北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.050

基于类别层次结构的多层文本分类样本扩展策略

李保利

河南工业大学计算机科学系，郑州 450001; E-mail: csblli@gmail.com

摘要 针对大规模多层文本分类训练样本获取代价高、类别分布不均衡等问题，提出并比较几种基于类别层次结构的大规模多层文本分类样本扩展策略，即利用类别层次体系中蕴含的类别名称、描述以及类别间的层次结构关系，从内涵和外延两方面入手构造或扩展类别训练样本。在首次大规模中文新闻信息多层分类评测数据集上，基于外延的局部样本扩展策略取得较好的性能。参测系统在第一级类别和第二级类别上宏平均F1分别为0.8413和0.7139，在10个参赛系统中位列第二。

关键词 多层文本分类；大规模中文新闻分类；中文新闻信息分类；类别层次体系

中图分类号 TP391

Expanding Training Dataset with Class Hierarchy in Hierarchical Text Categorization

LI Baoli

Department of Computer Science, Henan University of Technology, Zhenghou 450001; E-mail: csblli@gmail.com

Abstract As the number of classes is quite large in a hierarchical text categorization problem, it usually costs much to obtain a training dataset of reasonable size and sample distribution. Several strategies are proposed and compared to generate new training samples from the class hierarchy in a hierarchical text classification problem. These solutions try to make full use of the class hierarchy (including class names, their descriptions if any, and relationships between them), and derive new pseudo training samples based on connotations and extensions of classes. Experiments on the dataset of the first large scale Chinese News Categorization at NLPCC2014 show that the localized expanding strategy based on class extensions performs better. Our official system achieved MacroF1 0.8413 and 0.7139 at level 1 and level 2 respectively, which ranked our system the second place among the 10 participating systems.

Key words hierarchical text classification; large scale Chinese news categorization; classification of news in Chinese; class hierarchy

多层文本分类是大规模文本信息组织的关键技术，在Web信息索引、新闻出版、数字图书馆、专利管理等领域具有重要的应用价值。高精度的多层文本自动分类技术是当今大数据时代迫切需要的关键技术之一，已经成为近年来自动分类领域的研究热点。目前，国际上已经连续举办4次大规模多层文本分类评测。第三届国际自然语言处理与中文计算会议(NLPCC2014)举办的大规模中文新闻分类评测，是首次面向中文的大规模多层文本分类技术评测。

与普通的文本分类问题不同，多层文本分类需要考虑的类别总数会达到几千、上万甚至几十万、几百万。这众多类别之间往往还存在这样那样的依赖关系，并由此构成一个复杂的层次化的类别体系。比如在中文新闻信息分类中，分类体系主类表(2012年修订版报批稿)中共有6270个类别，分属5

河南省基础与前沿技术研究项目(112300410007)和河南工业大学高层次人才基金(2012BS027)资助
收稿日期: 2014–07–27; 修回日期: 2014–10–15; 网络出版时间: 2014-11-28 16:32

个不同的层次, 其中第一层有 24 类, 第二层有 340 类。图 1 给出中文新闻信息分类体系中“电子信息产业”这一类别所处的位置。

在解决实际的多层文本分类问题时, 要想取得较优的分类性能, 构建一个适度规模的、样本分布合理的训练样本集合是非常关键的, 同时也是非常困难的。一方面, 因为需要考虑的类别数量巨大, 所以需要标注的样本数量必然很大, 而准确标注较大规模的样本集合本身就很费时费力。另一方面, 由于问题的复杂性, 难以确定一个合理的样本分布, 但至少应当保证样本集合能覆盖所有类别。本文针对这一问题并结合 NLPCC2014 评测数据特点, 提出并比较多种训练样本扩展策略。实验结果表明, 基于外延的局部扩展策略效果较好。

1 相关研究

Silla 等[1]系统地总结了多层分类在文本处理、生物信息学和音频数据处理等多个领域的研究与应用, 给出多层分类的定义, 并在一个统一的框架下描述多层分类问题以及现有方法。何力等[2]全面总结了大规模多层文本分类问题的定义以及求解策略和方法, 并对典型的求解方法进行了对比。陆彦婷等[3]对基于自动生成的层次类别体系的多层分类方法进行了归纳和总结。

按照每次分类所考虑类别的数量, 可以把现有的多层分类方法分为局部策略和全局策略两大类。

局部策略: 也称自顶向下或分而治之策略。该策略构建一系列的分类器, 每个分类器只局部地考虑全部类别中的一小部分。分类时, 从分类体系的根结点出发, 自顶向下逐步确定样本的类别。Koller 等[4]首先采用这种方法。这也是目前使用最广泛的多层分类方法。在过去十多年里, 研究者们探索各种不同的策略来逐级构造分类器, 主要有: 1) 为除根结点之外的每个结点训练一个 2 值分类器, 可以有多种不同的策略用于构造每个 2 值分类器; 2) 为每个非叶结点训练一个多类分类器。

全局策略: 使用一个单一的、相对复杂的分类模型完成多层分类任务。一个最简单的特例是采取单层分类策略, 即完全忽略类别间的层次结构, 孤立地看待各个类别, 常称为 Flat 方法。另外一种策

图 1 中文新闻信息分类体系(部分)

Fig. 1 A part of the class hierarchy of Chinese news

略是改造已有的单层分类算法用于多层分类。

何力等[2]综合分析已有文献中的实验结果, 认为自顶向下策略优于简单的全局策略(Flat 方法)。Babbar 等[5]通过性能边界分析和实验验证, 发现自顶向下的策略适用于规模较大的、不平衡的类别体系, 而 Flat 方法更适合处理类别体系相对平衡的多层分类问题。

国际上, 自 2009 年下半年开始, 先后组织 4 次大规模多层文本分类的公开评测(http://l shtc.iit.demokritos.gr)。在这些评测中, 基于开放目录项目(open directory project, ODP, http://www. dmoz.org/)和维基百科建立测试数据, 对大规模 Web 网页多层分类进行评测。在这 4 次评测中, 最优系统的性能(即严格意义上的准确率)都低于 50%, 尚难以满足实际应用的需求。因此, 对于大规模多层文本分类问题, 非常有继续深入研究的必要性。

中文多层文本分类的研究已有十几年的历史。战学刚等[6]采用自顶向下的策略利用中心点算法(Centroid)以及特征规范化、逐级压缩策略对中文多层分类问题进行了研究。这是国内最早关于中文多层文本分类问题的研究。张志平[7]首先对基于《中文新闻信息分类与代码》的多层文本分类问题进行了研究。鉴于没有基于新闻分类体系的标注语料, 张志平[7]利用分类表中类目名称和类目说明, 获取描述各个类别的主题词作为特征, 然后采用基于中心点算法的自顶向下策略对新闻文本进行分类。吴碧军等[8]基于 SVM 算法也对中文新闻信息自动分类进行了研究, 提出在采用自顶向下分类策略时, 在不同层次重新计算特征向量。由于没有大规模的中文多层文本分类数据, 这些探索性研究与大规模多层文本分类的实际应用之间的距离还比较远。并且, 尚没有对中文多层文本分类技术的大规模比较研究。NLPCC2014 开展的首次大规模中文新闻信息分类评测, 将弥补以上不足。

从近期相关文献可以看出, 自顶向下的策略仍是主流的多层分类方法。近来, 研究者更关注优化类别体系、分阶段分类、重点解决深层类别和稀有类别分类问题、优化各结点分类器组合以及选择更适于多层文本分类的特征等方面, 这些已经成为多层文本分类研究的新趋势。本文结合 NLPCC2014 公开评测, 重点关注如何更好地利用类别层次体系来扩展和丰富训练样本集合, 以期提高分类系统性能。

2 大规模中文新闻信息分类评测任务

第三届国际自然语言处理与中文计算会议首次将大规模中文新闻信息自动分类作为评测任务, 试图通过提供大规模的数据集合和公开评测, 进一步推动中文多层文本分类的研究。实际上, 此次评测任务虽然数据规模相对较大, 但仍然只是一种受控的、简化的多层文本分类评测, 主要体现在: 1) 考虑的类别只是中文新闻信息分类体系的最上层的两级类别中的部分类别(247/340), 类别总数较少; 2) 每个样本只能归属于一个二级类别; 3) 训练样本与测试样本的类别分布基本一致。

训练数据集合中有 42689 篇新闻, 而测试集合中有 11577 篇新闻。训练集合和测试集合中的样本都来自同样的 247 个二级类别, 其分布分别如图 2 和 3 所示。其中, 类别 38.11 (疾病与治疗)所含样本数量最多, 分别有 2828 篇和 755 篇; 而在训练集合中样本数最少的 6 个类别(02.15 (社会公共安全)、14.02 (三农问题)、16.03 (节能)、33.02 (科研队伍)、37.06 (通讯社)、39.14 (体育奖))都只有 6 个样本, 在测试集合中这些类别分别有 2 个样本。图 4 给出在 24 个一级类别中的样本分布情况。类别 38 (医药、卫生)样本最多(分别为 4268 和 1143), 而类别 06 (灾难、事故)样本最少(分别为 26 和 8)。

NLPCC2014 使用在第一级和第二级类别上的准确率(Accuracy)、宏平均精确率(MacroP)、宏平均召回率(MacroR)和宏平均 F1 指数(MacroF1)评价各系统的性能, 这些都是评价单层文本分类系统的性能时常用的指标。

NLPCC2014 大规模中文新闻分类评测提供的训练和测试集合都以 XML 格式存储, 每篇新闻报道包含标题和正文。除训练和测试集合之外, 组织者还提供完整的中文新闻信息分类与代码标准(2012 修订版报批稿), 其中包含有所有类别的名称、编码、描述以及类别间的层次关系。这与国际上的大规模多层分类评测(LSHTC)只给出类别 ID 及类别间的上下位关系不同, 这个更完整的类别层次体系为我们提供了更为丰富的可用于分类的知识和信息。比如, 第一级类别 06 的名称为“灾难、事故”, 相应的描述信息为“包括各种灾难事故、救灾措施、救灾行动, 以及灾后重建的综合性报道”, 这

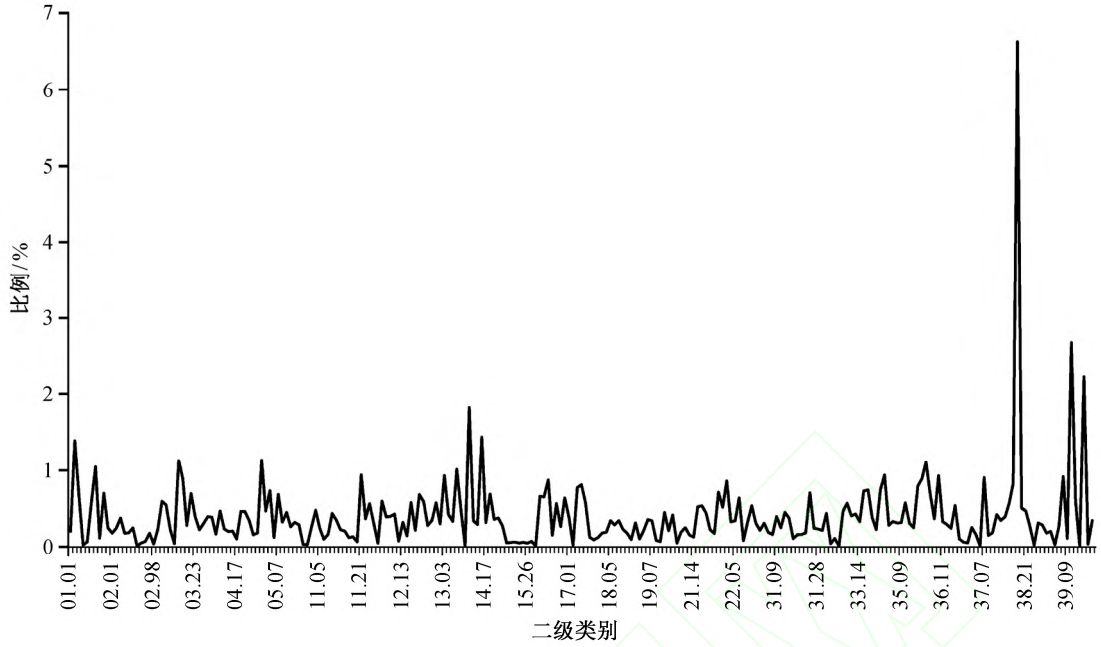

图 2　训练集合中样本类别分布

Fig. 2　Sample distribution of level two classes in the training dataset

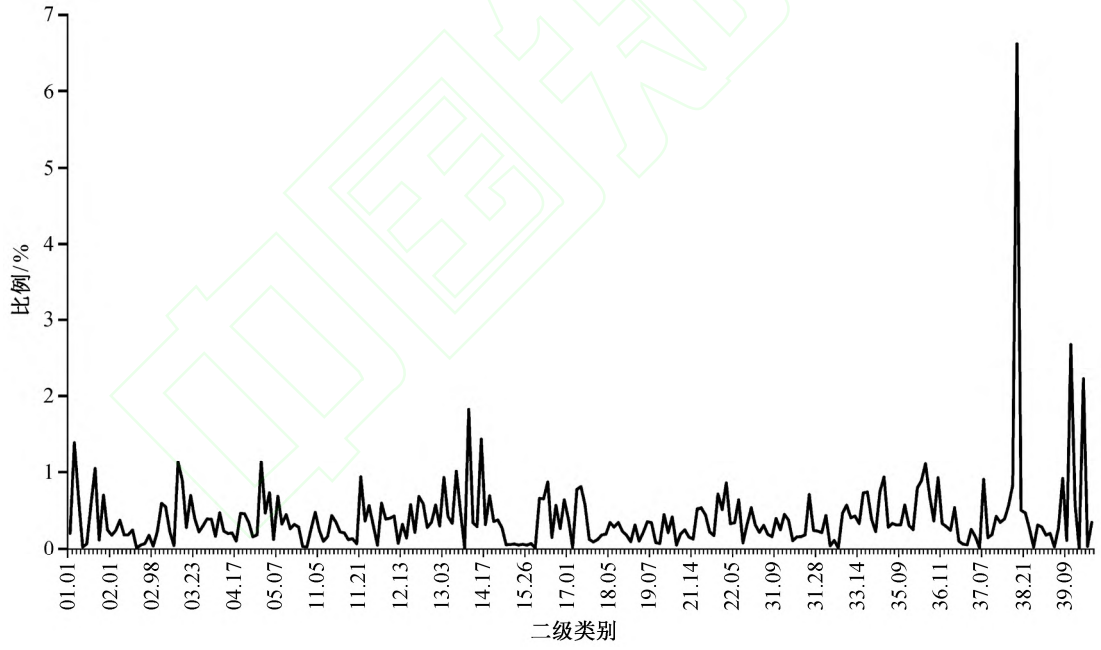

图 3　测试集合中样本类别分布

Fig. 3　Sample distribution of level two classes in the test dataset

里的类别描述信息提供了更全面的分类知识。

需要指出，与实际应用场景一样，在测试集合给出之前，我们一般无法预知测试集合的类别分布是否与训练集合一致，甚至不能臆测测试集合中是否有某些类别没有在训练集合中出现。

以上这些评测任务的特点决定了我们可能采用的解决方案。与此同时，我们也努力解决一些多层文本分类问题共有的难题，比如如何充分利用类别层次体系，如何扩展和构建训练样本集合，等等。

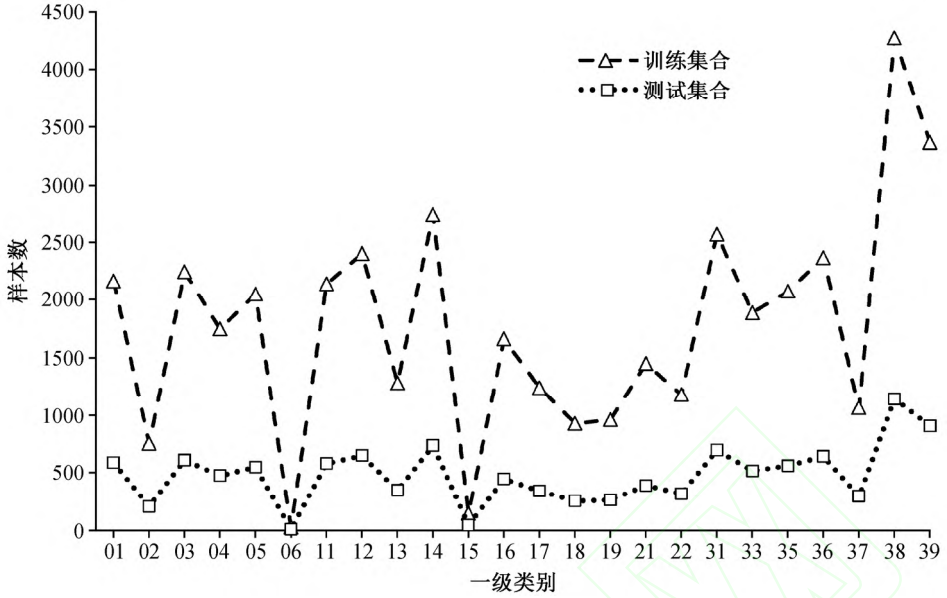

图 4　训练和测试集合中样本类别分布

Fig. 4　Sample distribution of level 1 classes in both the training and the test datasets

3　基于单层分类的中文新闻信息多层分类

通过对 NLPCC2014 大规模中文新闻信息分类评测任务特点的深入分析, 结合以往的文本分类实践, 我们确定采用单层分类的全局策略搭建参测的中文新闻信息多层分类系统, 即将 NLPCC2014 评测任务转化为一个有 247 个或 340 个类别的多类单标记文本分类任务。

图 5 给出我们的参测系统的处理流程。整个处理分为训练和测试两个过程, 其中共同的子处理过程有中文分词和向量化表示。我们使用斯坦福大学自然语言处理研究组开发的词语切分软件(Stanford Word Segmenter, 版本 3.3.1)[9□10](http://nlp.stanford.edu/software/segmenter.shtml)完成中文分词。该软件是基于条件随机场的统计模型构建而成, 我们选用中文宾州树库的切分标准(CTB)。在向量化表示阶段, 将切分后的文本数据样本数字化转换成特征空间上的向量表示, 其中特征即切分后得到的词语, 同时借助简单的基于文档频率的特征选择策略过滤掉所有文档频率为 1 的特征。特征权重的计算采用 TFIDF 的"ltc"格式, 公司如下:

$$f(w_i)=\frac{(1+\log TF_{w_i,d})\cdot\log(\frac{|D|}{DF_{w_i}})}{\sqrt{\sum_j((1+\log TF_{w_j,d})\cdot\log(\frac{|D|}{DF_{w_j}}))^2}}$$

其中, $|D|$为训练集合样本数, $TF_{w_i,d}$ 为词语(特征) w_i 在文档 d 中的出现频次, DF_{w_i} 为 w_i 在训练集合中出现的文档总数, 即文档频率。

在得到样本的数字化向量表示之后, 就可以利用各种不同的机器学习算法学习分类模型, 并据此预测新样本的类别。我们尝试了基于中心点的分类算法、朴素贝叶斯(二项式和多项式)[11]、k 近邻算法[12]、SVM 算法(libsvm[13], http://www.csie.ntu.edu.tw/~cjlin/libsvm)、线性分类算法(liblinear[14], http://www.csie.ntu.edu.tw/~cjlin/liblinear)等, 线性分类算法取得最优的分类性能。

NLPCC2014 评测组织者提供的训练集合中含有 247 个二级类别的标注样本, 但没有为其余 93 个二级类别提供任何训练数据。因此, 在测试集合样本类别分布未知的情况下, 我们期望构造的分类器也有能力识别来自这 93 个二级类别的样本。为解决此问题, 一个简单直接的策略就是扩充训练集

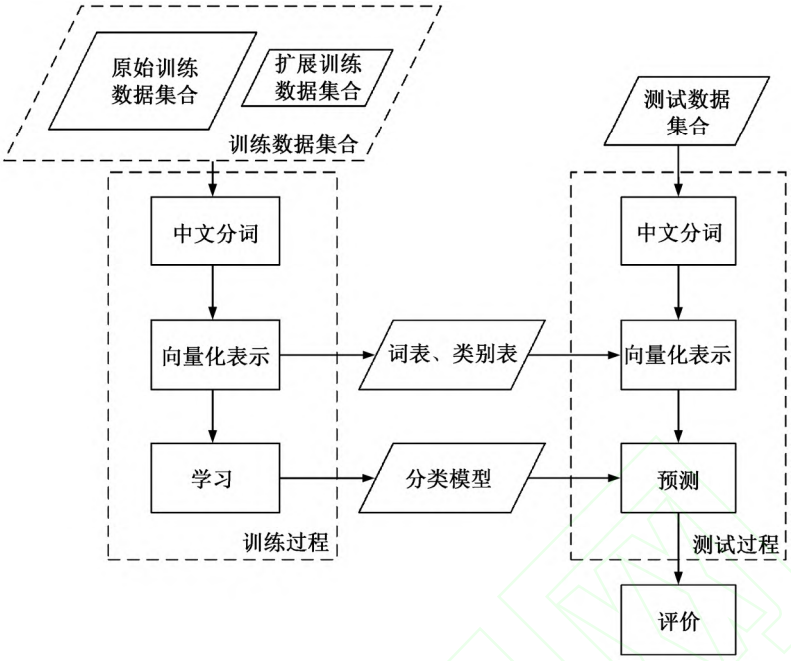

图 5　系统处理流程

Fig. 5　Processing Flow Chart of our system for the NLPCC2014 large-scale Chinese news categorization shared task

合，增加属于原有训练集合中没有出现的类别的样本。我们可以简单地利用类别名称，借助搜索引擎来扩展训练集合，但这样会使用外部的数据资源。尽管 NLPCC2014 评测任务规范里面没有明确禁止使用外部资源，却也没有像以往其他评测任务一样分为封闭测试(只使用组织者提供的数据资源)和开放测试(可以使用任何的数据资源)。我们确定只使用组织者提供的数据资源。组织者随训练数据提供的含有丰富分类知识和信息的中文新闻信息分类与代码标准(2012 修订版报批稿)，为我们简单地扩展训练样本集合提供了可能性。下面，我们将尝试设计多种可能的样本扩展策略，并将得到的新样本与原有的训练集合合并，得到更全面的训练集合。

4　基于类别层次结构的样本扩展策略

NLPCC2014 评测组织者提供的类别层次体系(中文新闻信息分类与代码标准)中包含所有类别的名称、编码、描述以及类别间的层次关系。类别名称是类别的简明表示，在类别描述中则提供了关于类别的更详细描述。显然，我们可以把它们合并起来作为一个特殊的属于该类别的伪样本，但这样得到的样本文本长度往往比较短。我们可以利用类别间的上下位关系，进一步扩展和丰富伪样本的内容。我们认为，一个类别及其包含的所有子孙类别的名称及其描述可以看做是对该类别外延的描述，而把一个类别及其祖先类别的名称及描述看做是对该类别内涵的刻画。由此，我们得到如下一些伪样本构造和训练集合扩展策略。

1) PDT_OFFSPRING: 遍历类别层次树/森林，由每个节点类别构造一个伪样本，内容为该类别及其所有子孙类别的名称和描述。由于评测任务只考虑第二级的类别，所以只考虑第二级以下的节点类别。可以把由第三级以下的节点类别得到的伪样本归入其所属的第二级类别中。

2) PDT_ANCESTOR: 与 PDT_OFFSPRING 类似，由每个节点类别构造的伪样本的内容为该类别及其所有直系祖先节点类别的名称和描述，其中直系祖先节点类别为类别树上从根节点到该节点的路径上的所有节点类别。

3) PDT_ALL: 相当于将以上两种策略合并，即由每个节点类别构造的伪样本的内容为该类别以及其所有直系祖先节点类别和所有子孙节点类别的名称和描述。

假设有一个类别层次树如图 6 所示。当考虑节点类别 K 时，以上 3 种策略包含的类别集合分别为

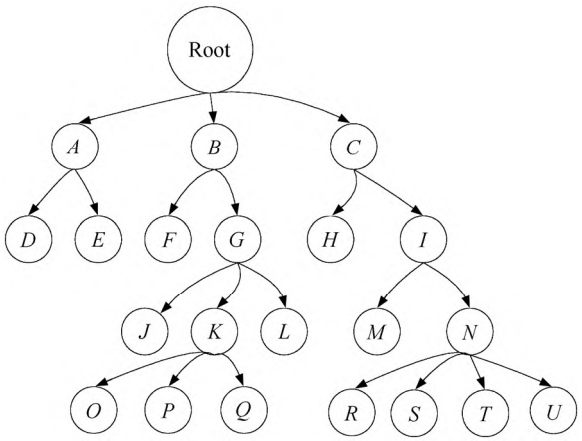

图 6　一个类别层次树
Fig. 6　A sample class hierarchy

{*K*, *O*, *P*, *Q*}, {*K*, *G*, *B*}, {*B*, *G*, *K*, *O*, *P*, *Q*}。

为了避免由下层叶子节点类别构造的伪样本内容过于具体, 我们将候选节点类别进一步限制在第二级类别及其直接子女节点类别(第三级类别), 由此得到以上 3 种策略相应的变种, 分别记为 PDT_OFFSPRING_V1, PDT_ANCESTOR_V1 和 PDT_ALL_V1。在实验中, PDT_OFFSPRING_V1 策略取得最优的性能。由策略 PDT_OFFSPRING, PDT_ANCESTOR 和 PDT_ALL 分别得到 6246 个伪样本, 而由策略 PDT_OFFSPRING_V1, PDT_ANCESTOR_V1 和 PDT_ALL_V1 分别可以得到 2513 个伪样本。

此外, 我们还尝试了只使用类别名称构建伪样本, 不过效果没有使用名称和描述构造伪样本好。

5　实验与分析

我们在 NLPCC2014 数据集合上做了大量实验, 比较了多种不同的分类算法、多种不同的样本扩展策略以及不同的多层分类方法。

5.1　不同分类算法的比较

我们使用单层分类策略来处理大规模中文新闻信息分类问题, 因此首先比较多种不同的分类算法, 试图发现解决此问题更合适的分类算法。

在实验中, 我们考虑的分类算法有以下几种。

CB: 基于中心点的分类算法。

NBB: 二项式朴素贝叶斯算法。

NBM: 多项式朴素贝叶斯算法。

kNN: *k* 近邻算法。*k* 分别取 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60 共 12 个不同的值。

SVM: 支持向量机算法, 使用 libsvm 默认参数。

LINEAR: 线性分类算法, 使用 liblinear 默认参数。

表 1 给出以上算法在测试数据集合上的性能[①], 使用的训练数据是原有的训练样本集合。从表 1 可以看出, 线性分类算法表现最好, 明显优于其他算法, 并且训练时间远小于 SVM 算法。二项式朴素贝叶斯算法 NBB 表现最差, 性能比多项式朴素贝叶斯算法 NBM 低不少。相比较而言, 基于中心点的 CB 算法简单、运行效率高, 性能接近 kNN 以及 SVM 算法(尤其是在第一级类别上), 在第二级类别上的宏平均 F1 指数(MacroF1)甚至优于 SVM 算法

表 1　不同分类算法的性能比较
Table 1　Performance comparison of different categorization algorithms

算法	一级类别				二级类别			
	MacroP	MacroR	MacroF1	Accuracy	MacroP	MacroR	MacroF1	Accuracy
CB	0.7705	0.7741	0.7723	0.7995	0.6565	0.6000	0.6270	0.6782
NBB	0.3234	0.0805	0.1289	0.1626	0.0531	0.0122	0.0198	0.1047
NBM	0.7058	0.5274	0.6037	0.6375	0.4546	0.2365	0.3112	0.4677
kNN (*k*=60)	0.7635	0.7664	0.7649	0.8025	0.6172	0.6106	0.6139	0.6901
SVM	0.8323	0.7468	0.7873	0.8087	0.6947	0.5439	0.6101	0.7039
LINEAR	0.8532	0.8256	0.8392	0.8586	0.7503	0.6616	0.7032	0.7656

① 这里以及其后的实验是在实际的测试集合上得到的结果。在实际测试集合及其标准答案发布前, 我们从原有训练集合中分出一小部分样本集合做验证用。因为得到的结论与这里报告的在实际的测试集合上得到的结论一致, 我们在这一部分报告的都是在实际的测试集合上得到的结果。

和 kNN 算法。

5.2 不同样本扩展方案的比较

在确定最优的分类算法之后, 我们比较了第 4 节提出的多种样本扩展策略。为简化起见, 在实验中, 仅仅使用由不同样本扩展策略得到的伪样本集合作为训练集, 使用测试集合做测试, 采用基于中心点的分类算法 CB, 得到的结果如表 2 所示。与其他实验不同, 由于训练样本集合相对较小, 我们在这个实验中没有用文档频率策略过滤低频特征。

从表 2 可以看出, 策略 PDT_OFFSPRING_V1 表现最好。相比较而言, 基于外延的扩展策略优于基于内涵的扩展策略。此外, 仅仅依靠伪样本集合就能够达到相当好的分类性能, 如策略 PDT_OFFSPRING_V1 在第一级和第二级类别上的准确率分别达到 0.5692 和 0.3491, 远超过猜测所有测试样本为最大类别 38.11 的基准系统的性能(0.0987 和 0.0652), 这也表明基于类别层次结构的样本扩展策略是有效的。

5.3 NLPCC2014 评测结果

基于以上实验结果, 我们确定了构建 NLPCC2014 参测系统的总体设计思路: 采用单层分类策略, 使用线性分类算法, 并借助 PDT_OFFSPRING_V1 策略扩展训练样本集合。

表 3 给出 NLPCC2014 大规模中文新闻信息分类评测任务的正式评测结果, 共有 10 个系统提交了结果。表 3 中每一行对应一个参测系统, 结果按照各系统在第二级类别上宏平均 F1 指数(MacroF1)和准确率(Accuracy)降序排列。系统 7 由于输出格式错误, 评测结果都为 0。我们的参测系统编号为 2, 总体性能在 10 个参测系统中排名第二, 其中在第一级类别上的 MacroF1 和 Accuracy 值分别为 0.8413 和 0.8604, 分别落后第一名(系统 9) 0.0266 和 0.0244, 高于第三名(系统 10) 0.0821 和 0.0700。在第二级类别上, 我们的参测系统的 MacroF1 和

表 2 不同样本扩展方案的比较

Table 2 Performance comparison of different pseudo sample generation strategies

方法	一级类别				二级类别			
	MacroP	MacroR	MacroF1	Accuracy	MacroP	MacroR	MacroF1	Accuracy
PDT_OFFSPRING	0.5416	0.5185	0.5298	0.5552	0.3782	0.2988	0.3338	0.3257
PDT_OFFSPRING_V1	0.5673	0.5474	0.5572	0.5692	0.4214	0.3136	0.3596	0.3491
PDT_ANCESTOR	0.5163	0.5023	0.5092	0.5026	0.3725	0.3082	0.3373	0.2964
PDT_ANCESTOR_V1	0.4955	0.4656	0.4800	0.4732	0.3487	0.2888	0.3159	0.2592
PDT_ALL	0.5360	0.5227	0.5293	0.5225	0.3884	0.3210	0.3515	0.3161
PDT_ALL_V1	0.5433	0.5336	0.5384	0.5473	0.4063	0.3202	0.3582	0.3350

表 3 NLPCC2014 大规模中文新闻信息分类评测结果

Table 3 The official result of the NLPCC2014 large-scale Chinese news categorization shared task

排名	系统编号	一级类别				二级类别			
		MacroP	MacroR	MacroF1	Accuracy	MacroP	MacroR	MacroF1	Accuracy
1	9	0.8725	0.8633	0.8679	0.8848	0.7772	0.7726	0.7749	0.8161
2	2	0.8513	0.8315	0.8413	0.8604	0.7487	0.6822	0.7139	0.7720
3	10	0.7422	0.7770	0.7592	0.7904	0.5646	0.6238	0.5927	0.6294
4	5	0.7336	0.7076	0.7204	0.7507	0.6024	0.5240	0.5604	0.6249
5	4	0.7260	0.7023	0.7140	0.7450	0.5922	0.5203	0.5539	0.6185
6	8	0.6536	0.6428	0.6481	0.7197	0.5073	0.4711	0.4885	0.5874
7	6	0.5817	0.4576	0.5123	0.5363	0.4577	0.2430	0.3174	0.3658
8	3	0.7389	0.6616	0.6981	0.7339	0.1352	0.1336	0.1344	0.1664
9	1	0.3758	0.2453	0.2969	0.2856	0.0761	0.0867	0.0892	0.0761
10	7	0	0	0	0	0	0	0	0

Accuracy 值分别为 0.7139 和 0.7720, 分别落后第一名 0.0610 和 0.0441, 高于第三名 0.1212 和 0.1426。结果显示, 我们选择的处理策略对解决 NLPCC2014 这一具体的大规模中文新闻信息分类任务是有效的。

对照表 3 与表 1 可以看到, 采用单层分类策略, 使用简单的基于中心点的分类算法, 就能取得优于实际参测系统第三名(系统 10)的结果。此外, 在包含伪样本的扩展训练集合上得到的结果略优于在原有训练集合上的结果(在第二级类别上的 MacroF1 和 Accuracy 取值分别为 0.7032 和 0.7656), 表明基于类别层次结构的样本扩展策略是有效的。鉴于实际的测试数据与训练数据的类别分布基本一致, 我们也尝试了只用属于已知类别的伪样本扩展训练集合, 可以略微提高一点系统的性能。

5.4 与其他多层分类方法的比较

除以上实验外, 在正式评测后, 针对这一多层分类问题, 我们还比较了基于单层分类的全局策略与自顶向下的局部策略。自顶向下的局部策略是沿类别层次体系构建一系列的分类器, 每个分类器只局部地考虑全部类别中的一小部分。分类时, 从分类体系的根结点出发自顶向下逐步确定样本的类别。主要有两种策略逐级构造分类器: 1) 为除根结点之外的每个结点训练一个 2 值分类器; 2) 为每个非叶结点训练一个多类分类器。第 1 种策略适合于多标记分类(一个样本可以属于多个类别), 在用来解决 NLPCC2014 分类任务时, 需要从多个可能类别中选择一个最可能的。另外, 在实际的实验中发现有很多样本无法分入任何类别, 这时需要引入一些复杂的机制进行处理。因此, 我们在这里只报告基于第 2 种策略的自顶向下多层文本分类方法的结果, 如表 4 所示。在本实验中, 我们尝试用两种不同算法在每个非叶节点构造多类分类器: 基于中心点的算法 CB 与线性分类算法 LINEAR。

对比表 4 与表 1, 可以发现基于 CB 和 LINEAR 的算法表现稳定, 使用基于多类分类器的自顶向下多层文本分类方法的性能基本与基于单层分类的方法一致, 这从另一个角度表明 NLPCC2014 评测任务还不是一个真正意义上的大规模多层文本分类任务。

6 结语

本文报告了我们参加 NLPCC2014 举办的首次大规模中文新闻信息分类评测的系统的设计思路以及实验和评测结果。根据对评测任务和数据的深入分析, 我们确定采用单层分类的策略搭建参测系统。此外, 为使训练得到的分类器能识别所有可能类别的样本, 我们提出并比较了几种基于类别层次结构的样本扩展方案, 并将构造的伪样本并入原有训练集合。正式的参测系统在第一级类别和第二级类别上宏平均 F1 分别为 0.8413 和 0.7139, 在 10 个参赛系统中位列第二, 表明我们确定的解决方案是有效的。

今后, 我们将尝试其他可能的样本扩展方案, 比如给予类别名称和描述不同的权重, 剔除描述中的噪音信息, 等等。另外, 我们也准备利用组织者提供的大规模中文新闻分类数据开展更多的相关研究工作, 如探索其他可能的多层文本分类策略等。当然, 我们更期待参加第二次大规模中文新闻信息分类评测。

参考文献

[1] Silla C N, Freitas A A. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, 2011, 22: 31–72

[2] 何力, 贾焰, 韩伟红, 等. 大规模层次分类问题研究及其进展. 计算机学报, 2012, 35(10): 2101–2115

[3] 陆彦婷, 陆建峰, 杨靖宇. 层次分类方法综述. 模式识别与人工智能, 2013, 26(12): 1130–1139

[4] Koller D, Sahami M. Hierarchically classifying documents using very few words // Proceedings of the 14th international conference on machine learning

表 4 基于多类分类器的自顶向下多层文本分类方法的性能

Table 4 Performance of two top-down hierarchical text classification algorithms

算法	一级类别				二级类别			
	MacroP	MacroR	MacroF1	Accuracy	MacroP	MacroR	MacroF1	Accuracy
CB	0.7712	0.7740	0.7726	0.7994	0.6569	0.5994	0.6282	0.6776
LINEAR	0.8534	0.8256	0.8393	0.8587	0.7515	0.6616	0.7037	0.7657

(ICML-1997). San Francisco: Morgan Kaufmann, 1997: 170−178

[5] Babbar R, Partalas I, Gaussier E, et al. On flat versus hierarchical classification in large-scale taxonomies // Burges C J C, Bottou L, Welling M, et al. Advances in neural information processing systems (NIPS-2013). Lake Tahoe: NIPS Foundation, 2013: 1824−1832

[6] 战学刚, 林鸿飞, 姚天顺. 中文文献的层次分类方法. 中文信息学报, 1999, 13(6): 20−25

[7] 张志平. 基于“中文新闻信息分类与代码”文本分类. 太原理工大学学报, 2010, 41(4): 402−405

[8] 吴碧军, 李涓子, 金鑫. 分层特征计算和错误控制的层次分类方法. 计算机科学, 2010, 37(10): 165−168

[9] Tseng H, Chang P, Andrew G, et al. A conditional random field word segmenter // Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. Jeju Island: 2005: 168−171

[10] Chang P C, Galley M, Manning C D. Optimizing Chinese word segmentation for machine translation performance // Proceedings of the Third Workshop on Statistical Machine Translation. Columbus: Association for Computational Linguistics, 2008: 224−232

[11] McCallum A, Nigam K. A comparison of event models for naive bayes text classification // Proceedings of the AAAI-1998 Workshop on Learning for Text Categorization. Madison: 1998: 41−48

[12] Li Baoli, Lu Qin, Yu Shiwen. An adaptive k-nearest neighbor text categorization strategy. ACM Transactions Asian Language Information Processing, 2004, 3(4): 215−226

[13] Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): Article No. 27

[14] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: a library for large linear classification. Journal of Machine Learning Research, 2008, 9: 1871−1874