大规模层次分类问题研究及其进展

何 力 贾 焰 韩伟红 谭 霜 陈志坤
(国防科学技术大学计算机学院 长沙 410073)

摘 要 随着信息技术的发展,互联网数据急剧增长.为了有效地组织和管理这些海量网页信息,通常按照一个大规模的概念或主题类别层次对网络上的信息进行分类,以更好地搜索和访问这些网络资源.在这个过程中,大规模层次分类问题研究如何将互联网上的网页文档准确地分到类别层次中的各个类别.该文对大规模层次分类问题进行了分析.首先,给出了大规模层次分类问题的定义,分析了大规模层次分类问题的求解策略;其次,对大规模层次分类问题的求解方法加以分类,在分类基础上,介绍了各种典型的求解方法并进行了对比;最后总结了各种大规模层次分类问题求解方法并指出了未来的研究方向.

关键词 文本分类;大规模层次分类;类别层次;类别层次树
中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2012.02101

Research and Development of Large Scale Hierarchical Classification Problem

HE Li JIA Yan HAN Wei-Hong TAN Shuang CHEN Zhi-Kun
(*School of Computer Science, National University of Defense Technology, Changsha* 410073)

Abstract With the development of information technology, Web information management and access become much difficult to some extent as rapid increase in Internet data. A large scale class hierarchy of concepts or topics was used to label the web information to make information access easier. In this process, large scale hierarchical classification problem researches how to classify the Web documents into the categories among the class hierarchy, which is surveyed in this paper. Firstly, a definition of large scale hierarchical classification problem is proposed, which is used to describe the problem in abstraction level. Meanwhile, strategies for conquering the problem are also investigated. Secondly, classification of solving methods for this problem is analyzed, and on the basis of the classification, many typical solving methods are introduced and compared. Lastly, future research trends of the solving methods for this problem are reviewed.

Keywords text categorization; large scale hierarchical classification; class hierarchy; tree-structured class hierarchy

1 引 言

随着信息技术的发展,互联网数据急剧增长.第29次CNNIC调查结果显示①,截至2011年12月底,中国网页数量为866亿个,比2010年同期增长44.3%.为了有效地组织和管理这些海量网页信息,通常的做法是按照一个概念或主题类别层次将这

收稿日期:2012-06-30;最终修改稿收到日期:2012-08-15.本课题得到国家"八六三"高技术研究发展计划项目基金(2010AA012505、2011AA010702、2012AA01A401、2012AA01A402)、国家自然科学基金(60933005)、国家科技支撑计划(2012BAH38B04)及国家242信息安全计划(2011A010)资助.**何 力**,男,1984年生,博士研究生,主要研究方向为网络与信息安全、数据库与数据挖掘.E-mail:hl19840507@163.com.**贾 焰**,女,1961年生,博士,教授,主要研究领域为网络与信息安全、数据库与数据挖掘、社会网络.**韩伟红**,女,1973年生,博士,副教授,主要研究方向为网络与信息安全、数据库与数据挖掘.**谭 霜**,男,1984年生,博士研究生,主要研究方向为网络与信息安全、数据库与数据挖掘.**陈志坤**,男,1985年生,博士研究生,主要研究方向为网络与信息安全、数据库与海量数据处理.

① 中国互联网网络信息中心(CNNIC), http://www.cnnic.cn/dtygg/dtgg/201201/W020120116337628870651.pdf

些网页信息组织为网络资源分类目录,以更好地搜索、访问和管理这些网络资源,例如开放目录专案(Open Directory Project,ODP 目录)①、雅虎目录(Yahoo!Directory)②等.要自动构建网络资源目录,就需要实现对互联网上未知类别信息的分类,这里的信息类别一般被组织为一个层次式结构,典型的是一棵树(tree)或者有向无环图(directed acyclic graph),这种类别层次一般规模巨大,其类别数目可以达到数万、甚至数十万之多.大规模层次分类(large scale hierarchical classification)就是按照这样一个规模巨大的类别层次,采用机器学习的方法指定网页在类别层次中所属的类别.除了构建网络资源目录外,大规模层次分类也是很多网络应用的基础,包括信息检索、网络资源管理、绿色上网、网络信誉管理、有害和敏感信息过滤等.

在大规模层次分类领域,目前有很多研究,但由于这些研究分散在不同领域,导致一个领域的研究人员往往不知道另一领域研究人员开发的方法.同时由于对大规模层次分类问题的概念、任务、目标没有明确定义,导致难以采用统一的实验和评价标准评估大规模层次分类方法.鉴于目前还没有研究者对大规模层次分类进行系统的总结,本文对大规模层次分类领域的相关研究作了总结和归类,以利于后续相关研究的开展.本文将对大规模层次分类问题的定义、大规模层次分类的解决策略和主要方法、不同方法的特点以及之间的区别等进行分析,最后讨论大规模层次分类领域存在的问题和未来研究方向.

2 大规模层次分类问题

2.1 大规模层次分类问题的定义

广义的大规模层次分类是指按照一个规模巨大的类别层次,指定未知类别对象在类别层次中所属的类别.这里的分类对象可以是文本对象,如维基百科中的文档,也可以是多媒体信息,如网页上的音乐、图像、视频等.分类方式可以是人工分类,也可以是基于机器学习的自动分类或者带有专家验证的自动分类.类别层次可以是专家编制,也可以是在分类过程中通过聚类方法学习产生.本文仅讨论狭义的大规模层次分类问题,即分类对象为文本对象,分类方式是基于机器学习的自动分类,类别层次由专家预先编制.

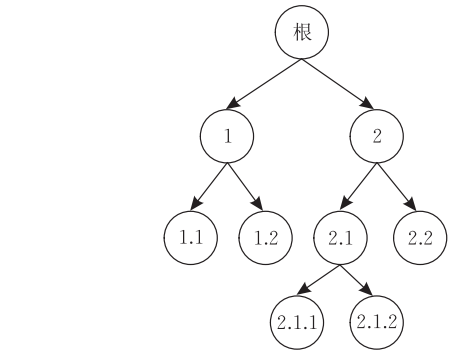
关于层次分类问题的定义,Silla 等人[1]提出可以用类别层次的结构类型、实例的类别数目、实例的标签路径深度 3 个属性描述一个层次分类问题,而在大规模层次分类中,类别往往被组织成一个多维度多层次的类别体系,并且类别之间的关系复杂,类别之间可能形成环路.针对大规模层次分类的这些特点,我们在层次分类问题定义的基础上进行扩展,从以下 4 个属性描述大规模层次分类问题.

属性 1. 类别层次的结构类型 H,H 代表类别层次中类别之间的关系,可能的取值有:

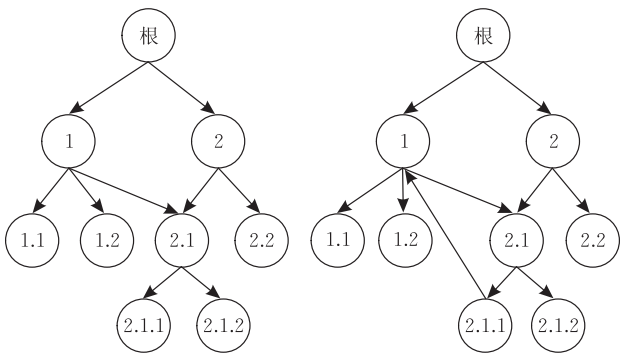
(1) T(树). 类别层次被组织为树形结构,如图 1(a)所示.

(2) DAG(有向无环图). 类别层次被组织为有向无环图的结构,如图 1(b)所示,显然,树是一类特殊的有向无环图.

(3) DCG(有向有环图). 类别层次被组织为有向有环图,即类别层次中可以出现环路,如图 1(c)所示.



(a) 树



(b) 有向无环图 (c) 向向有环图

图 1 类别层次结构图

属性 2. 类别层次的维度 P,P 描述类别层次是否包含多个维度,可能的取值有:

(1) SD(单一维度). 所有类别处于单一维度.

(2) MD(多个维度). 类别层次由多个维度的类别组成,例如 ODP 目录就包括主题、地区、语言 3 个

① http://www.dmoz.org/
② http://dir.yahoo.com/

类别维度，在进行多维度分类的时候应该按照各个不同维度分别进行分类的学习和预测.

属性 3.　实例的类别数目 L，L 描述是单标签分类问题，还是多标签分类问题. 在层次分类问题中，指一个实例可以被赋给类别层次中的一个标签路径，还是可以被赋给多个标签路径，可能的取值有：

(1) SPL(单标签路径). 单标签分类，所有实例只能被赋给类别层次中一条唯一的标签路径.

(2) MPL(多标签路径). 多标签分类，实例可以被赋给类别层次中的多条标签路径，例如在图 1(a)中，实例 d_1 同时属于类别 1.2 和类别 2.1.2，即 d_1 有两条标签路径，分别是“根-1-1.2”和“根-2-2.1-2.1.2”.

属性 4.　实例的标签路径深度 D，D 描述实例标签路径的深度，可能的取值有：

(1) FD(全深度标签). 所有实例必须被赋给全深度标签，每个实例的标签路径必须从根类别开始，以叶子类别结束，即所有的实例均处于叶子类别.

(2) PD(部分深度标签). 实例可以被赋给部分深度标签，实例的标签路径可以以非叶子类别结束，即实例可以处于类别层次中的中间节点上. 例如在图 1(a)中，实例 d_2 属于类别 2.1，但既不属于类别 2.1.1，也不属于类别 2.1.2，则 d_2 的类别标签路径为“根-2-2.1”，是一个部分深度标签. 部分深度标签分类问题是指可以将文档分到非叶子节点就停止分类，而不是必须分至叶子节点. 例如对于 d_2，在将 d_2 分至类别 2.1 时，分类就应该结束.

对于一个大规模层次分类问题 G，均可以利用 H、P、L、D 这 4 个属性进行描述. 因此，我们可以用四元组$\langle H,P,L,D\rangle$表示现有的任意一类大规模层次分类问题. 例如，$G=\langle T,SD,SPL,FD\rangle$描述的就是一类最常见的大规模层次分类问题，即类别层次是一个单一维度的树形结构，实例的类别唯一，并且所有实例均位于叶子节点. 由于这一类问题最具有代表性，因此在本文后面的讨论中，如果没有对问题进行特别说明，就是指这一类大规模层次分类问题.

2.2　大规模层次分类问题求解方法分类框架

大规模层次分类方法主要是针对大规模层次分类问题中类别层次结构巨大这一特点进行研究. 因此，大规模层次分类问题求解方法的不同主要体现在对大规模层次分类问题的处理策略上，目前有 3 种处理策略：全局处理策略(overall-conquer)、分而治之的策略(divide-and-conquer)、化繁为简的策略(reduce-and-conquer). 下面我们分别介绍这 3 种策略以及相应的分类方法.

整体处理策略：将所有类别作为一个整体，在整个数据集上进行分类学习，然后对待分类文档进行分类，我们将采用这种策略的方法称为全局分类方法.

分而治之策略：按照类别层次将一个大规模的全局分类问题分解为一个个小规模的局部分类问题，然后分别进行分类学习，对待分类文档进行自上而下的分类. 我们将采用这种策略的方法称为自上而下分类方法.

化繁为简的策略：通过搜索类别层次中所有与待分类文档相关的类别，然后在所有相关类别上进行分类学习和预测，将一个大规模的分类问题降低为一个小规模的分类问题. 我们将采用这种策略的方法称为收缩分类方法.

我们以问题的处理策略作为大规模层次分类问题求解方法的划分依据，更能体现这些方法的区别，同时这种划分方式也能很好地归类现有大规模层次分类方法，本文基于这一划分策略形成了比较完备的分类框架，如图 2 所示，其中，全局分类方法又分为平面分类方法和 Big-bang 分类方法，自上而下分类方法又分为基于二元分类器的自上而下分类方法、基于多元分类器的自上而下分类方法和基于类别层次优化的自上而下分类方法，这些方法将在第 3 节进行详细介绍. 本文的第 3 节将围绕大规模层次分类问题求解方法的分类框架展开.

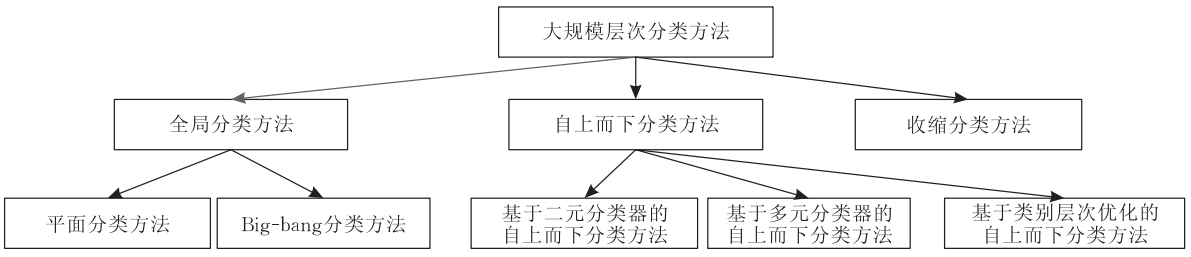


图 2　大规模层次分类问题求解方法分类框架

3 大规模层次分类问题求解方法

3.1 全局分类方法

全局分类(Global Approaches,Global)是指将所有类别作为一个整体,在整个类别层次上进行分类学习和预测.如果在分类学习过程中,不考虑类别间的层次关系,将所有类别看作相互独立的平级类别,采用传统的文本分类算法进行分类,则将这一类方法称为平面分类方法(Flat Approaches,Flat).如果在分类的学习过程中,考虑类别间的层次关系,利用类别之间的语义关系进一步帮助分类学习和预测,我们称这一类方法为 Big-bang 分类方法(Big-bang Approaches,Big-bang).下面我们从这两方面总结已有的全局分类方法.鉴于 Silla 等人[1]采用分类器节点分布图可以形象地描述层次分类方法,本文在图 3、图 4、图 5、图 7、图 8 中也采用了这种方式形象地表示大规模层次分类方法.

3.1.1 平面分类方法

平面分类方法不考虑类别层次,将类别树中所有叶子节点看作相互独立的平级类别,作为一个多类别分类问题(multi-class classification)进行处理,如图 3 所示,虚线框表示分类器的学习范围.

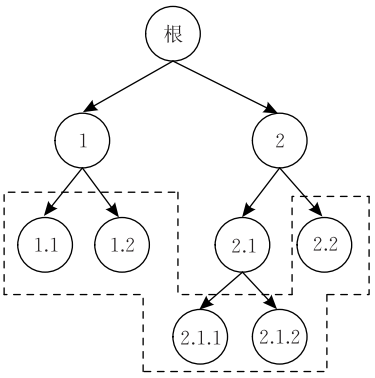


图 3 平面分类

平面分类方法可以直接采用传统的文本分类算法将文档分类至叶子节点,实现全深度标签分类.对于部分深度标签分类问题,我们可以通过为每一个包含有实例的非叶子节点引入一个新的子节点,然后将该非叶子节点上的实例转移到新的叶子节点上面,从而将部分深度标签分类问题转化为全深度标签分类问题,进而采用平面分类方法.例如,如果图 3 中的节点 2 包含有一个实例 d,则可以通过引入一个新的类别节点 2.3,将 d 转移到 2.3 上,从而转化为一个全深度标签分类问题,如图 4 所示.

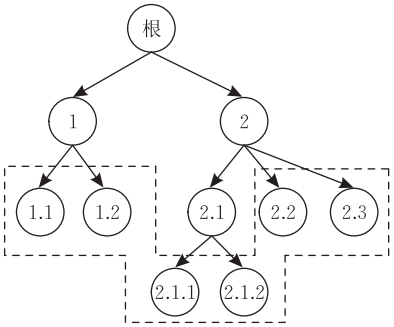


图 4 部分深度标签的平面分类

平面分类方法可以直接采用机器学习当中许多经典的分类算法,如基于质心的分类算法[2]、最近邻分类算法[3]、朴素贝叶斯分类算法[4]、Rocchio 算法[5]、基于人工神经网络的分类算法[6].在将这些算法应用到大规模层次分类问题时,由于数据倾斜和数据稀疏等问题,算法的分类性能往往会变差.数据倾斜问题是由于在大规模层次分类中类别数目巨大,如果将某一个类别的实例作为正样本,将其余类别的实例作为负样本,将导致负样本数量远远超过正样本数量,数据稀疏问题是由于大规模层次分类中的文档往往比较短,大量短文档导致文档特征矩阵稀疏.对于这些问题,研究者一般要对算法进行一些优化.各种基于传统文本分类算法的平面分类方法如表 1 所示.

表 1 平面分类方法

基本算法	平面分类方法
基于质心的分类算法	Guan 等人[2](2009)
最近邻分类算法	Wang 等人[3](2011)
朴素贝叶斯分类算法	Zhang 等人[4](2009)
基于人工神经网络的分类算法	Christophe[6](2011)
Passive-Aggressive (PA) 算法	Madani 等人[7](2010)

Hu 等人[2]采用近似文档频率逆类别频率的方法创建类别的质心权重向量,然后根据文档权重向量和类别质心向量之间的点积判断文档类别.对于类别 C_j,其质心向量中的每个元素 w_{ij} 是词 t_i 关于类别 C_j 的特征权重,若 t_i 在 C_j 中的文档频率为 DF_{ti}^{j},C_j 包括的文档总数为 $|C_j|$,训练集类别总数为 $|C|$,出现过 t_i 的类别数目为 $|CF_{ti}|$,则 w_{ij} 的计算方法如式(1)所示,b 是用来调节文档频率和逆类别频率在公式中重要程度的参数,即越多出现在类别 C_j,越少出现在其它类别的词,其权重越大

$$w_{ij}=b^{\frac{DF_{t_i}^{j}}{|C_j|}}\times\log\left(\frac{|C|}{CF_{t_i}}\right) \tag{1}$$

Wang 等人[3]提出了一种最近邻算法的投票策

略，每个实例投票的权重与该实例和测试文档之间的距离相关，对于一个测试文档，首先由相似性度量算法 BM25[3] 计算距离该文档最近的 k 个邻居，然后根据各个邻居同该文档之间的距离计算该文档的类别得分，距离测试文档越近的邻居其投票权重越大. 若测试文档为 S_e，邻居为 S_i，S_e 对类别 c 的得分为 $score(c)$，则 $score(c)$ 的计算方法如式(2)所示，其中 $\gamma(s_i,c)$ 判断 c 是否为 S_i 的类别标签，α 是一个正实数，用来调节各个邻居的投票权重.

$$score(c)=\sum_{i=1}^{k}\gamma(s_i,c)BM25(s_i,s_e)^{\alpha} \tag{2}$$

Zhang 等人[4] 针对朴素贝叶斯模型在 Web-Scale 分类中性能变差的问题，提出一种参数平滑方法，以提高朴素贝叶斯模型的分类准确率. 标准朴素贝叶斯模型采用拉普拉斯平滑方法计算词条在一个类别中的出现概率，对于词 w_i 和类别 c_u，若 w_i 在 c_u 中的出现次数为 N_u^i，c_u 中的词条总数为 N_c，记 $p(w_i|c)$ 为 w_i 在 c_u 中的出现概率，则拉普拉斯平滑计算法方法如式(3)所示

$$p(w_i\mid c)=\frac{N_c^i+1}{N_c+V} \tag{3}$$

在大规模层次分类当中，由于数据稀疏问题，拉普拉斯平滑可能导致一个类别中出现的词和未曾在这个类别中出现过的词的先验概率值比较接近，针对这个问题，Zhang 提出用 γ 平滑方法计算词条出现概率 z_u^i，如式(4)所示，其中 γ 是一个足够小的常数，由于 γ 取值可以足够小，相比拉普拉斯平滑，采用 γ 平滑一定程度上可以增强一个类别中出现词和未出现词的概率值的大小区别.

$$z_u^i=\begin{cases}\log p(w_i\mid c_u)=\log N_u^i-\log N_u, & \text{若 } N_u^i\neq 0\\ \gamma\Big/\sum\limits_{j;N_u^j=0}1, & \text{否则}\end{cases} \tag{4}$$

Madani 等人[7] 采用 PA 算法[8] 计算一个特征-类别权重矩阵，矩阵的每一行对应一个特征，每一列对应一个类别，通过对文档特征向量和权重矩阵之间进行点积运算计算文档同各个类别的相似性得分，算法通过迭代更新特征-类别权重矩阵来提高分类准确率. 还有其它基于传统文本分类算法的 Flat 分类方法，在此不一一详述.

3.1.2　Big-bang 分类方法

Big-bang 分类面向整个类别层次学习一个分类模型，如图 5 所示，Big-bang 方法可以将文档分到类别层次中任意一级节点，因此支持全深度标签分类和部分深度标签分类.

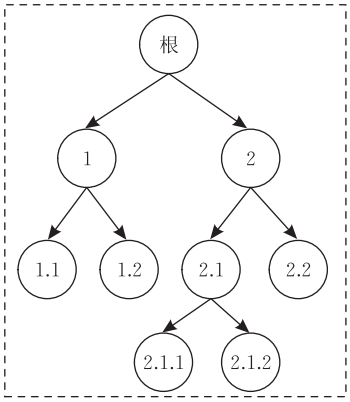


图 5　Big-bang 分类

Big-bang 方法在将传统的文本分类方法应用到大规模层次分类中的时候，会考虑整个类别层次，利用类别之间的父子派生关系进行分类的学习和预测，结合类别层次的特点对一般分类算法进行一些修改. 各种 Big-bang 分类方法如表 2 所示.

表 2　Big-bang 分类方法

基本算法	Big-bang 方法
基于质心的分类算法	Miao 等人①(2010)
基于关联规则的分类算法	Wang 等人[9](2001)
SVM 算法	Cai 等人[10](2004)
基于规则的分类算法	Sasaki 等人[11](1998)

Miao 等人① 根据文档向量和每个类别及其父类别向量的相似性之和进行分类预测，其主要工作是在计算文档和一个类别之间相似性的时候，将文档同该类别的父类别的相似性也考虑在内，即用一个类别与其父类别的质心向量之和代替该类别的质心向量来衡量其和测试文档之间的相似性. 对于类别 c，若 c 的父类别为 $p(c)$，c 和 $p(c)$ 的质心向量分别为 $\boldsymbol{w}_c$ 和 $\boldsymbol{w}_{p(c)}$，则文档 x 的预测类别 $\hat{w}$ 的计算方法如式(5)所示：

$$\hat{w}=\arg\max_c(\boldsymbol{w}_c+\boldsymbol{w}_{p(c)})\cdot x \tag{5}$$

Wang 等人[9] 利用关联规则挖掘算法处理层次文本分类问题，其主要工作是使基于关联规则的分类算法支持层次分类，他首先利用类别层次结构建立关联规则，并对规则进行排序和筛选，然后根据关联规则对文档进行分类，可以将文档直接分到类别层次中的任意一级节点.

Cai 等人[10] 基于 SVM 构造层次分类方法，其主要工作是利用了类别层次信息构造判别函数，即综合考虑祖先类别信息判断文档类别，首先由 SVM 模型计算文档在某个类别以及该类别所有祖先类别

① http://lib.iit.demokritos.gr/system/files/XipengQiu.pdf

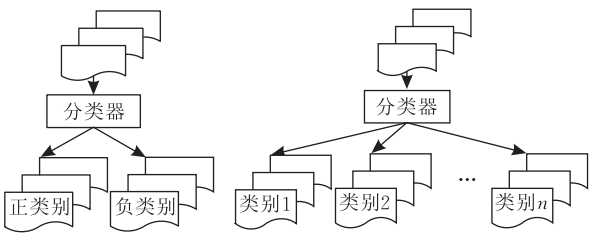
上的得分，然后将这些得分的加权之和作为文档在该类别的最终得分，以此判断文档类别. 还有其它 Big-bang 分类方法，在此不一一详述.

3.1.3 全局分类方法的优缺点

全局分类方法的优点是分类过程简单，因为全局分类将所有类别作为一个整体，为整个类别层次学习提供一个全局的分类模型，每次分类预测能将文档分至最终类别，因此避免了层次式迭代分类过程中的错误传播问题. 错误传播问题是指在自上而下分类方法中，上层节点的错误分类会传播到其后代节点的分类中去. 全局分类方法的缺点是分类学习的时间开销比较大，因为全局方法是在整个数据集上将进行学习和训练分类器，其学习的空间开销和时间开销比采用分而治之策略的方法要大许多. Liu 等人[12]通过实验对基于 SVM 实现的全局分类方法和自上而下分类方法进行了比较，以雅虎目录作为实验对象，实验结果表明，全局分类方法的学习时间高达 13 天，而自上而下分类方法的学习时间为 0.42 小时.

3.2 自上而下的分类方法

自上而下的分类方法(Top-down Approaches, Top-down)是指采用分而治之策略，将一个大规模的全局分类问题按照类别层次自上而下依次分解为一个个小规模的局部分类问题. 根据所采用分类器类型的不同，我们将自上而下的分类方法分为基于二元分类器的自上而下分类方法和基于多元分类器的自上而下分类方法. 二元分类器是指对实例在两个类别之间进行分类判断，如图 6(a)所示；而多元分类器则可以处理两个以上的类别，如图 6(b)所示.



(a) 二元分类器 (b) 多元分类器

图 6 分类器类型

除了这两类自上而下分类方法之外，还有一类自上而下方法试图通过优化类别层次结构来提高分类方法的性能，我们将这类方法称为基于类别层次优化的分类方法. 下面我们从这 3 方面总结已有的自上而下分类方法.

3.2.1 基于二元分类器的自上而下分类方法

基于二元分类器的自上而下分类方法(Binary Classifier based top-down Approaches, BC-top-down)是指为类别层次中除根节点以外的每个节点训练一个二元分类器，对文档进行自上而下的分类. 如图 7 所示，每个虚线框表示一个二元分类器，对于一个文档，自上而下进行分类预测，由每个节点上的本地分类器判断文档是否属于当前类别. 显然，这种分类方法可以将文档分到类别层次树中任一级上的多个节点上面，因此自然地支持部分深度标签分类、全深度标签分类、单标签分类以及多标签分类. 由于每个节点上面都是一个简单的二元分类问题，因此可以灵活地采用各种二元分类算法，常用的有 SVM[12-16]、朴素贝叶斯分类器[17]、决策树分类器[18]、Rocchio 分类器[19]等.

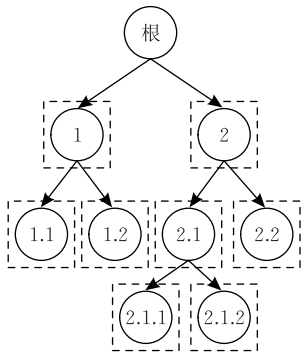


图 7 基于二元分类器的自上而下分类方法

Koller 等人[17]在 1997 年提出了这种基于二元分类器的自上而下层次分类方法，在分类预测过程中，仅当上一层的父节点输出为真时才在下一层的各个子节点进行分类预测，这种预测方法会产生“阻滞”问题. 就是说，如果文档被父节点的分类器拒绝了，则文档无法传给子层的分类器. 谭金波[16]试图解决自上而下层次分类中的“阻滞”问题，他在每个节点训练两个分类器，通过为分类预测建立双通道来减缓“阻滞”. Sun 等人[13]提出了基于类别相似性和基于类别距离两种判别文档和类别之间相关性的方法，并且支持部分深度标签分类和全深度标签分类. Liu 等人[12]以雅虎目录作为实验对象，从数学分析和实验两方面分析了基于 SVM 实现的全局分类方法和基于二元分类器的自上而下分类方法的计算复杂性，其主要工作是证明了在大规模层次分类中层次式 SVM 分类方法的学习时间是可以接受的，而平面式 SVM 分类方法的学习时间则难以接受.

在类别层次树中，相邻类别之间一般主题相关：纵向上，同一条路径上的类别之间具有派生关系；横向上，具有相同父亲节点的类别之间一般主题相近. 对此，一些分类方法试图利用类别之间的相关性帮助分类决策. 另外在大规模层次分类中，许多类别缺

少实例或者实例非常少，我们称之为稀有类别，而稀有类别容易导致分类器学习过分拟合. 通过利用相邻类别辅助分类决策，一定程度上可以解决稀有类别分类当中的过分拟合问题. 许多学者都对这种基于邻居的分类方法(Neighbor Based Approaches)进行了研究，例如 Susan 等人[14]在对一个文档分类的时候，以当前类别、父亲类别、祖父类别 3 个节点上的分类器分别计算该文档的相似性得分，然后以这 3 个得分的加权之和作为该文档的最终得分，以此进行分类.

下面我们对基于邻居类别的分类方法进行形式化描述. 首先定义类别层次树中一个类别的邻居类别的构成，如表 3 所示，邻居类别分为祖先类别、后代类别和兄弟类别 3 类. 对于一个类别 c_p 和一个文档 d，$P(c_p|d)$表示类别 c_p 上的分类器计算出的 d 属于 c_p 的概率或者得分. 基于邻居类别的分类方法在预测一个文档类别的时候，不仅考虑本地类别分类器的预测结果 $P(c_p|d)$，还综合考虑祖先类别的分类器预测结果 $P(Ca^p|d)$、后代类别的分类器预测结果 $P(Cd^p|d)$、兄弟类别上的分类器预测结果 $P(Cs^p|d)$，因此，基于邻居类别的分类方法计算文档 d 属于类别 c_p 的概率得分如式(6)所示. 其中，$P(c_p|d)$是 d 关于 c_p 的得分，$P(Ca^p|d)$是 d 关于 c_p 的祖先类别的得分，$P(Cs^p|d)$是 d 关于 c_p 的兄弟类别的得分，$P(Cd^p|d)$是 d 关于 c_p 的后代类别的得分，β_1、β_2、β_3、β_4 分别是 4 项的权重，$P_{final}(c_p|d)$是文档 d 关于类别 c_p 的最终得分，等于这 4 项得分的加权求和.

$$P_{final}(c_p|d)=\beta_1\times P(c_p|d)+\beta_2\times P(Ca^p|d)+\beta_3\times P(Cs^p|d)+\beta_4\times P(Cd^p|d) \tag{6}$$

表 3　邻居类别

邻居	定义	表示
祖先类别	从根节点到节点 c_p 的类别路径上的所有节点，节点 c_p 除外.	$\{Ca_i^p\|i=1,2,\cdots,k\}$，$i$ 是各个祖先类别 Ca_i^p 到 c_p 的路径长度，k 是根节点到 c_p 的路径长度. Ca_i^p 是节点 c_p 向上追溯 i 层的祖先类别，例如 Ca_1^p 表示 c_p 的直接父节点.
后代类别	以 c_p 为根节点的子树中的所有节点，节点 c_p 除外.	$\{Cd_{m,n}^p\|m=1,2,\cdots,M,\ n=1,2,\cdots,N_m\}$，$Cd_{m,n}^p$ 表示以 c_p 为根节点的子树中第 m 层的第 n 个节点. M 是子树高度，N_m 是第 m 层的节点个数. 例如，$Cd_{1,1}^p$ 表示 c_p 的第一个直接子节点.
兄弟类别	和 c_p 具有相同直接父节点的节点.	$\{Cs_j^p\|j=1,2,\cdots,t\}$，$t$ 是 c_p 兄弟节点个数. 例如，Cs_j^p 表示 c_p 的第 j 个兄弟节点.

$P(Ca^p|d)$是 c_p 的所有祖先类别对 d 的综合得分，其计算方法如式(7)所示. 其中 Ca_i^p 是 c_p 向上追溯 i 层的祖先类别，$P(Ca_i^p|d)$是 Ca_i^p 对 d 的得分，α_i 是 $P(Ca_i^p|d)$的权重系数，距节点 c_p 越近的祖先类别权重系数应该越大，因此有 $\alpha_1\geq\alpha_2\geq\alpha_3\geq\cdots\geq\alpha_k$.

$$P(Ca^p|d)=\frac{1}{k}\times\sum_{i=1}^{k}\alpha_i\times P(Ca_i^p|d) \tag{7}$$

$P(Cs^p|d)$是 c_p 的所有兄弟类别对 d 的平均得分，其计算方法如式(8)所示.

$$P(Cs^p|d)=\frac{1}{t}\times\sum_{j=1}^{t}P(Cs_j^p|d) \tag{8}$$

$P(Cs^p|d)$是 c_p 的所有后代类别对文档 d 的综合得分，其计算方法如式(9)所示，其中 $\frac{1}{N_m}\times\sum_{n=1}^{N_m}P(Cd_{m,n}^p|d)$是 c_p 的第 m 层后代类别对 d 的平均得分，γ_m 是第 m 层后代类别得分的权重系数，距离 c_p 越近的后代类别权重应该越大，因此有 $\gamma_1\geq\gamma_2\geq\gamma_3\geq\cdots\geq\gamma_M$.

$$P(Cd^p|d)=\frac{1}{M}\times\sum_{m=1}^{M}\left(\gamma_m\times\frac{1}{N_m}\times\sum_{n=1}^{N_m}P(Cd_{m,n}^p|d)\right) \tag{9}$$

权重系数 β_1、β_2、β_3、β_4、α_i($i=1,2,\cdots,k$)、γ_m($m=1,2,\cdots,M$)可以通过学习计算，也可以根据经验进行设置.

3.2.2　基于多元分类器的自上而下分类方法

基于多元分类器的自上而下分类方法(Multi-classifier based top-down Approaches，MC-top-down)是指根据类别层次树逐层为具有相同父节点的所有类别建立一个分类模型，即在类别层次树中所有非叶子节点上分别训练一个多类分类器，对文档进行自上而下的分类，如图 8 所示.

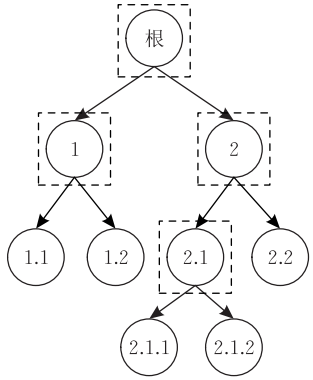


图 8　基于多元分类器的自上而下分类方法

在基于多元分类器方法方面已有一些工作.凌云等人[20]基于隐含语义模型在每个中间节点上建立多元分类器,试图通过在隐含语义空间上进行分类,以解决 LSA 模型中高维矩阵难以进行奇异值分解的问题.袁时金等人[21]分别基于贝叶斯方法、最近邻方法和 Boosting 方法构造多元分类器,并以 BBS 数据作为测试文档对这 3 类方法进行了实验测试,认为 Boosting 方法在学习能力方面表现更为优越.随着类别层次树层次增加,自上而下分类方法会产生比较严重的错误传播问题.对此,Wang 等人[22]采用扁平化策略对类别层次树进行扁平化处理,就是通过去除类别层次树中的一部分中间节点和概念节点降低树的高度,以达到减小错误传播的目的.

基于多元分类器分类方法和基于二元分类器分类方法的一个主要区别就是训练数据集的选择方法不同.多元分类器的训练集由每个类别的正样本集合组成,例如在图 8 中,对于节点 2 上的分类器,需要分别指定类别 2 的子类别 2.1、2.2 的正样本集合.二元分类器的训练集由当前类别的正样本集合和负样本集合组成,例如在图 7 中,对于节点 2 上的二元分类器,需要指定类别 2 的正样本集合和类别 2 的负样本集合.因此,对于基于多元分类器的自上而下分类方法,要在 c_p 建立多元分类器,就要为 c_p 的每个子类别 c_{pi} 指定正样本集合,常用的训练集选择方法有两种.如表 4 所示,第 1 种方法采用排斥策略,仅将 c_{pi} 自身实例作为 c_{pi} 的正样本,第 2 种方法采用包容策略,将 c_{pi} 以及 c_{pi} 所有后代类别的实例均作为 c_{pi} 的正样本.

表 4 MC-top-down 的训练集选择方法

c_p 的训练集选择方法	c_{pi} 的正样本集合
排斥策略[23]	c_{pi} 的实例
包容策略[23]	c_{pi} 的实例以及 c_{pi} 所有后代类别的实例

对于基于二元分类器的自上而下分类方法,要在节点 c_p 建立分类器,就要为 c_p 指定正样本集合和负样本集合,常用的训练集选择方法有两种,具体如表 5 所示.第 1 种方法采用同胞策略,将 c_p 以及 c_p 所有后代类别的实例作为 c_p 的正样本,将 c_p 所有兄弟类别以及这些兄弟类别所有后代类别的实例作为 c_p 的负样本,第 2 种方法采用排斥的同胞策略,仅将 c_p 的实例作为正样本,将 c_p 所有兄弟类别的实例作为负样本.

表 5 BC-top-down 的训练集选择方法

节点 c_p 的训练集选择方法	c_p 的正样本集合	c_p 的负样本集合
同胞策略[24]	c_p 的实例以及 c_p 所有后代类别的实例	c_p 的兄弟类别以及这些兄弟类别所有后代类别的实例
排斥的同胞策略[25]	c_p 的实例	c_p 的兄弟类别的实例

3.2.3 基于类别层次优化的自上而下分类方法

在大规模层次分类中,类别层次是由人工订制的,因此,这种类别层次结构可能并不适合进行自动分类.针对这个问题,一些学者提出了基于类别层次优化的自上而下分类方法(Hierarchy Adaptation based top-down Approaches,HA-top-down),这种方法根据样本数据,通过聚类或者逐步优化的方式修改类别层次结构,获得一棵更加适合自动分类的类别层次树,然后利用新的类别层次树对文档进行自上而下的分类,并且根据原类别层次树和新类别层次树之间的节点映射关系,得到文档在原类别层次树中的类别.

类别层次优化方法分为聚类和逐步优化两种方式,聚类方法[26]是先将类别层次树分解为一组平级类别,然后采用聚类方法将这些类别构建为一棵更加适合自动分类的类别层次树.逐步优化方法[27-28]则是通过提升、降级、合并等基本操作修改类别层次结构,对每次修改后产生的类别层次,进行一次分类学习和测试,评估新产生类别层次的分类性能,再进行下一次的迭代优化,直到获得满意的分类性能.

Lei 等人[27-28]最早提出了主题类别层次优化方法,其目标是对定义好的类别层次,采用数据驱动方式产生一个更加适合自动分类的新类别层次,但这种迭代优化方法收敛比较慢,因此在应用到大规模层次分类中时会产生非常大的计算开销,对此,Nitta 等人[29]提出了一种改进方法,即在算法的迭代过程中,通过限制每个节点的最大子节点数目以降低计算开销.Qi[30-31]采用遗传算法模拟类别层次的迭代进化过程,优化类别层次以提高分类准确率.

重构方法可以获得最佳性能的类别层次树,但完全重构可能导致类别层次树的结构不平衡从而增大训练开销,而逐步优化方法则可以控制类别层次树的结构使其类别分布的更加平衡.

基于类别层次优化的自上而下分类方法的优点是采用数据驱动方式可以产生一个更加适合进行自动分类的类别层次,缺点是类别层次的迭代优化过程一般收敛比较慢,而且对每次迭代产生的类别层

次都需要进行一次分类的学习和测试，以评估新产生类别层次的分类性能，因此，随着类别层次规模的扩大，优化类别层次的时间开销可能会难以控制.

3.2.4　自上而下分类方法的优缺点

自上而下分类方法采用分而治之策略，按照类别层次将一个大规模全局分类问题分解为一个个小规模的局部分类问题，然后进行分类学习和预测. 因此，该方法的主要优点是分类学习和预测的计算开销比较小，最明显的缺点是错误传播问题. 随着类别层次树高度的增加，错误传播问题加剧，深层类别的分类准确率一般会明显下降，因此一些方法试图通过降低类别层次树的高度来减少错误传播带来的分类误差.

3.3　收缩分类方法

收缩分类方法(Reduce Approaches，简称 Reduce)是一种先搜索再分类的方法，即先根据待分类文档在所有类别中搜索与该文档相关的候选类别，然后根据候选类别的样本训练分类器并对待分类文档进行分类. 因此，这种方法又被称为两阶段分类方法，其核心思想是通过减小分类器学习的类别数目以提高分类准确率. 两阶段方法基于这样一个假设：在一棵大规模类别层次树中，给定一个文档，其相关类别数量远少于不相关类别.

Xue 等人[32-33]提出了一种两阶段分类方法，如图 9 所示. 两阶段分类方法分为搜索和分类两个阶段，在搜索阶段采用 KNN 算法进行相关性计算，确定待分类文档的候选类别，然后在分类阶段采用朴素贝叶斯算法根据候选类别训练分类器，最后对待分类文档分类. 在此基础上，Oh 等人[34]也实现了一种两阶段分类方法，他采用全文索引检索工具 Lucene① 计算文档的候选类别，在分类阶段同样采用了朴素贝叶斯算法，他的主要工作是针对大规模层次分类问题中稀有类别缺少训练样本的问题，提出了一个训练集的样本扩展策略，利用类别层次树中的邻居类别增加候选类别的训练样本. 而 Malik[35]则在搜索阶段采用了基于 SVM 的自上而下分类方法计算候选类别，其主要工作是试图通过降低类别层次树的高度，减少自上而下分类方法中的错误传播，提高候选搜索方法的性能. 另外，Rao② 和 Han[36]进一步将两阶段分类方法应用于多标签分类问题.

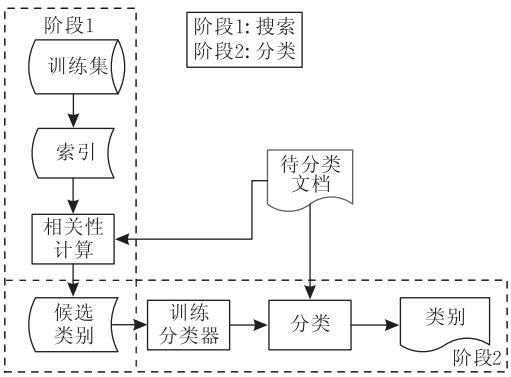


图 9　两阶段分类方法[30]

在两阶段分类方法中，由于第 2 阶段的分类依赖于第 1 阶段候选搜索的准确性，要确保分类正确，就应当使计算出来的候选类别集合包含待分类文档的真实类别. 而已有的两阶段分类方法均未对候选类别搜索方法的有效性进行评估，对此，本文提出评价指标 V_Γ，以此评价一个候选搜索算法 Γ 的有效性，对于待分类文档 d，由算法 Γ 搜索的候选类别集合为 E，对于单标签分类问题，V_Γ 的定义如式(10)所示. 对于多标签分类问题，V_Γ 的定义如式(11)所示，其中 a 是 E 中出现的 d 真实类别数目，l_d 是 d 真实类别总数.

$$V_\Gamma(d)=\begin{cases}1, & \text{如果 } E \text{ 包含 } d \text{ 的真实类别}\\ 0, & \text{否则}\end{cases} \tag{10}$$

$$V_\Gamma(d)=\begin{cases}a/l_d, & \text{如果 } a>0\\ 0, & \text{否则}\end{cases} \tag{11}$$

对于候选搜索算法 Γ，若训练样本集合为 $I=\{\langle d_1,v_1\rangle,\langle d_2,v_2\rangle,\cdots,\langle d_n,v_n\rangle\}$，文档 d_i 的真实类别为 v_i，样本总数为 $|I|$，则 V_Γ 在 I 上的平均值 $\overline{V_\Gamma}$ 的定义如式(12)所示，可以用 $\overline{V_\Gamma}$ 衡量候选搜索算法 Γ 在样本集合 I 上的性能. $\overline{V_\Gamma}$ 值对分类非常关键，$\overline{V_\Gamma}$ 值越大，说明候选搜索算法性能越好，分类越容易进行. 因此，在研究两阶段分类方法时，应该使候选搜索方法的 $\overline{V_\Gamma}$ 取值最大.

$$\overline{V_\Gamma}=\frac{1}{|I|}\sum_{i=1}^{n}V_\Gamma(d_i) \tag{12}$$

基于效率考虑，大规模层次分类中的候选搜索方法一般采用信息检索技术搜索待分类文档 d 的候选类别，利用向量空间模型计算文档 d 和不同类别的相关性得分，对相关类别进行排名，如果将候选集合 E 大小设置为 k，则将前 k 个类别作为候选. 因此，候选搜索问题是指给定一个样本集合 I，并且指定候选集合 E 的大小，求解具有最大 $\overline{V_\Gamma}$ 值的最优解 Γ. 下面证明候选搜索问题是一个 NP-难解问题.

① http://lucene.apache.org

② http://lshtc.iit.demokritos.gr/system/files/Vaijanath.pdf, 2011

定理 1. 候选搜索问题是一个 NP-难解问题.

证明. 为了便于证明,此处用布尔向量表示文档 d 的特征权重向量,即将在文档 d 中出现了的特征权重置为 1,未出现的特征权重置为 0;另外假设每个特征 f 最多和 m 个类别相关,此处将 m 取值为 1;另外设置候选集合 E 的大小 k 为 1,即取得分最高的类别作为候选. 显然,实际的候选搜索问题比加入以上这些限定的候选搜索问题更加困难. 因此,如果该特定候选搜索问题是 NP-难解的,则一般搜索问题必然是 NP-难解的. 而文献[37]证明了可以将集合覆盖问题规约到该特定候选搜索问题,由于集合覆盖问题是一个 NP 完全问题,因此该特定候选搜索问题是一个 NP-难解问题. 所以,候选搜索问题是一个 NP-难解问题. 证毕.

由于候选搜索问题是 NP-难解的,因此对于大规模层次分类问题,求解最优解 Γ 的时间开销将会难以接受. 因此,可以研究采用启发式算法求解候选搜索问题.

总之,两阶段分类方法的优点是通过候选搜索有效减小了数据规模,因此可以灵活地选择分类方法和分类器,分类准确率比较高;缺点是测试分类的计算开销比较大,因为对于每个测试样例,首先要搜索和计算候选类别,然后根据候选类别的所有样本训练分类器,再对文档进行分类.

4 大规模层次分类方法的评价和比较

4.1 评价标准

目前大多数的层次分类方法依旧采用 Flat 分类方法中的评价标准衡量算法的分类性能,经常使用的有准确率(*accuracy*)、正确率(*precision*)、召回率(*recall*)以及 F_1 值(F_1-*score* 或 *F-measure*). 准确率是指分类正确的样本占所有分类测试样本的比例;召回率是指分类正确的样本占所有目标类中样本的比例;正确率是指分类正确的样本占所有被指定为目标类的样本的比例;F_1 是召回率和正确率的调和平均数. 对于类别 c,其测试结果邻接表如表 6 所示,则准确率可表示为 $(TP+TN)/(TP+FN+TN+FP)$,正确率可表示为 $TP/(TP+FP)$,召回率可表示为 $TP/(TP+FN)$,F_1 计算公式如式(13)所示:

$$F_1=\frac{2\times precision\times recall}{precisoin+recall} \tag{13}$$

准确率反映了分类器对整个样本的判定能力;正确率反映了分类器对负样本的判定能力;召回率反映了分类器对正样本的判定能力;F_1 则综合考虑了分类器对正样本和负样本的判定能力. 实验中通常采用宏平均 F_1(Macro-F_1)和微平均 F_1(Micro-F_1)作为衡量算法分类性能的评价标准.

表 6 测试结果邻接表

类别 c		人工判断	
		True	False
分类器判断	Positive	*True positive* (*TP*)	*False positive* (*FP*)
	Negative	*True negative* (*TN*)	*False negative* (*FN*)

针对大规模层次分类方法,一些学者也提出了相应的评价标准,以更好地评价层次分类方法的分类性能,但目前还没有一个统一的评价标准. 在这些评价标准中,树误差[38]、层次式度量[39]、分级度量[13]是应用较为广泛的几个评价标准. 文献[38]提出了树误差这一评价标准来计算树分类的整体误差,如式(14)所示,其中 D 是测试文档总数,c_d 是文档 d 的预测类别,t_d 是文档的真实类别,$Path\text{-}length(c_d,t_d)$ 是 c_d、t_d 相应两个节点在类别层次树中之间的距离. 树误差可以有效评价树分类算法的分类性能,但是不适用于有向无环图分类,因为在有向无环图类别层次中,节点之间的距离并不唯一.

$$Tree_{error}=\frac{\sum_{d=1}^{D}Path\text{-}length(c_d,t_d)}{D} \tag{14}$$

文献[39]提出了一个同时适用于树分类和有向无环图分类的评价标准:层次式度量,包括层次式正确率(*Hierarchical Precision*, *HP*)、层次式召回率(*Hierarchical Recall*, *HR*)、层次式 *F* 度量(*Hierarchical F-measure*, *HF*),*HP*、*HR*、*HF* 的定义如式(15)~(17)所示,其中 $\hat{P}_i$ 是包括文档 i 预测类别以及该类别所有祖先类别在内的集合,$\hat{T}_i$ 是包括文档 i 真实类别以及该类别所有祖先类别在内的集合. 层次式度量适用于所有满足偏序关系的类别层次,因此可以用于树分类和有向无环图分类,而不能用于有向有环图分类.

$$HP=\frac{\sum_i \hat{P}_i\cap\hat{T}_i}{\sum_i|\hat{P}_i|} \tag{15}$$

$$HR=\frac{\sum_i \hat{P}_i\cap\hat{T}_i}{\sum_i|\hat{T}_i|} \tag{16}$$

$$HF=\frac{2\times hP\times hR}{hP+hR} \tag{17}$$

树误差和层次式度量均是从整体上评价算法的分类性能，为了能够分析算法在类别层次中不同层级上的分类性能，Liu 等人[12]提出了一种分级度量方法，分别计算分类算法在类别层次中每一级上的准确率，例如，若分类算法判断测试文档 d 在图 1(a)中的类别为 2.1.2，那么我们在评价第 1 级分类性能的时候只关心 d 是否真属于类别 2，当评价第 2 级分类性能的时候只关心 d 是否真属于类别 2.1，以此类推.分级度量由于能够直观评价算法在不同层级上的性能表现，因此应用较为广泛，文献[12，32-34，40]均采用分级度量评价算法性能.

总之，层次式度量适用于对单标签路径分类、多标签路径分类、全深度标签分类、部分深度标签分类、树形分类、有向无环图分类等大多数层次分类方法进行性能评价，是一个适用性较强的整体评价方法，而分级度量则是一种适用于树形分类的简单直观的评价方法.

4.2　不同分类方法的比较

在衡量一个分类算法优劣的时候，除了分类性能之外，还需要考虑算法的易用性、时间开销、推广能力等因素.下面我们对第 3 节介绍的 6 种分类方法，从算法的易用性、时间开销、分类性能等方面进行一个概括的比较.表 7 列出了这 6 种分类方法各自的特点以及自上而下分类方法的共同特点.

表 7　不同分类方法的特点

方法	优点	缺点
Flat	实现简单.	忽略类别层次；分类学习的时间开销比较大；数据倾斜问题.
Big-bang	在分类的学习和预测中，考虑类别层次关系.	要对原始分类算法进行修改；分类学习的时间开销比较大.
BC-top-down	实现简单；自然支持多标签分类.	数据倾斜问题；需要维护大量的分类器.
MC-top-down	实现简单；相比基于二元分类器的自上而下方法分类器数量较少.	仅支持树形类别层次的分类问题.
HA-top-down	产生一个更加适合进行自动分类的类别层次.	优化类别层次的时间开销大.
Top-down 方法的共同特点	利用类别层次生成训练集；利用类别层次进行迭代分类；灵活，可以采用一般分类算法；分类学习和预测的计算开销比较小.	错误传播问题；“阻滞”问题.
Reduce	灵活，可以采用一般分类算法.	分类预测的时间开销比较大.

对于分类性能，有没有一种分类方法普遍优于其它分类方法？对此，我们基于已有相关工作中的结论，对不同分类方法的分类性能进行了比较，如表 8 所示，符号＞、～、＜分别表示相应行的方法的分类性能优于、相似、差于相应列的方法，行上的方法是相关工作作者提出的新算法，而列上的方法则为基准算法.从表 8 可以看出，相比 Flat 方法，Top-down 方法和 Big-bang 方法应该更有竞争力.HA-top-down 方法通过优化类别层次结构可以进一步提高 Top-down 方法的分类性能，而 Reduce 方法通过减小分类问题的规模，也可以提高 Flat 方法和 Top-down 方法的分类性能.由于大规模层次分类问题

表 8　不同分类方法的性能比较

方法	相关工作	Flat	Big-bang	BC-top-down	MC-top-down	HA-top-down	Reduce
Flat	Guan 等人[2](2009)	＞					
	Zhang 等人[4](2009)	＞					
	Manani 等人[7](2010)		＞				
	Wang 等人[3](2011)		＞				
Big-bang	Wang 等人[9](2001)	＞					
	Cai 等人[10](2004)	＞					
	Miao 等人④(2010)				～		
BC-top-down	Susan 等人[14](2000)	～					
	Liu 等人[12,41](2005)	＞					
	谭金波[19](2007)	＞		＞			
	Vens 等人[18](2008)		＜				
	Wang 等人[22](2010)			＞			
MC-top-down	袁时金等人[21](2004)	＞					
	凌云等人[20](2005)	＞					
HA-top-down	Tang 等人[28](2006)				＞		
	Nitta[29](2010)					～	
	Qi[30-31](2011)	＞			＞		
Reduce	Xue 等人[32](2008)			＞			
	Oh 等人[34](2010)						＞
	Rao 等人⑥(2011)	＞					
	Malik 等人[42](2011)	＞					

目前还没有公认的测试数据集和评价标准,因此表8中的相关工作是在不同的数据集上对算法进行性能测试,在这些工作当中,一些研究者采用的是从ODP目录、雅虎目录采集到的数据集,还有一些研究者采用的是特定领域的数据集,而且所采用的评价标准也不尽相同,因此难以对这些分类方法进行一个比较全面的评价和比较.所以,哪种大规模层次分类方法更好依旧是一个悬而未决的问题.

下面我们分析这6种分类方法对不同大规模层次分类问题的适用性.对于一个大规模层次分类问题 $G=\langle H,P,L,D\rangle$,其中 H 的可能取值有T、DAG和DCG,P 的可能取值有SD和MD,L 的可能取值有SPL和MPL,D 的可能取值有FD和PD.由于多维度分类问题可以分解为单维度分类问题处理,并且所有的分类算法均适用于单维度分类,所以此处不再考虑分类算法对于属性 P 的适用性.表9列出了6种分类方法对T、DAG、DCG、SPL、MPL、FD、PD这7种分类问题的适用性,√表示适用,×表示不适用.Top-down方法由于要求类别层次满足偏序关系,因此3类Top-down方法均不适用于有向有环图分类问题.基于多元分类器的Top-down方法用于有向无环图分类时会产生不一致问题,因为在分类过程中,具有多个父节点的类别可能导致分类结果的不一致,例如在图1(b)中,2.1有两个父节点1和2,在分类过程中,若节点1上的分类器判断文档 d 属于2.1,而节点2上的分类器判断文档 d 不属于2.1,则结果互相矛盾,因此基于多元分类器的Top-down方法不适用于有向无环图分类.Flat方法由于要求所有实例均位于叶子类别,因此不适用于部分深度标签分类,除非将部分深度标签分类问题转化为全深度标签分类问题,否则不能用于部分深度标签分类.

表9 不同分类方法的适用性

方法	T	DAG	DCG	SPL	MPL	FD	PD
Flat	√	√	√	√	√	√	×
Big-bang	√	√	√	√	√	√	√
BC-top-down	√	√	×	√	√	√	√
MC-top-down	√	×	×	√	√	√	√
HA-top-down	√	√	×	√	√	√	√
Reduce	√	√	√	√	√	√	√

5 未来研究方向

大规模层次分类问题的研究尚处于起步阶段,结合实际应用需求和大规模层次分类问题研究现状,我们认为未来的研究方向主要有如下几个方面:

(1)类别层次中稀有类别的分类方法研究

在大规模层次分类中,稀有类别在类别层次中非常普遍.在ODP目录、雅虎目录等主流Web分类目录中,70%左右类别的实例个数不足10个.由于稀有类别的实例非常少,这使得难以发现稀有类别的规律性,降低了分类器的学习效果.对此,现有方法利用邻居类别来增加稀有类别的训练样本,即将稀有类别在类别层次中的邻居类别的实例也作为稀有类别的训练样本.这种方法以稀有类别所在的类别路径或者子树的特征代表稀有类别的特征,而由于稀有类别自身实例相对稀少,这将导致稀有类别淹没到其所在的类别路径或者子树当中,最终导致分类结果发生漂移.因此,稀有类别分类是大规模层次分类中一个亟待解决的问题.

(2)类别层次中深层类别的分类方法研究

在大规模层次分类中,类别层次的深度一般比较深,称之为深层次结构(deep hierarchy).例如,雅虎目录的类别层次深度为16,ODP目录的类别层次深度为12.类别层次的这种结构特点,导致深层类别的分类准确率下降明显.现有方法在处理深层类别类问题的时候,通常采用扁平化策略,通过去除一部分中间节点和概念节点以降低类别层次高度、减少错误传播而提高分类准确率,这类方法通过降低类别层次高度,一定程度上可以减小错误传播.导致深层类别分类性能比较差的原因除了错误传播问题之外,另外还有两个因素:一是深层类别中许多类别是稀有类别;二是随着目录深度的增加.相邻类别之间的相似性增强.现有研究并未将这两个因素考虑在内.因此,未来对深层类别分类问题的研究可以从这两个方面入手.

(3)半监督的类别层次结构调整方法研究

在类别层次的使用过程中,用户往往会根据需要调整类别层次结构,例如增加或删除一些节点.尤其是对于网络资源目录这种类别层次,随着互联网的快速发展,信息分类目录需要不断扩展,在目录结构扩展过程中,需要将已有的实例重新指定到新目录中,这个过程如果由人工完成,则会产生巨大的工作量.因此,在人工调整目录结构之后,需要将目录中已有实例自动指定到新目录中的类别,这就需要一种半监督的类别层次结构调整方法,以保证类别层次的可扩展性.

(4)有向无环图和有向有环图的分类方法研究

目前的大规模层次分类方法研究主要针对树形

类别层次的分类问题，对于有向无环图和有向有环图两种类别层次的分类问题研究很少，而在实际应用中，类别往往会被组织成有向无环图，甚至更复杂的有向有环图. 例如，维基百科分类目录①的类别体系就是一个复杂的有向有环图，而 ODP 目录则是一个有向无环图，因为该目录中的一些类别节点有多个父节点. 所以在面向实际应用时，就需要对有向无环图和有向有环图这两种类型的大规模层次分类问题进行研究.

6　结束语

大规模层次分类是一个新研究热点. 从树形类别层次的分类问题到有向无环图和有向有环图的分类问题、从单一维度的分类问题到多个维度的分类问题、从单标签路径分类问题到多标签路径分类问题、从全深度标签分类问题到部分深度标签分类问题都受到了大家的关注. 可以说，大规模层次分类问题是一个非常活跃的方向. 从整体上讲，目前在大规模层次分类问题方面的研究还不成熟，尚未建立起一套完整的理论体系，而且从理论的完善到算法的具体应用还有很大的差距.

本文回顾了近年来学术界在大规模层次分类领域的主要成果，对大规模层次分类问题的各方面进行了综述，详细介绍了各种典型的求解方法并加以对比，最后对各种求解方法进行了总结并指明了未来研究的方向.

致　谢　在此，我们向对本文的工作给予支持和建议的同行，尤其是国防科技大学计算机学院软件所 613 教研室的老师和同学表示感谢！

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, 2010, 22(1-2): 31-72

[2] Guan Hu, Zhou Jing-Yu, Guo Min-Yi. A class-feature-centroid classifier for text categorization//Proceedings of the 18th international conference on World Wide Web. Madrid, Spain, 2009: 201-210

[3] Wang Xiao-Lin, Zhao Hai, Lu Bao-Liang. Enhance K-Nearest neighbor algorithm for large-scale multi-labeled hierarchical classification//Proceedings of the 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Athens, Greece, 2011: 58-66

[4] Zhang Cong-Le, Xue Gui-Rong, Yong Yu et al. Web-scale classification with Naive Bayes//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 1083-1084

[5] Labrou Y, Finin T W. Yahoo! as an ontology: Using Yahoo! Categories to describe documents//Proceedings of the 8th International Conference on Information and Knowledge Management. Kansas City, USA, 1999: 180-187

[6] Christophe Brouard. ECHO at the LSHTC pascal challenge 2//Proceedings of the 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Athens, Greece, 2011: 49-57

[7] Madani O, Huang Jian. Large-scale many-class prediction via flat techniques//Proceedings of the Large-Scale Hierarchical Classification Workshop in the 32nd European Conference on Information Retrial. Milton Keynes, UK, 2010: 1-6

[8] Crammer K, Dekel O, Keshet J et al. Online passive-aggressive algorithms. Journal of Machine Learning Research, 2006 (7): 551-585

[9] Wang Ke, Zhou Sen-Qiang, He Yu. Hierarchical classification of real life documents//Proceedings of the 1st Society for Industry and Applied Mathematics International Conference on Data Mining. Chicago, USA, 2001: 1-16

[10] Cai Li-Juan, Hofmann T. Hierarchical document categorization with Support Vector Machines//Proceedings of the 13th ACM International Conference on Information and Knowledge Management. Washington, USA, 2004: 78-87

[11] Sasaki M, Kita K. Rule-based text categorization using hierarchical categories//Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics. San Diego, USA, 1998: 2827-2830

[12] Liu Tie-Yan, Yang Yiming, Wan Hao et al. Support vector machines classification with a very large-scale taxonomy. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 36-43

[13] Sun Ai-Xin, Lim E-P. Hierarchical text classification and evaluation//Proceedings of the 2001 IEEE International Conference on Data Mining. California, USA, 2001: 521-528

[14] Susan Dumais, Hao Chen. Hierarchical classification of Web content//Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Athens, Greece, 2000: 256-263

[15] Nicolo Cesa-Bianchi, Claudio Gentile, Luca Zaniboni. Hierarchical classification: Combining bayes with SVM//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 177-184

[16] Marath Sathi T. Large-scale web page classification[Ph. D. dissertation]. Halifax, Nova Scotia: Dalhousie University, 2010

[17] Koller D, Sahami M. Hierarchically classifying documents using very few words//Proceedings of the 14th International Conference on Machine Learning. Nashville, USA, 1997: 170-178

[18] Vens C, Struyf J, Schietgat L et al. Decision trees for hierarchical multi-label classification. Machine Learning, 2008,

① http://zh.wikipedia.org/wiki/Wikipedia:分类索引

73(2): 185-214

[19] Tan Jin-Bo. An improved hierarchical document classification method. New Technology of Library and Information Service, 2007(2): 56-59(in Chinese)
(谭金波. 一种改进的文档层次分类方法. 现代图书情报技术, 2007(2): 56-59)

[20] Ling Yun, Liu Jun, Wang Xun. Multi-hierarchical classification of Web text. Journal of the China Society for Scientific and Technical Information, 2005, 24(6): 684-689(in Chinese)
(凌云, 刘军, 王勋. 多层次 Web 文本分类. 情报学报, 2005, 24(6): 684-689)

[21] Yuan Shi-Jin, Li Rong-Lu, Zhou Shui-Geng et al. Hierarchical Chinese document categorization. Journal of China Institute of Communications, 2004, 25(11): 55-63(in Chinese)
(袁时金, 李荣陆, 周水庚等. 层次化中文文档分类. 通信学报, 2004, 25(11): 55-63)

[22] Wang Xiao-Lin, Lu Bao-Liang. Flatten hierarchies for large-scale hierarchical text categorization//Proceedings of the 5th International Conference on Digital Information Management. Thunder Bay, Canada, 2010: 139-144

[23] Eisner R, Poulin B, Szafron D et al. Improving protein function prediction using the hierarchical structure of the gene ontology//Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. La Jolla, USA, 2005: 1-10

[24] Fagni T, Sebastiani F. On the selection of negative examples for hierarchical text categorization//Proceedings of the 3rd Language Technology Conference. Poznan, Poland, 2007: 24-28

[25] Ceci M, Malerba D. Classifying web documents in a hierarchy of categories: A comprehensive study. Journal of Intelligent Information Systems, 2007, 28(1): 37-78

[26] Li Tao, Zhu Sheng-Huo. Hierarchical document classification using automatically generated hierarchy. Journal of Intelligent Information Systems, 2007, 29(2): 211-230

[27] Tang Lei, Liu Huan, Zhang, Agarwal N et al. Topic taxonomy adaptation for group profiling. ACM Transactions on Knowledge Discovery from Data, 2008, 1(4): 1-26

[28] Tang Lei, Zhang Jian-Ping, Liu Huan. Acclimatizing taxonomic semantics for hierarchical content classification from semantics to data-driven taxonomy//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2006: 384-393

[29] Nitta K. Improving taxonomies for large-scale hierarchical classifiers of web documents//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 1649-1652

[30] Qi Xiao-Guang, Davison Brian D. Optimizing hierarchical classification through tree evolution. Bethlehem, USA: Department of Computer Science and Engineering. Lehigh University, Technical Report: LU-CSE-11-002, 2011

[31] Qi Xiao-Guang. Web page classification and hierarchy adaptation [D]. Halifax, Nova Scotia: Lehigh University, 2012

[32] Xue Gui-Rong, Xing Di-Kan, Yang Qiang et al. Deep classification in large-scale text hierarchies//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 619-626

[33] Xing Di-Kan, Xue Gui-Rong, Yang Qiang et al. Deep classifier: automatically categorizing search results into large-scale hierarchies//Proceedings of the 1st ACM International Conference on Web Search and Data Mining. New York, USA, 2008: 139-148

[34] Oh Heung-Seon, Choi Yoonjung, Myaeng Sung-Hyon. Combining global and local information for enhanced deep classification//Proceedings of the 25th ACM SIGAPP Symposium on Applied Computing. Sierre, Switzerland, 2010: 1760-1767

[35] Malik H. Improving hierarchical SVMs by hierarchy flattening and lazy classification//Proceedings of the Large-Scale Hierarchical Classification Workshop in the 32nd European Conference on Information Retrial. Milton Keynes, UK, 2010: 1-12

[36] Han Xiao-Gang, Liu Jun-Fa, Shen Zhi-Qi et al. An optimized *K*-Nearest Neighbor Algorithm for large scale hierarchical text classification//Proceedings of the 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Athens, Greece, 2011: 2-12

[37] Madani O, Connor M, Greiner W. Learning when concepts abound. Journal of Machine Learning Research, 2009, 10: 2571-2613

[38] Dekel O, Keshet J, Singer Y. Large margin hierarchical classification//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 209-216

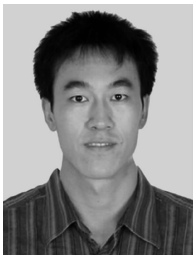
[39] Kiritchenko S, Matwin S. Functional annotation of genes using hierarchical text categorization//Proceedings of the Joint ACL/ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Detroit, USA, 2005: 1-5

[40] He Shi-Zhu, Wang Ming-Wen, Zhou Jun-Jun et al. Research on large-scale text hierarchies combining relevant category information. Journal of Shandong University, 2011, 46(5): 58-62(in Chinese)
(何世柱, 王明文, 周军军等. 结合相关类别信息的大规模文本层次分类研究. 山东大学学报, 2011, 46(5): 58-62)

[41] Liu Tie-Yan, Yang Yi-Ming, Wan Hao et al. An experimental study on large-scale web categorization//Proceedings of the 14th International Conference on World Wide Web. Chiba, Japan, 2005: 1106-1107

[42] Mlik H, Fradkin D, Moerchen F. Singe pass text classification by direct feature weighting. Knowledge and Information Systems, 2011, 28(1): 79-98

[43] Kosmopoulos A, Gaussier E, Paliouras G et al. The ECIR 2010 large scale hierarchical classification workshop. ACM SIGIR Forum, 2010, 44(1): 23-32



HE Li, born in 1984, Ph. D. candidate. His research interests include network and information security, database and data mining.

JIA Yan, born in 1961, professor, Ph. D. supervisor. Her research interests include network and information security, database and data mining, social networks.

HAN Wei-Hong, born in 1973, Ph. D., associate professor. Her research interests include network and information security, database and data mining.

TAN Shuang, born in 1984, Ph. D. candidate. His research interests include network and information security, database and data mining.

CHEN Zhi-Kun, born in 1985, Ph. D. candidate. His research interests include network and information security, database and massive data processing.

Background

With the continuous development of network information technology, computer networks expand rapidly in the world, and the number of Web pages in the Internet showing the exponential growth. Then the challenges that we face is how to organize and deal with these vast amounts of information effectively, how to search, filter and manage these network resources better. Therefore, Web text classification has become an important technology. Web taxonomies (i. e. the Yahoo! Directory and the Open Directory Project) often have a large scale, hierarchical and multi-dimensional taxonomy. However, the actual number of Web documents far exceeds the number that has been manually placed into the taxonomies. This combined with the fast-growing pace of the Web as well as dynamically generated Web-pages argues for the need for the large scale hierarchical classification methods that can automatically place Web pages into Web taxonomy. Thus the problem of classifying a Web document to a large scale hierarchical taxonomy presents a new challenge which becomes a popular topic recently. According to the LSHTC experiment in the Large Scale Hierarchical Classification Workshop of the ECIR 2010[43], the maximum classification accuracy by all the 19 methods attended is 0. 467. Obviously, the performance can not meet the actual demands. Therefore, the large scale hierarchical text classification becomes the current hot spots. Large scale hierarchical classification problem researches how to classify the Web documents into the categories among the class hierarchy, which is surveyed in this paper. Firstly, a definition of large scale hierarchical classification problem is proposed, which is used to describe the problem in abstraction level. Meanwhile, strategies for conquering the problem are also investigated. Secondly, classification of solving methods for this problem is analyzed, and on the basis of the classification, many typical solving methods are introduced and compared. Lastly, future research trends of the solving methods for this problem are reviewed. At the National High Technology Research and Development Program of China (No. 2011AA010702, No. 2010AA012505), there is a need that automatic classification of Web pages in the Internet according to the multi-level, multi-dimensional criteria of network services for the national network security management needs. My work can provide a solid theoretical foundation and methodological guidance for this program to some extent.