件样本的类别标签。与传统的分类问题不同的是，一个案件往往可以适用多项法律条文，因此法律适用的自动识别问题是一个多标签分类问题[3] [4]。在多标签学习中，每个实例可以对应多个类别标签，使得学习问题更加复杂。更进一步地，法律条文的组织呈现为树状结构，如图1.1所示。一个案件不仅可能适用多项法律条文，这些法律条文的具体程度也可能不

同，即案件适用的法律条文可能位于树结构的叶节点，也可能位于树结构的内部节点。如果忽略类别标签空间的树形结构特征，无疑会损失重要的分类信息[5] [6]，学习的模型无法达到满意的预测性能。因此，如何利用法律适用识别问题中类别标签的结构信息，是本文的重要研究内容。本质上，法律适用自动识别问题是一个层次多标签学习问题[7] [8]，其中样本的特征需要通过文本挖掘手段从文本中提取，而类别标签的结构呈树形。

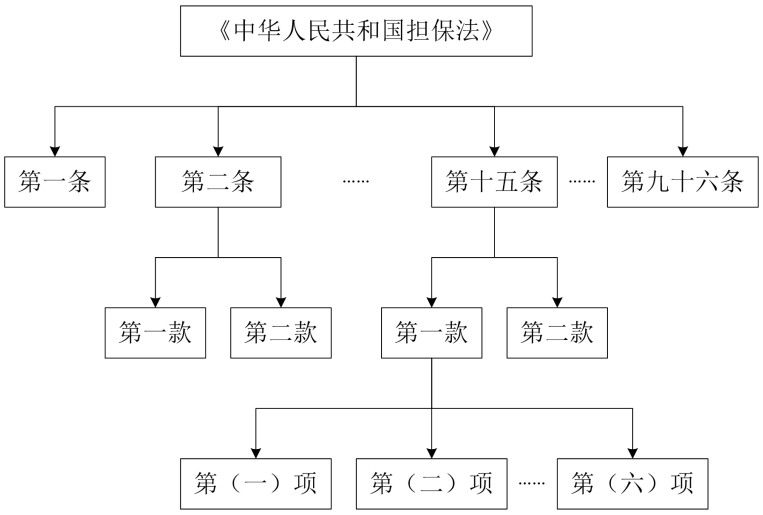

图 1.1: 法律条文树形结构示例

综上，本文研究的目标是解决案件适用法律的自动识别问题，研究的方法是首先利用文本挖掘技术对海量裁判文书进行处理分析，得到案件事实的结构化表示，即样本特征，样本集中的每个实例可以对应于标签空间中的多个类别标签，标签空间以树状结构组织。在此样本集上通过层次多标签学习算法构建预测模型，实现对未判案件适用法律的识别。

1.3　论文组织

本文组织如下：

第一章阐述本文研究的背景和意义，法律适用自动识别在当前社会法律活动中的重要辅助作用，并提出研究内容和目标，指出面临的问题及解决方向。

第二章主要对文本挖掘相关技术进行介绍，包括文本的表示模型，特征词权重计算和特征选择等。

第三章介绍层次多标签学习，从多标签学习和层次学习两方面对层次多标签学习的问题定义、评价指标、学习算法等进行介绍。

第四章详述本文提出的一种新的层次多标签学习方法$LocalBalance$，包括算法的主体思想和细节分析。

第五章为实验部分，即运用层次多标签分类方法构建案件适用法律自动识别模型，并对其预测性能进行评估的过程，包括对裁判文书文本的处理，以及模型预测效果的分析。

第六章对本文工作进行总结，并提出改进的方向。

第二章　文本挖掘技术

文本挖掘是一门交叉性学科，涉及数据挖掘、机器学习、模式识别、统计学、计算机语言学等多个领域。文本挖掘旨在从大量文本中发现隐含的知识和模式，它从数据挖掘发展而来，但又与传统的数据挖掘有许多不同。文本挖掘的对象是海量、异构、分布的文本，文本内容是人类使用的自然语言，缺乏计算机可理解的语义。传统数据挖掘所处理的数据是结构化的，而文本挖掘所处理的文本都是非结构化或半结构化的。所以，文本挖掘面临的首要问题是如何合理地表示文本，使之既能包含足够的信息以充分反映文本的特征，又不至于过于复杂使学习算法无法处理。在浩如烟海的网络信息中，80%的信息是以文本的形式存在的。

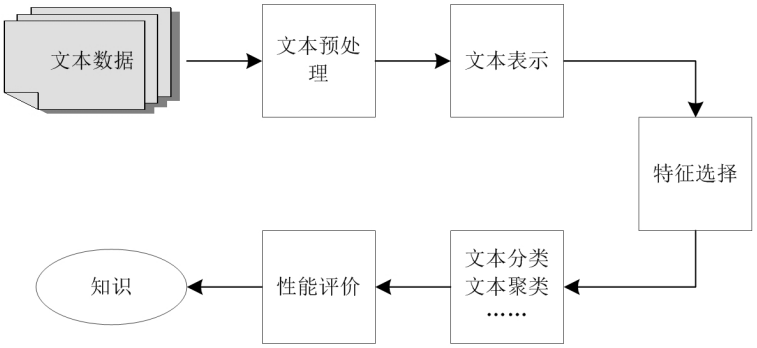

图 2.1: 文本挖掘流程

文本挖掘的一般流程如图2.1所示，其中文本的表示及其特征项的选取是文本挖掘、信息检索的核心问题，其目的是将非结构化的文本数据转化为结构化的计算机可以识别处理的数据。目前有关文本挖掘的研究主要集中于文本表示模型和特征选择算法的研究上。用于表示文本的特征项必须具备一定的特性：(1)特征项要能够确切标识文本的内容；(2) 特征项具有区分不同文本的能力；(3)特征项的个数不能过多；(4)特征项的提取要比较容易实现。在中文文本中可以采用字、词或短语作为表示文本的特征项。相比较而言，词比字具有更强的表达能力，而词和短语相比，词的切分难度比短语的切分难度小得多。因此，目前大多数中文文本分类系统都采用词作为特征项，称作特征词。这些特征词作为文档的中间表示形式，用来实现文档之间的相似度计算。如果把切分得到的所有词都作为特征词，那么特征向量的维数将过于巨大，从而导致后续计算量太大，在这样的情况下，要完成文本分类等任务几乎是不可能的。特征选择

的目的是在保留文本重要信息的情况下尽量减少特征词的数量，以此来降低特征向量空间维数，从而简化计算，提高文本挖掘的速度和效率。文本特征选择对文本内容的过滤和分类、聚类、自动摘要以及用户兴趣模式发现、知识发现等有关方面的研究都有非常重要的影响。通常根据某个特征评估函数计算各个特征词的评分值，然后按评分值对这些特征词进行排序，选取若干个评分值最高的特征词作为特征。

2.1　文本表示

2.1.1　文本表示模型

目前，最常用的文本表示模型有两种，一种是布尔模型，另一种是向量空间模型(Vector Space Model, VSM)。经典的向量空间模型由Salton等人[9]于上世纪七十年代提出，并成功地应用于著名的SMART文本检索系统。向量空间模型概念简单，将文档d表示成m维特征向量，即$d = (w_1, w_2, ..., w_m)$，特征空间$\{t_1, t_2, ..., t_m\}$ 由从文档集合中选取的m个特征词组成，$w_i(1 \leq i \leq m)$表示特征词t_i 在文档d中的权重。布尔表示模型可以看做向量空间模型的一个特例，区别在于布尔表示模型中特征词的权重由该特征词在文档中是否出现决定，若出现则权重为1，否则为0；而向量空间模型中，特征词的权重通常为该特征词在文档中出现频率的函数。

利用布尔模型和向量空间模型表示文本，假定了文本中的字或词的出现是相互独立的，即文本中字或词出现的次序对文本的分类没有影响。在实际情况中，这一假设往往是不成立的，忽略字词出现的顺序将丢失大量的文本内容信息，造成模型的表示能力不足。尽管存在以上问题，向量空间模型的简洁实用，以及在大量实际应用中取得的良好效果，使得其成为了目前应用最为广泛的文本表示模型。

通过上述的向量空间模型，文本数据就转换成了计算机可以处理的结构化数据，即文本特征向量，两个文档之间的相似性问题转变成了两个向量之间的相似性问题，可以利用欧氏距离、余弦相似度等进行衡量。

2.1.2　特征权重计算

向量空间模型中，特征词的权重计算有不同的方法，特征词的权重取值，会在较大程度上影响文本分类、聚类算法的性能。目前常用的特征词权重计算方法有布尔权重法、对数权重法、TF-IDF 权重法、平方根权重法和基于熵的权重法等，其中TF-IDF算法[10]是应用最为普遍的算法。

TF-IDF(Term Frequency-Inverse Document Frequency)是一种应用于信息检索与文本挖掘的常用加权技术。TF-IDF 是一种统计方法，用以评估字或词对于一个文档集合或文档集合中某个文档的重要程度。TF-IDF的主要思想是：一个字或词在特定文档中的重要性随着其在文档中出现的频率成正比提高，同时会随着其在文档集合中出现的频率成反比降低，即如果某个字词或短语在一份文档中出现的频率高，并且在文档集合中的其他文档中很少出现，则认为该字词或者短语具有很好的类别区分能力，适合用来分类，应该具有更高的权重。

利用TF-IDF计算特征词的权重考虑了三方面的因素：

- 词频tf(Term Frequency)：特征词在文档中出现的频率。
- 反文档频率idf(Inversed Document Frequency)：特征词在文档集合中分布情况的量化，度量特征词在文档集合中出现的频繁程度。常用的计算方法是$\log(N/n+0.01)$，其中N 是文档集合中文档的总数，n是文档集合中出现特征词t的文档数量。
- 归一化因子(Normalization Factor)：对文本向量中的各个分量进行归一化，归一化后的文本向量为单位向量，从而减轻文本长度对文本向量的影响。

综上，利用TF-IDF算法计算特征词t_j在文档d_i中权重的公式如下：

$$tf\text{-}idf_{ij} = \frac{tf_{ij} \times \log(N/n_j + 0.01)}{\sqrt{\sum_{t_j \in d_i}[tf_{ij} \times \log(N/n_j + 0.01)]^2}} \tag{2.1}$$

其中，tf_{ij}表示特征词t_j在文档d_i中出现的频率，idf_j表示特征词t_j在文档集合中的反文档频率，N表示文档集合中的文档总数，n_j表示特征词t_j在文档集合中的文档频率，即文档集合中出现特征词t_j的文档数目，分母为归一化因子。

2.2　文本特征选择

目前人们通常采用向量空间模型来表示文本，通过把从文本中抽取出的特征词进行量化来表示文本信息，但是如果直接用分词算法和词频统计方法得到的所有特征词构成文本特征集合，那么文本向量的维度将会非常大。这种未经处理的文本向量表示不仅给后续文本挖掘任务带来巨大的计算开销，使整个处理过程的效率非常低下，而且会损害分类、聚类算法的精确性，无法得到满意的结果。因此，必须对原始文本向量进行简化，在保证尽可能少地损失文本信息的基础上，选取最具代表性的文本特征来表示文本向量。为了解决这个问题，最有效的办法就是通过特征选择来降维。

特征降维方法可以分为两类：特征选择(Feature Selection)和特征提取(Feature Extraction)。特征选择是从原始的d维特征空间中，选择为我们提供信息最多的k个维，这k 个维是原始特征空间的子集；特征提取则是将原始的d维特征空间映射到k维空间中，新的k维空间不属于原始特征空间。在文本挖掘与文本分类中，通常采用特征选择方法进行维度约简，原因在于文本的特征一般都是词，具有语义信息，使用特征选择得到的k维特征，仍然是以单词作为特征，保留了语义信息，而特征提取得到的k维特征，会失去原有的语义信息，影响模型的可解释性。

常用的文本特征选择方法主要基于文档频率(DF)，信息增益(IG)，互信息(MI)，卡方统计(CHI)等衡量指标[11] [12] [13]。

2.2.1　文档频率

文档频率(Document Frequency, DF)通过统计特征词在文档集合中出现的文档数量，来衡量某个特征词的重要性。对于文档集合中的每个特征词，计算其文档频率值，文档频率低于某个设定的阈值的词将从特征空间中移除。使用文档频率进行特征选择的原理是，如果某些特征词在文档中经常出现，那么这个词就可能很重要。而对于在文档中出现很少的特征词，携带的信息量很少，甚至可能是“噪声”，这些特征词，对分类器学习影响很小。文档频率特征选择方法属于无监督的学习算法，仅考虑了频率因素而没有考虑类别因素，因此，文档频率算法将会考虑一些频率较高但没有分类意义的词，如中文里的“的”、“是”，“个”等。这类词常常具有很高的文档频率，但是对分类影响很小。

基于文档频率的特征选择是最为简单的方法，往往无法选取最具分类信息的特征词，但是由于其计算复杂度低，随文本集合中文档的数量线性增长，因此适合海量数据的处理。

2.2.2　信息增益

信息增益(Information Gain, IG)方法是通过比较当得知一个文档中某个特征词存在或缺失时，所获得的用于类别预测的信息量，来衡量某个特征词的重要性。同样地，令$\{c_i\}_{i=1}^{m}$ 表示目标空间中的标签集合，信息增益的定义为：

$$\begin{aligned} IG(t) = & -\sum_{i=1}^{m} P_r(c_i) \log P_r(c_i) \\ & + P_r(t) \sum_{i=1}^{m} P_r(c_i|t) \log P_r(c_i|t) \\ & + P_r(\bar{t}) \sum_{i=1}^{m} P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t}) \end{aligned} \tag{2.2}$$

上述定义是信息增益的一种通用形式，可以应用到多分类情形中，考虑到文本分类往往不是简单的二分类问题，通用形式的信息增益可以在衡量特征词的重要性时考虑到所有类别。对于文档集合中的每个特征词，我们计算其信息增益，信息增益值低于某个设定的阈值的词将从特征空间中移除。

2.2.3　互信息

互信息(Mutual Information, MI)用于表征特征词与文档类别之间的相关性。考虑特征词t和文档类别c，令A为包含特征词t且属于类别c的文档数量，B为包含特征词t 而不属于类别c 的文档数量，C 为不包含特征词t而属于类别c的文档数量，N为总文档数量，则特征词t和类别c的互信息定义为

$$MI(t,c) = \log \frac{P_r(t \wedge c)}{P_r(t) \times P_r(c)} \tag{2.3}$$

可以由上述定义的A, B, C, N进行估算：

$$MI(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)} \tag{2.4}$$

当特征词t和类别c独立时，其互信息为0。为了衡量特征词的重要性，需要综合各个类别的互信息，综合的方法有取均值和取最大值两种：

$$MI_{avg}(t) = \sum_{i=1}^{m} P_r(c_i) MI(t,c_i) \tag{2.5}$$

$$MI_{max}(t) = \max_{i=1}^{m} MI(t,c_i) \tag{2.6}$$

互信息的缺点在于容易受到特征词的边缘概率影响，即如果某个特征词的频率很低，那么其互信息值就会很大，因此互信息倾向于选择“低频”的特征词，而词频很高的词互信息值就会变低，如果这词携带了很高的信息量，互信息法就会变得低效。

2.2.4　卡方统计

卡方统计(χ^2 Statistic, CHI)特征选择算法利用了统计学中的“假设检验”的基本思想：首先假设特征词与类别是不相关的，如果利用CHI分布计算出的检验值偏离阈值越大，那么更有信心否定原假设，接受原假设的备则假设：即特征词与类别有着很高的关联度。考虑特征词t和文档类别c，令A为包含特征词t且属于类别c的文档数量，B为包含特征词t而不属于类别c 的文档数量，C为不包含特征词t而属于类别c的文档数量，D为不包含特征词t且不属于类别c的文档数量，N为总文档数量，则特征词t和类别c的卡方统计量定义为：

$$\chi^2(t,c) = \frac{N \times (AD - BC)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{2.7}$$

当特征词t和类别c独立时，其卡方统计量为0。与互信息类似，针对某个特征词，我们计算其关于各个类别的卡方统计量，然后分别用取均值和取最大值两种方式进行综合：

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P_r(c_i)\chi^2(t,c_i) \tag{2.8}$$

$$\chi^2_{max}(t) = \max_{i=1}^{m} \chi^2(t,c_i) \tag{2.9}$$

卡方统计和互信息是两种衡量特征词和类别关联性的方法，卡方统计相比于互信息的主要优点在于它是归一化的值，因此可以更好地衡量同一类别中的不同特征词。

2.3　本章小结

文本信息在网络以及现实世界中的广泛存在，使得文本挖掘的重要性与日俱增，其中文本表示和文本特征选择又是文本挖掘过程中的重要步骤。尽管有很多新的文本表示方法和特征选择算法提出，本章介绍的几种常用方法和算法依然由于其计算效率和性能，成为目前的主流，得到了普遍的应用。本文的案件适用法律识别问题作为一个文本分类问题，除了采用以上文本表示方法和特征选择算法对文本进行处理，还涉及到分类算法的研究。传统的文本分类算法有决策树、朴素贝叶斯、神经网络、支持向量机算法[2] 等，但是由于法律适用识别问题的层次多标签特性，这些算法无法直接用来解决本文问题，需要研究更为适合的层次多标签学习方法。

第三章　层次多标签学习

层次多标签分类问题与传统机器学习中的分类问题有两方面不同：(1)一个样本可以同时属于多个类别；(2)样本的类别标签集合以层次结构组织，属于一个类别的样本自动也属于该类别的父类别。层次多标签分类问题在很多领域都有出现，包括文本分类、功能基因组学、对象识别等[14]。传统的机器学习方法往往无法直接处理层次多标签分类问题，或者由于没有利用到类别标签集合的层次结构特征而无法取得最优的学习效果。层次多标签分类就是希望通过利用类别之间的关联和层次结构，提高学习得到的预测模型的性能。本章从多标签学习[3] [4] 和层次分类[14] 两方面对层次多标签分类问题进行定义，并介绍常用的层次多标签评价指标和分类算法。

3.1　问题定义

3.1.1　多标签学习

传统的监督学习框架下，样本由实例及其关联的类别标签组成，实例往往具有明确而单一的语义，即样本的类别标签唯一。学习系统在包含了足量训练样本的集合上利用学习算法学习得到预测模型，对未见实例的类别进行预测。现实世界中的大多数机器学习任务可以由上述学习框架实现，因此传统的监督学习被广泛研究，取得了巨大的成功。然而在某些场景中，实例对象往往不只有唯一的语义，可能具有多义性：如一篇新闻报道可以同时属于“经济”、“体育”、“NBA”等类别；一幅图片可以同时包含“大海”、“日落”等元素；一个基因可能同时具有“新陈代谢”和“蛋白质合成”等功能，等等。这种情况下，传统的只考虑明确而单一语义的监督学习框架难以取得好的效果。为了直观地反映多义性对象所具有的多种语义信息，一种很自然的方式就是为该对象显式地赋予一组合适的类别标签，即标签子集。基于上述考虑，作为一种多义性对象学习建模工具，多标签学习框架应运而生。在此框架下，每个对象由一个实例描述，实例具有多个而不再是唯一的类别标签，学习的目标是将所有合适的类别标签赋予该实例。

假设$\mathcal{X} = \mathbb{R}^d$代表$d$维的实例空间，$\mathcal{Y} = \{y_1, y_2, ..., y_q\}$代表包含$q$个类别的标签空间。给定多标签训练集$\mathcal{D} = \{(x_i, Y_i)|1 \leq i \leq m\}$，其中$x_i \in \mathcal{X}$为$d$维的特征向量$(x_{i1}, x_{i2}, ..., x_{id})^{\mathrm{T}}$，而$Y_i \subseteq \mathcal{Y}$为与$x_i$对应的一组类别标签，学习系统的任务

是从中学习得到一个多标签分类器$h : \mathcal{X} \to 2^{\mathcal{Y}}$。对于未见实例$\mathbf{x} \in \mathcal{X}$，分类器预测该实例的类别标签集合为$h(x) \subseteq \mathcal{Y}$。

在许多情况下，学习系统的输出往往对应于某个实值函数$f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$，其中$f(\mathbf{x}, y)$可以看作实例$\mathbf{x}$具有类别标签$y$的“置信度”。对于给定的实例$\mathbf{x}$及其对应的类别标签集合$Y$，一个有效的学习系统将在属于$Y$的类别标签上输出较大的值，而在不属于$Y$的类别标签上输出较小的值，即有$f(\mathbf{x}, y') > f(\mathbf{x}, y'')(y' \in Y, y'' \notin Y)$成立。此外，实值函数$f(\cdot, \cdot)$还可以转化为一个排序函数$rank_f(\cdot, \cdot)$，该排序函数将所有的实值函数输出$f(\mathbf{x}, y)(y \in \mathcal{Y})$映射到集合$\{1, 2, ..., q\}$上，使得当$f(\mathbf{x}, y') > f(\mathbf{x}, y'')$成立时$rank_f(\mathbf{x}, y') < rank_f(\mathbf{x}, y'')$也成立。

上述的多标签分类器$h(\cdot)$其实可以由实值函数$f(\cdot, \cdot)$转换而来：给定阈值函数$t : \mathcal{X} \to \mathbb{R}$，则$h(\mathbf{x}) = \{y | f(\mathbf{x}, y) > t(\mathbf{x})\}$。换句话说，给予阈值$t(\mathbf{x})$学习系统将标签空间二分为“相关”标签集合和“不相关”标签集合。阈值函数通常设为“常量函数”。

如果限定每个实例只对应一个类别标签，则传统的监督学习框架可以看作多标签学习框架的特例。多标签学习的一般性使得解决该问题的难度大大增加。总的来看，多标签学习所面临的最大挑战在于其输出空间过大，即输出空间的类别标签集合数将随着标签空间的增大而成指数规模增长。例如，当标签空间具有20个类别标签时(q=20)，则可能的类别标签集合数将超过一百万(2^20)。

为了有效应对标签空间过大所造成的学习困难，学习系统需要充分利用标签之间的相关性来辅助学习过程的进行。例如，如果已知一副图像具有类别标记“狮子”和“草原”，则该图像具有类别标签“非洲”的可能性将会增加；如果已知一篇新闻报道具有类别标签“娱乐”，则该新闻报道同时属于类别标签“政治”的可能性将会降低。因此，如何充分利用标签之间的相关性是构造具有强泛化能力多标签学习系统的关键。基于考察标签之间相关性的不同方式，已有的多标签学习问题求解策略大致可以分为以下三类：

1. “一阶（first-order）”策略：该类策略通过逐一考察单个标签而忽略标签之间的相关性，如将多标签学习问题分解为个独立的二类分类问题，从而构造多标签学习系统。该类方法效率较高且实现简单，但由于其完全忽略标签之间可能存在的相关性，其系统的泛化性能往往较低。

2. “二阶（second-order）”策略：该类策略通过考察两两标签之间的相关性，如相关标签与无关标签之间的排序关系，两两标签之间的交互关系等等，从而构造多标签学习系统。该类方法由于在一定程度上考察了标签之间的相关性，因此其系统泛化性能较优。然而，当真实世界问题中标签之间具

有超越二阶的相关性时，该类方法的性能将会受到很大影响。

3. “高阶（high-order）”策略：该类策略通过考察高阶的标签相关性，如处理任一标签对其它所有标签的影响，处理一组随机标签集合的相关性等等，从而构造多标签学习系统。该类方法虽然可以较好地反映真实世界问题的标签相关性，但其模型复杂度往往过高，难以处理大规模学习问题。

由上可见，不同的多标签学习问题求解策略具有各自的优缺点。本章第3节将针对不同的求解策略，介绍几种具有代表性的多标签学习算法。

3.1.2　层次分类

一直以来，数据挖掘、机器学习相关的研究大多集中在扁平(flat)分类问题，即标准的二分类和多分类问题。然而，现实世界中很多重要的分类问题都以层次分类问题出现，即要预测的类别以层次结构组织，这些层次结构由人为地预先指定，典型的有树结构(Tree)和有向无环图(Direct Acyclic Graph, DAG)两种。例如在文本分类中，网页、专利、维基百科等文本的类别都以树形结构组织；在生物信息学应用如蛋白质功能预测[7]中，基因功能类别以有向无环图的形式组织。层次学习问题的定义主要依赖于类别层次的定义。

一般地，类别的层次可以用二元组$(C, \prec_h)$表示，其中C是类别的集合，$\leq_h$是表示类别父子关系的偏序。偏序关系$\prec_h$可以理解为“属于”(”IS-A”)关系，具有非对称性、非自反性和传递性，即：

1. 根节点R为层次中的唯一最大元素，即任意$c_i \in C$且$c_i \neq R$，有$c_i \prec R$
2. 任意$c_i, c_j \in C$，如果有$c_i \prec c_j$，那么$c_j \nprec c_i$
3. 任意$c_i \in C$，有$c_i \nprec c_i$
4. 任意$c_i, c_j, c_k \in C$，$c_i \prec c_j$且$c_j \prec c_k$，则有$c_i \prec c_k$

类别组织结构满足上述四个特征的分类问题都可以认为是层次分类问题。上述定义不仅适用于树形结构，也适用于有向无环图结构。在有向无环图中，一个节点可能有多个父节点，相比树形结构更为复杂，算法的设计难度更大，因此目前层次分类方面的研究主要针对树形层次结构。

层次分类问题包括不同形式，[14]提出可以根据以下三个属性对层次分类问题进行分类：

- ϒ 表示类别层次结构，即用来表示类别及其相互关系的图的类型，如可能的取值有
  - 树T：表示预测的类别组织成树结构，如图3.1(a)所示；
  - 有向无环图DAG：表示预测的类别组织成有向无环图，如图3.1(b)所示。

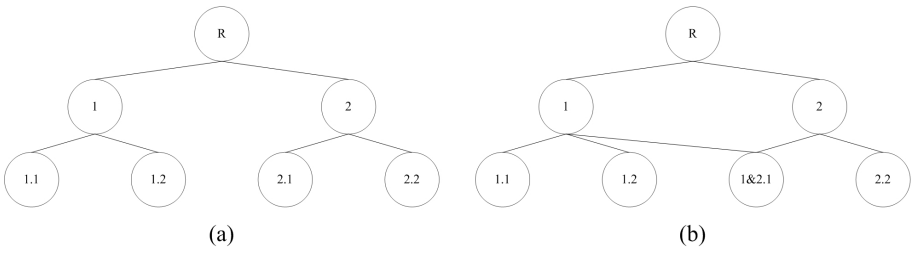

图 3.1: 类别层次结构类型

- Ψ 表示样本在类别层次中是否可以对应多条路径，即样本是否可以对应多个类别标签。如在图3.1(a)的树形结构中，如果一个样本同时对应标签1.2和2.1，那么这个样本就对应于多条标签路径。Ψ的取值有两种：
- Φ 表示样本的标签深度

3.1.3　层次多标签分类

层次多标签分类综合了多标签分类和层次分类两方面的特点。

3.2　评价指标

在多标签学习问题中，由于每个对象可能同时具有多个类别标签，因此传统监督学习中常用的单标签评价指标，如精度（accuracy）、查准率（precision）、查全率（recall）等，无法直接用于多标签学习系统的性能评价。因此，研究者们相继提出了一系列多标签评价指标，总的来看可分为两种类型，即“基于样本”的评价指标（example-based metrics）以及“基于类别”的评价指标（label-based metrics）。

基于样本的多标签评价指标首先衡量分类器在单个测试样本上的分类效果，然后返回其在整个测试集上的“均值（mean value）”作为最终的结果。基于上节的符号表示，给定多标签分类器以及多标签测试集，其中为隶属于示例的相关标签集合。常用的基于样本的多标签评价指标包括：

Subset accuracy:

其中，对于任意的谓词，当成立时取值为1，否则取值为0。该评价指标用于考察预测的标签集合与真实的标签集合完全吻合的样本占测试集的比例情况。该指标取值越大则系统性能越优，其最优值为。值得注意的是，当标签空间中包含大量类别标签（q很大）时，学习系统往往难以给出与真实的标签集合完全吻合的预测，此时该评价指标的取值将会很低。

Hamming loss:

其中，算子用于度量两个集合之间的“对称差（symmetric difference）”，算子用于返回集合的“势（cardinality）”。该评价指标用于考察样本在单个标签上的误分类情况，即相关标签未出现在预测的标签集合中或无关标签出现在预测的标签集合中。该指标取值越小则系统性能越优，其最优值为。值得注意的是，当中的每个样本仅含有一个类别标签时，hamming loss 的取值即为传统分类误差的倍。

One-error:

其中，为与多标签分类器对应的实值函数。该评价指标用于考察在样本的类别标签排序序列中，序列最前端的标签不属于相关标签集合的情况。该指标取值越小则系统性能越优，其最优值为。值得注意的是，当中的每个样本仅含有一个类别标签时，one-error即为传统的分类误差。

Coverage:

其中，为与实值函数对应的排序函数。该评价指标用于考察在样本的类别标签排序序列中，覆盖所有相关标签所需的搜索深度情况。该指标取值越小则系统性能越优，其最优值为。

Ranking loss:

其中，为集合在标签空间中的“补集（complementary set）”。该评价指标用于考察在样本的类别标签排序序列中出现排序错误的情况，即无关标签在排序序列中位于相关标签之前。该指标取值越小则系统性能越优，其最优值为。

Average precision:

该评价指标用于考察在样本的类别标签排序序列中，排在相关标签之前的标签仍为相关标签的情况。该指标取值越大则系统性能越优，其最优值为。值得注意的是，该指标最先出现于信息检索领域，用于度量给定查询下检索系统返回文档的排序性能。

与基于样本的多标签评价指标不同，基于类别的多标签评价指标首先衡量分类器在单个类别上对应的“二类分类（binary classification）”效果，然后返回

其在所有类别上的“均值（macro-/micro averaged value）”作为最终的结果。给定多标签测试集，对于第个类别而言，分类器在该类别上的二类分类性能可由如下四个统计量进行刻画：

（“真”正例的个数，# true positive instances）：

（“伪”正例的个数，# false positive instances）：

（“真”负例的个数，# true negative instances）：

（“伪”负例的个数，# false negative instances）：

据式(7)-式(10)可知，成立。值得注意的是，绝大部分二类分类性能指标均可由以上四个统计量导出，例如:

基于此，令代表由所定义的统计量求得的某种二类分类性能指标，则基于类别的多标签评价指标可采用如下两种方式获得：

Macro-averaging:

Micro-averaging:

其中，macro-averaging首先基于统计量求得在各个类上的分类性能，然后再将所有类上的均值作为最终结果，其基本思想是为各个类赋予相同的权重。相应地，micro-averaging首先将各个类上的统计量相加，然后再将求得的分类性能作为最终结果，其基本思想是为各个样本赋予相同的权重。

总的来看，已有的各种基于样本的或者基于类别的多标签评价指标是从不同的侧面来衡量学习系统的泛化性能。目前并不存在适用于所有问题的“通用的（general-purpose）”多标签评价指标，其选择依赖于具体的学习任务。例如，对于“分类（classification）”任务而言，采用基于样本的评价指标如hamming loss可能比较合适；而对于“检索（retrieval）”任务而言，采用基于类别的评价指标如micro-averaged precision可能比较合适。除此之外，各评价指标之间的关系尚不明确，如优化其中一些指标是否意味着同时优化其他一些指标等，关于这方面的研究工作目前还比较少。

3.3　学习算法

3.3.1　多标签学习算法

一般而言，算法研究是机器学习研究的核心课题，这一点对于多标记学习而言也不例外。目前已经涌现出了大量的多标记学习算法，总的来看大致可以分为两类：a) “问题转换（problem transformation）”方法：该类方法的基本思想是通过对多标记训练样本进行处理，将多标记学习问题转换为其它已知

的学习问题进行求解。代表性学习算法有一阶方法Binary Relevance[8]，该方法将多标记学习问题转化为“二类分类（binary classification）”问题求解；二阶方法Calibrated Label Ranking[26]，该方法将多标记学习问题转化为“标记排序（label ranking）”问题求解；高阶方法Random k-labelsets[23]，该方法将多标记学习问题转化为“多类分类（multi-class classification）”问题求解。b) “算法适应（algorithm adaptation）”方法：该类方法的基本思想是通过对常用监督学习算法进行改进，将其直接用于多标记数据的学习。代表性学习算法有一阶方法ML-kNN[27]，该方法将“惰性学习（lazy learning）”算法k 近邻进行改造以适应多标记数据；二阶方法Rank-SVM[11]，该方法将“核学习（kernel learning）”算法SVM 进行改造以适应多标记数据；高阶方法LEAD[22]，该方法将“贝叶斯学习（Bayes learning）算法”Bayes 网络进行改造以适应多标记数据。换句话说，问题转换方法的核心是“改造数据适应算法（fit data to algorithm）”，本章3.1 小节将介绍前述的三种基于该方法的代表性算法；算法适应方法的核心是“改造算法适应数据（fit algorithm to data）”，本章3.2 小节将介绍前述的三种基于该方法的代表性算法。

3.3.1.1　问题转换算法

3.2.1 Binary Relevance该算法的基本思想是将多标记学习问题转化为q 个独立的二类分类问题，其中每个二类分类问题对应于标记空间中的一个类别标记[8]。基于2.1 节的符号表示，给定多标记训练集，其中为隶属于示例的相关标记集合。具体来说，对于第个类别而言，Binary Relevance 算法首先构造与该类别对应的二类训练集：基于此，Binary Relevance 算法采用某种二类学习算法训练二类分类器，即。由此可见，对于任一多标记样本，示例将参与q 个二类分类器的学习。其中，对于相关标记而言，在构造二类分类器时对应于正例；对于无关标记而言，在构造二类分类器时对应于反例。该训练策略亦称为“交叉训练（cross-training）”法[8]。在测试阶段，对于未见示例，Binary Relevance 算法通常采用如下方式预测其类别标记集合：值得注意的是，当所有二类分类器的输出均为负值时，将会导致算法预测的标记集合为空。为了避免这种情况的发生，可以采用如下的T-Criterion 准则[8]来进行预测：此时，当所有二类分类器输出为负时，预测的标记集合中将含有输出值“最大（least negative）”的类别标记。除了上述的T-Criterion 准则之外，Boutell 等人[8]还给出了其它一些基于各二类分类器输出确定测试样本标记集合的准则，具体细节可参见相应文献。3.2.2 Calibrated Label Ranking该算法的基本思想是将多标记学习问题转化为标记排序问题，其中标记排序采用“成对比较（pairwise comparison）”的

方式实现[26]。对于具有q 个类别的标记空间而言，针对每一个可能的标记配对，采用成对比较的方式将产生共计个二类分类器。具体来说，对于标记配对而言，成对比较法首先构造与该配对对应的二类训练集：其中，算子的定义如式(16)所示。基于此，成对比较法采用某种二类学习算法训练二类分类器，即。由此可见，对于任一多标记样本，示例将参与个二类分类器的学习。其中，对于的情况而言，在构造二类分类器时对应于正例；对于的情况而言，在构造二类分类器时对应于反例。在测试阶段，对于未见示例，Calibrated Label Ranking 算法首先将其提交给已训练的个二类分类器，得到该示例在各个类别标记上的“投票（votes）”：据式(20)可知，成立。基于2.1 节的符号表示，令，则可根据相应的排序函数对标记空间中的所有类别标记实现排序。当时，标记与的相对排序位置随机确定。值得注意的是，利用成对比较法虽然可以得到函数，实现对所有标记的排序。但如2.1 节所示，为了得到最终的多标记分类器，仍需确定相应的阈值函数，从而将标记的排序序列“二分（bipartition）”为相关标记集合与无关标记集合。为了在成对比较的框架下实现该目标，Calibrated Label Ranking 算法为每个多标记样本加入一个“虚拟标记（virtual label）”，该虚拟标记的作用是在的相关标记集合与无关标记集合之间加入一个“人工分割点（artificial splitting point）”。换句话说，在标记排序序列中，虚拟标记应位于所有相关标记之后，并位于所有无关标记之前。此时，针对每一个新的标记配对，Calibrated Label Ranking 算法将在原有的个二类分类器基础上，额外训练个二类分类器。具体来说，对于标记配对而言，首先构造与该配对对应的二类训练集：基于此，Calibrated Label Ranking 算法采用二类学习算法训练与虚拟标记对应的二类分类器，即。基于新求得的二类分类器，可在式(20)的基础上更新未见示例在各个类别标记上的投票：此外，进一步计算未见示例在虚拟标记上的投票：基于2.1 节的符号表示，令且，可得到所需的多标记分类器：值得注意的是，对照式(21)与式(16)的定义，训练集即为Binary Relevance 算法所使用的训练集。因此，Calibrated Label Ranking 算法可以看作是在常规标记配对求得的个二类分类器基础上，进一步引入Binary Relevance 算法求得的q 个二类分类器，以辅助学习任务的完成[26]。3.2.3 Random k-Labelsets该算法的基本思想是将多标记学习问题转化为多类分类问题的“集成（ensemble）”，集成中的每一个基分类器对应于标记空间的一个随机子集，并采用“Label Powerset”的方式进行构造[23]。简单地说，Label Powerset（简记为LP）是一种直观地将多标记学习问题转化为多类分类问题的方法。对于包含q 个类别的标记空间而言，给定多标记训练集，我们可以将训练集中出现的每一种标记组合看作一个“新类（new class）”。不失一般性，设为标记空间的“幂空间”至自然数空间的“单射函数

（injective function）”，而为与对应的“逆函数（inverse function）”。首先，LP 方法将原始的多标记训练集转化为如下的多类（单标记）训练集：其中，数据集中含有新类：显然，成立。基于此，LP 方法采用某种多类学习算法训练多类分类器，即。由此可见，对于任一多标记样本，示例的标记集合首先被映射为一个新类，然后参与多类分类器的学习。在测试阶段，对于未见示例，LP 方法采用如下方式预测其类别标记集合。(27)值得注意的是，虽然上述LP 方法可以将多标记学习问题转化为多类分类问题进行求解，但是该方法存在两个主要缺陷。首先，由式(26)及式(27)可知，LP 仅能预测在训练集中出现过的类别标记集合（即），对于其真实标记集合在训练集中未出现的测试示例无法正确预测；其次，当标记空间较大时，往往会导致新类集合过大，从而导致部分新类在中的训练样本不足且多类分类器的训练复杂度过高。为了充分发挥LP 方法简单直观的优势并同时克服其存在的缺陷，Tsoumakas 与Vlahavas 提出了Random k-Labelsets 算法，结合“集成学习（ensemble learning）”技术与LP 方法求解多标记学习问题。其算法核心是每次仅针对一个随机“k-标记集（k-labelsets）”调用LP 方法，并将多次调用所得的多类分类器进行集成以得到最终的输出。其中，“k-标记集”是标记空间的一个子集，包含k 个类别标记。设为标记空间的所有“k-标记集”所构成的集合，其中。不失一般性，记中的第l 个“k-标记集”为，则，,。采用与式(25)相同的符号表示，对于而言，在调用LP 方法时首先构造与之对应的训练集：其中，数据集中含有新类：基于此，采用某种多类学习算法训练多类分类器，即。此外，假设Random k-Labelsets 算法所考察的集成大小为n，即针对n 个随机“k-标记集”分别调用LP 方法，得到相应的多类分类器。基于此，对于未见示例，针对各个类别标记计算如下统计量：其中，用于统计基于集成在类别标记上的最大投票数，而用于统计基于集成在类别标记上的实际投票数。基于2.1 节的符号表示，令且，则可得到所需的多标记分类器：换句话说，当集成在上的实际投票数超过最大投票数的半数时，该标记即被认为是未见示例的相关标记。一般而言，对于由n 个“k-标记集”生成的集成而言，每个类别标记所能得到的最大投票数的平均值为。Random k-Labelsets 算法推荐的默认设置为，此时各类别标记最大投票数的平均值为6。本小节分别介绍了三种具有代表性的“问题转换”类型的多标记学习算法，即一阶方法Binary Relevance[8]、二阶方法Calibrated Label Ranking[26]以及高阶方法Random k-Labelsets[23]。除此之外，目前还存在其他一些基于“问题转换”的一阶[4][28]、二阶[4][29]以及高阶[30][31][32][33][34]多标记学习算法。限于篇幅，这里不再做一一介绍。

3.3.1.2　算法适应算法

3.3.1 ML-kNN该算法的基本思想是采用“k 近邻（k-nearest neighbors）”分类准则，统计近邻样本的类别标记信息，通过“最大化后验概率（maximum a posteriori，简记为MAP）”的方式推理未见示例的标记集合[27]。给定多标记训练集以及未见示例，假设代表在训练集中的k个近邻样本构成的集合。对于第个类别而言，ML-kNN 算法将计算如下的统计量：由上可知，统计了中将作为其相关标记的样本个数。进一步地，设代表具有类别标记这一事件，代表当中有个样本具有类别标记时，成立的后验概率。相应的，代表当中有个样本具有类别标记时，不成立的后验概率。基于2.1 节的符号表示，令且，可得到所需的多标记分类器：换句话说，当后验概率大于后验概率时，即将标记赋予示例。基于贝叶斯定理，函数可重写为：(35)其中，与分别代表事件成立与不成立的先验概率，与分别代表事件成立与不成立时，中有个样本具有类别标记的条件概率。值得注意的是，上述先验概率以及条件概率可基于训练集通过“频率计数（frequency counting）”的方式进行估计。具体来说，先验概率可以通过如下方式估计而得：其中，“平滑（smoothing）”参数s 用以控制“均匀分布（uniform prior）”在概率估计时的权重，通常s 设置为1 对应于Laplace 平滑。与先验概率的估计不同，条件概率的估计过程要相对复杂一些。对于第个类别而言，ML-kNN 算法首先确定两个数组以及，其中每个数组各含有如下k+1 个元素：其中，与式(33)类似，式(39)中定义的统计了第i 个训练样本的k 近邻中，将作为其相关标记的近邻个数。相应地，统计了具有标记且其k 近邻中恰好有r 个近邻具有标记的训练样本个数，统计了不具有标记且其k 近邻中恰好有r 个近邻具有标记的训练样本个数。基于此，条件概率可以通过如下方式估计而得：此时，将所得先验概率以及条件概率(式(40)与式(41))代入式(35)，即可基于式(34)得到所需的多标记分类器。值得一提的是，虽然ML-kNN 算法采用了一阶策略来求解多标记学习问题，即在模型构建过程中忽略标记之间的相互影响。然而，基于该算法的基本思想，可以方便地将其扩展至高阶策略予以实现。例如，在确定事件是否成立时，可以根据中蕴含的信息来进行MAP 推理（而非仅仅考察的取值），即。近期，德国学者E. Hüllermeier 教授指导学生W. Cheng 沿上述思路专文对ML-kNN 算法进行改进，该论文获2009 年欧洲机器学习会议最佳学生论文奖并被推荐到权威期刊《Machine Learning》发表[35]。3.3.2 Rank-SVM该算法的基本思想是采用“最大化间隔（maximum margin）”策略，定义一组线性分类器以最小化式所示的ranking loss 评价指标，并通过引入“核技巧（kernel trick）”处理非线性分类问题[11]。设学习系统由q 个线性分类器组成，其中为与第j 类对应的“权值向量（weight vector）”，而为与第j 类

对应的“偏置（bias）”。基于2.1 节的符号表示，令，算子返回向量内积。给定多标记训练集，Rank-SVM 算法首先按如下方式定义学习系统在样本上的分类间隔：其中，对于相关标记以及无关标记而言，其对应的分类超平面为，因此式(42)考察样本在各“相关-无关”标记配对情况下至分类超平面的距离，将其最小值定义为样本的分类间隔。基于此，学习系统在训练集上的分类间隔对应于：理想情况下，假设上式定义的训练集分类间隔取值为正，即成立。进一步地，通过对线性分类器的参数进行适当的缩放，从而使成立，且存在样本及使该式取等号。此时，最大化式(43)所示的训练集分类间隔可表述为如下优化问题：设训练样本足够充分，即对于所有类别标记，存在使得。此时，上式的优化目标即对应于，相应的优化问题转化为：为了克服式(45)中的算子对优化造成的困难，Rank-SVM 算法将该算法子用求和算子加以近似，进一步地将上述优化问题转化为：为了反映真实情况下式所示的约束无法完全满足的情况，可引入“松弛变量（slack variables）”改写算法对应的优化问题：其中，为松弛变量集合。由上可见，式中所示的目标函数由两个求和项组成。其中，第一项对应于学习系统在训练集上的分类间隔（model complexity），第二项对应于学习系统在训练集上的经验误差（ranking loss），参数C 用于平衡上述两项对目标函数的影响。值得注意的是，式(47)对应于一个具有凸目标函数和线性约束条件的“二次规划（quadratic programming）”问题，但仅仅假设了线性模型用于样本分类。为了使得系统具有非线性分类能力，可以通过引入核技巧将式(47)转化成其“对偶形式（dual form）”求解，具体细节可参见文献[36]。基于2.1 节的符号表示，Rank-SVM 算法还采用了特殊的方式确定阈值函数。具体来说，设为线性函数。其中，为q 维属性向量，其分量对应于分类系统在各类别标记上的输出。相应地，为q维权值向量，为偏置。给定训练集，Rank-SVM 使用线性最小二乘法求解相应参数：通常，可能的取值对应于一个实数区间，算法取该区间的中值用以最小化式(48)。最终，基于式(47)与式(48)的解，则可得到所需的多标记分类器：

3.3.2　局部学习方法

3.3.2.1　节点局部法

3.3.2.2　父节点局部法

3.3.2.3　层局部法

3.3.3　全局学习方法

3.4　本章小结

本章主要介绍了多多标签学习框架的定义、面临的主要问题、常用的评价指标、以及几种代表性学习算法。

第四章　LocalBalance算法

针对案件适用法律的自动识别问题，本文提出了一种局部的层次多标签分类算法。

4.1　算法描述

4.2　本章小结

第五章　法律适用自动识别

本章介绍利用本文提出的层次多标签学习算法进行案件适用法律自动识别的过程，包括裁判文书文本的获取和处理，使用的实验平台，采用的评价指标以及算法取得的实验效果等。

5.1　实验数据

5.1.1　数据获取

本文实验数据取自浙江法院公开网公开的浙江省各级人民法院裁判文书，所有裁判文书均以文本形式存在。为了获取足量的裁判文书样本用于预测模型的学习，本文利用基于jsoup 的爬虫技术实现裁判文书的快速自动获取。jsoup是一款用于处理实际HTML的Java库，它提供了非常方便的API，能够充分利用DOM，CSS以及类似JQuery的方式来提取和操作数据。jsoup 实现了WHATWG HTML 规范，可以将HTML解析成与现代浏览器一样的DOM。jsoup的主要功能包括：

- 从URL、文件或字符串中解析HTML
- 使用DOM遍历器或CSS选择器来查找和提取数据
- 操作HTML元素、属性、文本
- 基于安全的白名单对用户提交的内容进行清理来预防跨站脚本攻击
- 输出整洁的HTML

通过jsoup提供的API，我们首先从浙江法院公开网获取了裁判文书文本所在页面的URL，然后从URL中解析HTML，提取出其中的裁判文书文本，并以文本文件的形式保存至本地。

5.1.2　数据预处理

从浙江法院公开网获取的裁判文书均为文本形式，为非结构化数据，因此需要进行数据的预处理。数据预处理主要完成以下几项工作：

- 提取案件事实描述和适用法律，将裁判文书文本转化为半结构化数据

- 对案件适用法律中的错误和格式不一致进行修正
- 对案件事实描述进行分词和词性标注

浙江省平湖市人民法院
民 事 判 决 书
（2011）嘉平乍商初字第 18 号

原告：XXX。
委托代理人：YYY。
被告：ZZZ。

原告 XXX 为与被告 ZZZ 买卖合同纠纷一案，本院于 2010 年 12 月 31 日立案受理，依法组成合议庭，于 2011 年 7 月 11 日公开开庭进行了审理。原告 XXX 的委托代理人 YYY 到庭参加诉讼，被告 ZZZ 经本院传票合法传唤，无正当理由拒不到庭。本案现已审理终结。

原告诉称，被告于 2007 年在嘉兴港区承接钢贸城水电工程时，向原告赊购管道材料。截止 2008 年 12 月 18 日，被告结欠原告材料款计人民币 62000 元。当时被告约定于 2008 年 12 月 31 日付清，并约定被告如逾期付款由平湖市人民法院审理。但被告未能按时支付，直到 2009 年 1 月 19 日，被告以银行卡汇付了 30000 元，本金 32000 元……

本院认为，合法的买卖关系应受法律保护，被告向原告购买货物后，未及时支付货款，显属欠理，现应承担立即支付货款并负担逾期付款损失的义务。原告的诉讼请求，符合法律规定，本院予以支持。据此，依照《中华人民共和国合同法》第一百六十一条、第一百零七条及《中华人民共和国民事诉讼法》第一百三十条之规定，判决如下：

被告 ZZZ 于本判决生效后十日内支付原告 XXX 货款 32000 元及逾期付款损失……

审判长　AAA
审判员　BBB
审判员　CCC

二〇一一年七月十一日
书记员　DDD

图 5.1: 裁判文书样例

一份典型的裁判文书主要由首部、案件事实描述、裁判理由、裁判结果和尾部组成。图5.1 给出了裁判文书的样例，其中直线下划线标注部分为案件事实描述部分，曲线下划线标注部分为案件适用的法律条文。本文数据预处理的第一步就是要从裁判文书中提取案件事实描述和案件适用的法律条文，前者用于生成表示样本的特征向量，而后者构成了样本的类别标签。通过分析裁判文书的行文结构，我们发现案件事实描述在裁判文书中的位置是固定的，其所在段落的前一段落通常以“本案现已审理终结”或“本案依法缺席审理”结尾，而后一段落通常以“本院认为”开头；类似的，裁判理由部分多以“本院认为”开头，而以“判决如下”结尾，其中案件适用的法律条文部分多以“依

照”、“依据”、“根据”、“按照”等开头，以“之规定”、“的规定”等结尾。根据以上规律，我们可以从裁判文书文本中提取出案件事实描述和适用法律两部分内容，将原始的裁判文书文本初步转化为半结构化数据。

由于裁判文书的书写、电子化等过程基本是人为实现，因此难免会出现笔误、格式错误等现象，在对上述半结构化数据进行进一步处理之前，需要对数据进行清理，减少数据中的错误。与案件事实描述不同，案件适用法律来自既已成文的法律条文，有规范可循，因此对数据的清理主要是对案件适用法律的校验。一种常见错误是法律名称错误或格式不一致，比如“《中华人民共和国民事诉讼法》”被错误写成“《中华人民共和国民诉法》”，“最高人民法院《关于适用〈中华人民共和国婚姻法〉若干问题的解释（二)》”与“《最高人民法院关于适用〈中华人民共和国婚姻法〉若干问题的解释（二)》”格式不一致等；另一种常见错误是法条编号的错误或格式不一致，比如编号中出现非数字、文字编号和数字编号混乱等；第三种错误是间隔符错误，由于案件适用的法律法条通常有多个，一般不同法律之间以逗号间隔，同一法律不同法条之间以顿号间隔，这在后续的标签处理过程中非常重要。由于本文需要处理的裁判文书数量较大，因此我们只对出现次数较多的错误进行了处理，而对出现较少的错误出于时间成本的关系选择了忽略。

本文采用向量空间模型对案件事实描述文本进行表示，首先对案件事实描述的内容进行分词，然后选取特征词将案件事实描述结构化地表示为特征向量。本文选用了哈工大的语言技术平台（Language Technology Platform, LTP）作为语言处理工具。LTP 是一整套中文语言处理系统，制定了基于XML的语言处理结果表示，并在此基础上提供了一整套自底向上的丰富而且高效的中文语言处理模块（包括词法、句法、语义等六项中文处理核心技术），以及基于动态链接库（Dynamic Link Library, DLL）的应用程序接口、可视化工具，并且能够以网络服务（Web Service）的形式进行使用。本文主要利用LTP语言处理系统进行案件事实描述的分词和词性标注，原始的文本内容经过处理转化为一系列词及其词性的列表。分词产生的词即特征词，经过特征选择用于结构化地表示案件事实，词性标注则有助于我们去除一些无意义的词，减少特征词的数量，减少后续的计算量。

5.1.3　标签处理

案件适用法律的自动识别作为一个层次多标签分类问题，其标签空间由案件适用的法律条文组成，根据法律条文的具体程度不同，呈树形结构组织，如图5.2所示。标签树的根节点是一个虚拟节点，其子节点为“法律”节点，如

“《中华人民共和国民事诉讼法》”、“《中华人民共和国担保法》”等；“法律”节点的子节点为“条”节点，每个子节点分别对应于该法律的一条，如“《中华人民共和国民事诉讼法》第十八条”、“《中华人民共和国担保法》第十五条”等；“条节点”的子节点为“款”节点，每个子节点对应于该条的一款，如“《中华人民共和国民事诉讼法》第十八条第一款”、“《中华人民共和国担保法》第十五条第二款”等，若该条没有进一步细分到款，则该“条节点”为叶节点；“款节点”的子节点为“项”节点，每个子节点对应于该款的一项，如“《中华人民共和国民事诉讼法》第十八条第一款第（一）项”、“《中华人民共和国担保法》第十五条第一款第（六）项”等，若该款没有进一步细分到项，则该“款节点”为叶节点；“项节点”均为叶节点。

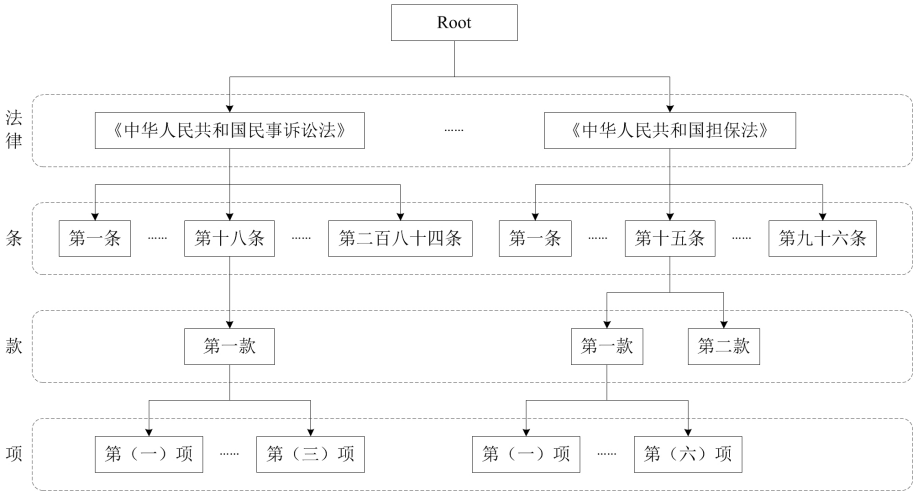

图 5.2: 标签空间树形结构

由于案件可以同时适用多项法律条文，因此样本的类别标签为标签空间的一个子集，包含了该案件样本适用的法律条文对应的标签。考虑标签空间的树形层次结构，如果一个案件适用了某项法律条文，则该案件也适用该法律条文的所有祖先法律条文，因此每个样本的类别标签可以看作标签空间树的一棵子树，并且具有相同的根节点。

在数据预处理的过程中，我们对案件适用法律中的几种错误和不一致情况进行了处理，然而处理后的结果无法直接对应到标签空间中的标签，需要进一步的整理。如前文所述，在裁判文书中，一个案件可能适用多项法律条文，不同法律之间以逗号间隔，同一法律不同条款之间以顿号间隔，标签处理的主要目的是将案件适用法律切分为一项或多项完整的法律条文，从而可以直接对应到标签空间中的标签元素。通过对裁判文书中法律适用的分析，我们总结了不同情形下案件适用法律的书写格式：

- 适用单项法律条文：如“《中华人民共和国合同法》第二百零六条”
- 适用单项法律的多项条款：如“《中华人民共和国刑法》第七十二条第一、二、三款”
- 适用多项法律的单项条款：如“《中华人民共和国合同法》第二百零六条，《中华人民共和国民事诉讼法》第一百四十四条”
- 适用多项法律的多项条款：如“《中华人民共和国合同法》第二百零五条、第二百零六条、第二百零七条，《中华人民共和国民事诉讼法》第一百四十四条”

上述几种情形下，案件适用单项条款时比较容易处理，而适用多项条款时由于会出现省略“第”、“条”、“款”、“项”关键字的情况，简单的切分无法处理，需要从其前后项找到上述关键字来确定其完整法律条文。经过标签处理，原本书写格式多样的法律适用中每一项都转换为完整法律条文，如“《中华人民共和国合同法》第二百零五条、第二百零六条、第二百零七条”将转换为“《中华人民共和国合同法》第二百零五条；《中华人民共和国合同法》第二百零六条、《中华人民共和国合同法》第二百零七条”；“《中华人民共和国刑法》第七十二条第一、二、三款”将转换为“《中华人民共和国刑法》第七十二条第一款；《中华人民共和国刑法》第七十二条第二款；《中华人民共和国刑法》第七十二条第三款”，从而可以直接对应到标签空间的标签元素。

理想地，标签空间由所有既已成文的法律条文构成，然而一方面这会造成标签空间维度过高，使得学习算法在时间和空间上耗费太多资源；另一方面大多数法律条文在数据集中没有出现或者出现极少，无法训练得到有效的预测模型。因此，需要对标签空间进行约简，选择合适的标签组成标签空间。本文基于标签频率进行标签的选择，即选择出现频率达到一定阈值的标签，构成标签空间。这样处理的原因还在于，上一节中对案件适用法律中的错误和格式不一致无法做到完全修正，基于阈值的标签选择可以很容易地将噪声数据过滤。

经过上述处理过程，案件适用法律可以以标签向量的形式表示，向量的每一维代表标签空间中的一个标签元素，即一项具体的法律条文。如果案件适用了某项法律条文，则对应的向量分量值为1，否则为0。

5.1.4　特征处理

本文采用向量空间模型作为文本表示模型，采用TF-IDF方法计算特征词的权重，因此案件事实描述将被表示为文本向量，特征空间由特征词组成，并采

用信息增益算法进行特征词的选择。

在预处理过程中，我们对案件事实描述文本进行了分词和词性标注，词性标注的目的在于对分词得到的词进行初步筛选。在案件事实描述中，动词和名词往往携带了大量信息，而其他类型的词往往缺少实际意义，对分类影响不大，因此我们通过词性标注，去除文本中非动词且非名词的词，这样可以大大减少分词之后文本中词的数量，简化后续计算。

特征词数量	动词名词数量	比例
1000	0	0
2000	0	0
3000	0	0
4000	0	0
5000	0	0
6000	0	0
7000	0	0
8000	0	0

表 5.1: 特征词中动词名词比例

5.2　实验平台

Mulan[15]

5.3　评价指标

传统的分类算法评价指标有

5.4　实验结果

给出算法在实验数据集上的预测表现，与已有算法的性能比较等。

第六章　总结与展望

总结本文的贡献和不足，提出后续的改进方案。

6.1　工作总结

6.2　改进方向

参考文献

[1] 向李兴. 基于自然语义处理的裁判文书推荐系统设计与实现[D]. [S.l.]: 南京大学, 2015.

[2] AGGARWAL C C, ZHAI C. Mining text data[M]. [S.l.]: Springer Science & Business Media, 2012.

[3] TSOUMAKAS G, KATAKIS I. Multi-label classification: An overview[J]. Dept. of Informatics, Aristotle University of Thessaloniki, Greece, 2006.

[4] ZHANG M-L, ZHOU Z-H. A review on multi-label learning algorithms[J]. Knowledge and Data Engineering, IEEE Transactions on, 2014, 26(8): 1819 – 1837.

[5] ROUSU J, SAUNDERS C, SZEDMAK S, et al. Kernel-based learning of hierarchical multilabel classification models[J]. The Journal of Machine Learning Research, 2006, 7: 1601 – 1626.

[6] BI W, KWOK J T. Hierarchical multilabel classification with minimum bayes risk[C] // Data Mining (ICDM), 2012 IEEE 12th International Conference on. 2012: 101 – 110.

[7] BARUTCUOGLU Z, SCHAPIRE R E, TROYANSKAYA O G. Hierarchical multi-label prediction of gene function[J]. Bioinformatics, 2006, 22(7): 830 – 836.

[8] VENS C, STRUYF J, SCHIETGAT L, et al. Decision trees for hierarchical multi-label classification[J]. Machine Learning, 2008, 73(2): 185 – 214.

[9] SALTON G, WONG A, YANG C-S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613 – 620.

[10] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513 – 523.

[11] YANG Y, PEDERSEN J O. A comparative study on feature selection in text categorization[C] // ICML: Vol 97. 1997: 412 – 420.

[12] CHEN W, YAN J, ZHANG B, et al. Document transformation for multi-label feature selection in text categorization[C] // Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. 2007 : 451 – 456.

[13] SPOLAÔR N, TSOUMAKAS G. Evaluating feature selection methods for multi-label text classification[J]. BioASQ workhsop, 2013.

[14] SILLA JR C N, FREITAS A A. A survey of hierarchical classification across different application domains[J]. Data Mining and Knowledge Discovery, 2011, 22(1-2) : 31 – 72.

[15] TSOUMAKAS G, SPYROMITROS-XIOUFIS E, VILCEK J, et al. Mulan: A java library for multi-label learning[J]. The Journal of Machine Learning Research, 2011, 12 : 2411 – 2414.

致　　谢

首先感谢XXX