中国科学: 信息科学 2018 年 第 48 卷 第 5 期: 487–500
SCIENTIA SINICA Informationis
人工智能优青专刊 · 评述

大规模分类任务的分层学习方法综述

胡清华[1*], 王煜[1], 周玉灿[1], 赵红[1], 钱宇华[2], 梁吉业[2]

1. 天津大学计算机科学与技术学院, 天津 300350
2. 山西大学计算机与信息技术学院, 太原 030006
* 通信作者. E-mail: huqinghua@tju.edu.cn

收稿日期: 2017–12–03; 接受日期: 2018–03–12; 网络出版日期: 2018–05–11

国家自然科学基金 (批准号: 61432011, U1435212, 61732011) 资助项目

摘要 分层分类是一种利用数据类别间层次结构关系进行分类的任务, 可以高效地组织和处理大规模数据. 近些年来, 在这个受到越来越多关注的领域中涌现出许多重要的工作. 本文介绍分层分类的定义, 并按照不同种类的问题解决策略, 对大规模分层分类任务中的几个基本问题的研究进行总结. 首先, 给出层次结构的形式化定义. 其次, 分别阐述如何设计分层评价指标、如何构建层次结构、如何利用层次结构信息进行特征选择、如何利用层次结构信息训练分类器以及如何面向层次结构设计停止机制, 并介绍具有代表性的相关工作. 最后, 对大规模分层分类任务进行总结, 并展望未来可能的研究方向.

关键词 分层学习, 层次结构构建, 分层分类器学习, 分层分类停止机制, 分层特征选择, 分类

1 引言

由于互联网和物联网的快速普及, 数据的产生和收集速度正在急剧提升, 可供分析挖掘的数据样本、特征维度都在爆炸式增长. 与此同时, 人们面临的分类学习任务也变得越来越复杂, 需要学习的类别数量迅速增加, 从原来的若干类, 到几十类, 发展到现在的数万类学习任务. 这些超大规模的极多类学习问题的出现, 给分类建模带来了新挑战 [1].

超大规模学习任务具有以下特点. 首先, 待建模的类别数非常大, 建模更加困难. 类别数量从垃圾邮件判别的二类、手写数字的十类、恒星光谱的数百类 [2], 到 ImageNet 图像数据和网页数据的数万个类别 [3]. 过去, 多分类任务往往转换成多个二分类任务来解决, 当类别数量非常大时, 特征空间中将出现大量的不可分区域, 对分类学习产生无法忽视的影响. 第二, 真实数据的海量类别往往呈现长尾分布特性, 传统方法难以有效处理大量尾部类别的建模问题 [4,5]. 数据中大量的样本分布在少数的几个类别上, 而其他大部分类别上样本占比很少, 这就是俗称的 “二八定律”, 即 80% 的样本分布在 20%

引用格式: 胡清华, 王煜, 周玉灿, 等. 大规模分类任务的分层学习方法综述. 中国科学: 信息科学, 2018, 48: 487–500, doi: 10.1360/N112017-00246
Hu Q H, Wang Y, Zhou Y C, et al. Review on hierarchical learning methods for large-scale classification task (in Chinese). Sci Sin Inform, 2018, 48: 487–500, doi: 10.1360/N112017-00246

的类别上. 大量的拥有极少数样本的尾部稀有类别难以有效建模, 被错分到样本数量多的常见类别中. 然而稀有类别在恒星分类、疾病与故障诊断等学习任务中比常见类别更加重要, 这些样本被错分可能会带来巨大的损失. 第三, 海量类别条件下传统特征选择无法解决维数灾难问题 [6]. 当前, 随着采样精度和频度的提高以及传感器的多样化, 描述分类任务的特征也急剧增加, 导致了维数灾难问题 [7]. 传统特征选择方法可以从高维特征中挑选出能同时区分多个类别的少数特征, 但是当类别数很大时, 所挑选出的判别性特征数目仍然巨大, 学习任务仍然在一个超高维空间中进行.

面对这种超大数量、超多类、超高维的学习问题, 人们通常将这些数据类别按照从抽象到具体的方式组织成一个层次结构进行记忆和检索 [8,9]. 如对于大量动植物类别的分类, 公元前 300 多年亚里士多德首次提出生物分类学概念, 1735 年卡尔 · 冯 · 林奈提出树状界、门、纲、目、科、属、种的物种分类法, 体现出人类利用分层结构处理超大规模学习问题的方式. 2004 和 2006 年, *Science*, *Nature* 等期刊论文表明, 即使在蛋白质和基因尺度上也存在着类似于树和有向无环图的层次结构 [10,11], 因此可以利用数据的层次结构解决大规模分类任务的学习问题. 事实上网页的类别和 ImageNet 这样的大规模分类任务中, 往往存在着层次结构, 从 100 多年前的社会现象分类 [12] 开始到今天的 ImageNet 图像分类 [3], 分层分类已经广泛地应用在众多场景中.

利用层次结构进行大规模分类学习将带来一系列优势. 首先, 将超多类问题分解为多个子类学习任务, 能有效降低建模的难度. 其次, 长尾分布中大量尾部类别的学习, 可通过借用与其在层次结构中相似类别的样本以减小数据不平衡带来的困难. 再次, 借助分层结构将大量尾部类别汇聚成上层样本数较多的大类, 在下层分类的过程中再对这些包含少量样本的类别进行分类, 从而更好地解决类别不均衡问题. 最后, 由于每一层需要区分的类别数大大减少, 对不同类别进行判别所需的特征数目也随之显著下降, 可有效地解决维数灾难问题. 如何针对这类学习任务的特点, 设计有效的学习算法应对新的挑战, 正在引起机器学习 [13,14]、自然语言处理 [15,16] 和计算机视觉领域 [17,18] 学者的关注.

在大规模分层分类任务中, 有几个基本问题需要进行探讨. 第一, 分层分类性能的评价. 传统分类任务评价指标是否同样适用于分层分类任务? 如不适用, 如何设计分层评价指标? 第二, 分层分类的层次构建. 某些学习任务中已存在语义的分层结构, 但这个结构是否适合当前的分类任务, 如不适合如何进行调整? 如果没有层次结构信息, 如何建立适合数据分类任务的层次结构? 第三, 分层分类的特征选择. 给定层次结构之后, 如何利用分层结构信息来设计分层特征选择算法以解决海量类别下的维数灾难问题? 第四, 分层的分类器学习. 如何利用已知分层结构的信息训练分类器改进分类性能? 第五, 停止机制设计. 传统分层分类从上层的根节点类别开始, 逐层细分进行到底层叶节点停止. 但是, 如果信息不充分时, 往下的细分可能会产生误判. 能否设计停止机制将样本停止在中间节点以保证结果的正确性? 基于以上的基本问题, 图 1 中概述了大规模分类任务的分层学习方法研究路线图. 虽然文献 [19] 从分类的角度对分层分类进行了总结, 但是本文将对以上 5 个基本问题全面地进行分析, 并综述现有的相关工作, 展望有待解决的问题.

2 类别的层次结构

层次结构一共分为两种, 一种是树结构, 一种是有向无环图结构. 它们的共同点是都具有 “从属” 关系 [20]. 具体而言, 这种 “从属” 关系可以被归纳为三个特性: 不可逆性、反自反性和传递性 [19]. 定义一个层次结构为一个 $(D, \prec)$ 对, 其中 D 是标签集合, 表示 “从属” 关系. 3 个特性的形式化描述为

(1) 不可逆性: 若 $d_i \prec d_j$, 则对于 $\forall d_i, d_j \in D$, 有 $d_j \nprec d_i$;

(2) 反自反性: 对 $\forall d_i \in D$, 有 $d_i \nprec d_i$;

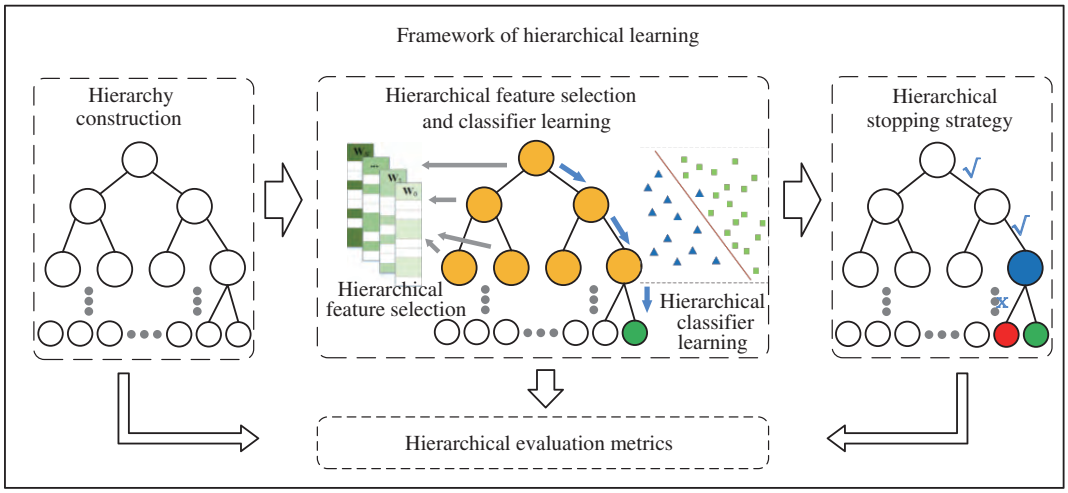

图 1 (网络版彩图) 分层学习研究整体架构图. 图中绿色节点代表样本的真实标签, 红色节点代表错误的预测, 黄色节点代表进行分层特征选择和分类器学习的节点, 蓝色节点代表避免错误预测的正确停止策略结果

Figure 1 (Color online) The framework of hierarchical learning researches. The green node is the ground truth of a sample, the red node is the wrong prediction, the yellow node represents that hierarchical feature selection and classifier learning are done on this node, and the blue node is the right result produced by stopping strategy to avoid misclassification

(3) 传递性: 若 $d_i \prec d_j$ 且 $d_j \prec d_k$, 则对 $\forall d_i, d_j, d_k \in D$, 有 $d_i \prec d_k$.

一般来说, 在一个层次结构中, 有几种类型的节点. 对层次结构中的某节点:

(1) 其父节点用 p_i 表示;

(2) 其孩子节点用 C_i 表示, $|C_i|$ 表示其孩子节点的数量;

(3) 其祖先节点表示该节点的所有从属的节点, 用 $\mathrm{An}(d_i)$ 表示, $|\mathrm{An}(d_i)|$ 表示其祖先节点的数量;

(4) 其对应的叶子节点用 $\mathrm{Le}(d_i)$ 表示, $|\mathrm{Le}(d_i)|$ 表示其对应叶节点的数量. 整个层次结构的叶节点用 L 表示, $|L|$ 表示整个层次结构的叶节点数.

树和有向无环图两种层次结构都具有上述的特性, 其区别在于, 每一个节点在树结构中有且仅有一个父节点 (根节点的父节点可以看作是其本身); 而在有向无环图中, 节点可以拥有多个父节点.

3 分层分类的性能评价

传统的多分类任务评价指标如 F 分数能够度量分类器对于不同类别分类的能力, 但是在分层结构中却不能恰当地描述错分的程度. 图 2 是一个类别层次结构示例, 绿色的矩形表示某样本的真实标签, 红色的三角形分别表示两个分类器的预测标签. 假设该样本的真实标签是小汽车, 如果一个分类器将其预测为公交车, 另一种将其预测为鸟类, 其错误的程度显然是不同的, 而传统的分类评价指标却无法体现出这种特点, 因此需要设计基于层次结构的分层分类评价指标.

应用较为广泛的评价指标主要是基于集合的度量, 其共同点是将真实标签和预测标签扩展为加入各自对应的所有祖先节点的集合. 用 Y 表示样本 X 的真实标签, 用 $\hat{Y}$ 代表 X 的预测标签. 真实标签扩展集和预测标签扩展集分别为

$$Y_{\mathrm{aug}} = Y \cup \mathrm{An}(y_1) \cup \mathrm{An}(y_2) \cup \cdots \cup \mathrm{An}(y_N), \tag{1}$$

其中, $(y_1,\ldots,y_N)$ 是样本 X 对应真实标签的 N 个祖先节点类别.

$$\widehat{Y_{\text{aug}}} = \widehat{Y} \cup \text{An}(\widehat{y_1}) \cup \text{An}(\widehat{y_2}) \cup \cdots \cup \text{An}(\widehat{y_M}), \tag{2}$$

其中, $(y_1,\ldots,y_M)$ 是样本 X 对应预测标签的 M 个祖先节点类别.

树诱导损失是度量样本预测类别与真实类别在树结构中的距离 [21]. 它用从预测节点到真实节点需要走的步数来度量. 由于树诱导损失能够反映样本在树结构上的错误程度, 因此应用非常广泛. 其形式化表示如下:

$$\mathit{\Gamma}_{\text{TIE}}(Y,\widehat{Y}) = \sum_{\text{E}}(Y,\widehat{Y}), \tag{3}$$

其中, $\sum_{\text{E}}(a,b)$ 表示连接树结构中 a 和 b 节点的总边数. 但是, 以树诱导损失作为目标函数无法通过凸优化的方法进行求解 [22].

分层对称差损失由传统多标记分类问题中的度量指标发展而来, 通过计算假阳性样本和假阴性样本度量了预测扩展集和真实扩展集节点数目的差异 [23]. 其形式化描述如下:

$$l_{\mathit{\Delta}}(Y_{\text{aug}},\widehat{Y_{\text{aug}}}) = |(Y_{\text{aug}}\backslash\widehat{Y_{\text{aug}}}) \cup (\widehat{Y_{\text{aug}}}\backslash Y_{\text{aug}})|. \tag{4}$$

另一种由传统度量指标发展的分层评价指标是分层准确率和召回率. 它将传统的准确率和召回率扩展到了分层分类任务中 [23]. 分层准确率表示为预测扩展集与真实扩展集中共同的类别数与预测扩展集的比值, 表征预测为正的类别中正确的比例; 分层召回率表示为预测扩展集与真实扩展集中共同的类别数与真实扩展集的比值, 表征应当被预测为正的类别中有多少被预测出来. 形式化表示为

$$P_H = \frac{|\widehat{Y_{\text{aug}}} \cap Y_{\text{aug}}|}{|Y_{\text{aug}}|}, \quad R_H = \frac{|Y_{\text{aug}} \cap \widehat{Y_{\text{aug}}}|}{|Y_{\text{aug}}|}. \tag{5}$$

分层准确率和召回率是应用较为广泛的一种评价指标, 因为其能够反映出层次结构中类别的从属关系, 从而度量出错误发生的程度. 例如, 上层发生的错误要比下层发生错误更加严重. 但是, 在某些深度较高的层次结构中, 由于分层准确率和召回率会将从根节点起的所有节点都考虑进来, 因此不能明显地反映出下层不同错误之间的差异. 另外, 对有向无环图中有多个父亲的节点, 它会将多条路径上的父节点都加入扩展集中, 无法有效度量有这种分类任务的性能. 因此, 2015 年 Gaussier 等 [23] 提出了最小公共祖先的评价指标. 该方法只关注以预测节点和真实节点的最小公共祖先为根节点的子结构, 因此对较深树结构中的下层节点错误的表达能力更强, 能同时处理树和有向无环图两种结构. 但是, 它在多标记分类任务上会忽视兄弟节点错误, 即同一个父节点下的多个子节点错误只会被计算一次.

4 层次结构的构建

在分层分类中, 许多学者利用 ontology 知识进行分类学习, 如图像分类的 WordNet 语义结构 [24, 25] 和网页分类的分层结构 [26, 27] 等. 但这些基于知识库定义的结构与数据的特征空间可能会存在不一致性, 严重影响分类任务的效果. 如图 3 所示, 人类和鲸鱼都是哺乳动物而鲨鱼不是, 因此从语义结构上人类和鲸鱼之间更为相似; 但从视觉信息上说, 鲨鱼和鲸鱼的相似度显然更高, 它们在分类任务中分到同一大类下更加合理. 以上分类任务给定训练数据时也给出了类的分层结构, 但更多的情况是数据中不包含层次结构, 这就需要基于数据自动构建合理的层次结构. 因此在分类任务中, 对于已知层次

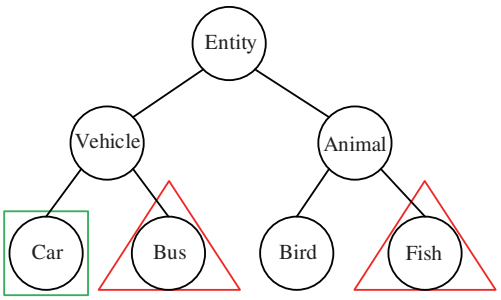

图 2 (网络版彩图) 分层结构上的错分. 假设绿色矩形表示某样本的真实类别, 红色三角形分别代表两个分类器的预测结果. 将轿车预测为鱼类比预测为公交的错误更严重

Figure 2 (Color online) Misclassification on the hierarchy. Assume green rectangle represents the ground truth of a certain sample, and red triangle represents the predictions of the two classifiers. It is more severe for misclassifying a car to a fish than a bus

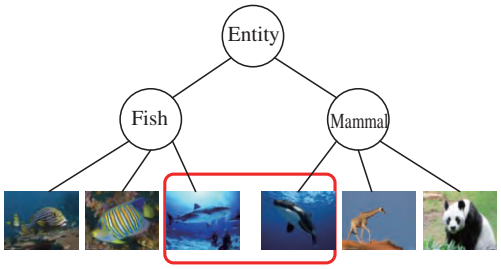

图 3 (网络版彩图) 语义鸿沟. 鲨鱼和鲸在语义结构上分别从属于鱼类和哺乳动物两个大类目下, 但它们在视觉特征上非常相似, 更应该从属于同一个类目下, 即语义空间与特征空间不一致的问题

Figure 3 (Color online) The semantic gap. Sharks and whales belong to the category Fish and Mammal, but it is more reasonable to classify them into the same superclass. This is the inconsistence between the semantic space and the feature space

结构的数据, 需判断其是否与数据特征空间一致, 如不一致则需根据数据对其进行调优; 对于未知层次结构的数据, 需基于数据自动构建适合分类任务的层次结构.

层次结构构建相关的工作大致可以分为 3 类. 第 1 类工作将层次构建的问题看作一个递归的分类问题, 学习 N 个一对多的分类器, 再建立一个混淆矩阵, 将不容易分开的类别归到同一个大类中. 2008 年, Griffin 等 [28] 建立了一个二叉树层次结构以提高性能. 2010 年, Bengio 等 [29] 对多分类问题设计了标签嵌入的树结构. 2013 年 Liu 等 [30] 对于大规模数据分类建立了概率标签树结构. 与这些基于树结构假设的工作不同, Deng 等在 2011 年提出 [31] 建立一个有向无环图结构能比树结构对不易区分的类别获得更好的分类性能. 这一类工作的主要不足在于, 层次结构的好坏依赖于分类器的性能, 而且在大规模数据条件下训练 N 个多分类器的时间复杂度高, 很难进行实际应用. 另外, 当数据分布不均衡甚至呈现长尾分布特性时, 由基于均匀分布假设的传统分类器得到的混淆矩阵是不可靠的.

第 2 类工作将层次结构的建立看作是相似性类别的划分问题, 首先设计并学习类别之间的相似性度量, 再利用聚类建立起层次结构. 2013 年, Zhou 等 [32] 利用了近邻传播进行分层结构的聚类, 2014 年, Lei 等 [33] 采用了谱聚类构建层次结构. 2015 年, Fan 等 [34] 利用类样本的均值作为代表点度量类别之间的相似关系来建立层次结构, 虽然实现了快速的层次构建方法, 但是却丢失了较多的信息. 因此在 2016 年, Qu 等 [17] 针对大规模数据提出了快速高效的层次结构建立方法, 通过类样本的均值和方差巧妙地设计了类别相似度矩阵, 再通过分层谱聚类的方法实现. 与传统的欧氏距离不同, Zheng 等 [35] 在 2017 年提出利用了 Hausdorff 距离以获得更加精确的相似度矩阵, 并通过主动采样降低算法的时间复杂度. 虽然运用主动采样的方式减少了整体算法的计算复杂度, 但产生了如何能够使得采样表征整体数据样本的新问题. 这类方法的性能高度依赖于所选择的特征空间的表达能力及相似性度量的设计, 而现有的大多数方法是启发性算法, 缺乏严格的理论基础.

通过利用已有知识和其他策略建立层次结构, 我们可以获得从不同角度描述类别关系的结构, 因此第 3 类工作通过考虑这些不同维度的层次结构, 综合性地刻画类别之间的结构关系. 2012 年, Hwang 等 [36] 提出对于多种标记结构分别建立一个核函数树, 再学习针对标记的多核融合方法, 从多个核函数树中逐层选择对识别某个特定的类别相关的结构关系. 2013 年, Wang 等 [37] 从集成方法的角度提

出在每个标记结构上均构造一个由基分类构成的分类树, 测试样本的预测结果由多个分类树的计算结果集成得到. 2016 年, Zhao 等 [38, 39] 借鉴特征空间上的多核融合方法直接对多个标记结构定义的核矩阵进行优化融合. 这两种策略的不足在于: 基于多核融合的方法融合后的标记结构可解释性差, 用特征空间上的融合方式去处理标记空间上的多重结构的有效性缺乏理论保证, 并且依赖于标记结构上距离定义; 基于集成策略的方法在构造基分类器时只考虑了单个结构, 限制了基分类器的性能.

另外, Deng 等 [40] 和 Jernite 等 [41] 将层次结构与分层分类器联合学习, 能够对于特定的任务实现快速有效的分层学习. 其中, 文献 [40] 优化了精确度和效率的权衡, 而文献 [41] 则最大化了每个节点上分类的纯度和均衡性. 2018 年, Zhou 等 [42] 为了解决长尾分布问题提出了一个端到端的学习模型. 他们将组稀疏的策略融合到卷积神经网络中, 把视觉相似度高的类别划分到同一组中, 从而实现了对不同类别更加精确地判别.

5 分层分类的特征选择

传统的特征选择 [43, 44] 能够获得分类任务中特征较为紧致的表示 [45]. 但是在大规模数据中, 为了能够同时区分海量类别, 所挑选出的特征数目仍然巨大, 无法解决大数据下维数爆炸的问题. 如果借助分层结构将复杂的问题分而治之, 在每一层对于当前较少类别在低维子空间上挑选出具有判别性的特征, 可以有效地解决该问题.

传统的特征选择算法假定所有的类别是相互独立的, 而且对于所有的类别都选出一组相同的特征子集. 不过, Freeman 等 [46] 发现某些特征对一些类别具有较好的判别性, 但是对于其他类别则起不到相应的作用. 基于类似思想的启发, 一些分层特征选择的模型相继被提出. 2011 年, Freeman 等 [47] 学者提出利用遗传算法对于分层分类器和特征选择进行联合优化. 随后 2013 年他们又在分层分类中的不同分类任务中独立地选取不同的特征子集 [46]. 2012 年, Grimaudo 等 [48] 利用 Peng 等 [49] 于 2005 年提出的 mRMR 特征选择算法, 在每一层上独立地选取不同的特征子集. 2015 年, 则是对于分层文本的特定分类任务提出了特征选择算法. 他们对于结构中的不同节点选择不同的特征, 但是没有将层次结构中的关系信息嵌入到特征选择任务中. 这些工作的共同点是对于不同的节点独立的选择不同的特征子集, 但是他们都没有考虑到不同类别之间的依赖关系.

在一个层次结构中, 父子关系和兄弟关系是两种最主要的依赖关系. 其中, 具有父子关系的类别彼此之间较为相似, 会共享某些特征; 而互为兄弟关系的节点需要不同的特征来进行区分. 基于以上的问题和发现, 2017 年, Zhao 等 [6] 研究者基于递归正则化设计了一种分层特征选择算法对不同的子分类任务选择出不同的特征子集. 他们通过 Hilbert-Schmidt 独立性准则度量兄弟类别间的相关性, 对兄弟节点上筛选出的特征之间的依赖性进行惩罚, 使得最终筛选出的特征子集在具有父子关系的类别上较为相似, 在互为兄弟关系的类别上较为不同. 表 1 归纳了不同策略特征选择方法在分层分类应用中的特点和不足.

6 分层分类器学习

给定层次结构之后, 分类器的学习可以通过引入层次结构信息提高性能. 通常来说, 叶子节点对应着所有的样本标签, 因此常用的方法是将样本从根节点开始向下找到对应正确的下层节点, 每一个节点基于贪心准则进行划分, 即在每一层都选择概率最大的分支, 逐步向下细化, 直至叶子节点, 如 Griffin 等 [28]、Feng 等 [50]、Wen 等 [51]、Bengio 等 [29]、Gao 等 [18] 和 Lei 等 [33] 学者的工作. 这种自顶

表 1 不同类型特征选择方法比较

Table 1 Comparison between different kinds of feature selection methods

	特点	不足
传统特征选择	用一组特征子集区别所有类别.	无法解决海量类别数据中的维数灾难问题; 假设类别之间相互独立.
独立分层特征选择	利用分治的思想在不同结点上分别选择特征子集, 有效降低建模难度. 由于每一层区分更少的类别, 使得子分类任务所需的特征数量大大减少.	未考虑类别间的关系, 准确率偏低.
联合分层特征选择	充分考虑类别之间的关系, 将多分类问题划分成若干个小规模分类问题, 并利用类别间的关系进一步优化每个结点所选特征, 对不同类别进行判别所需的特征数目也随之显著下降, 有效地解决了维数灾难问题.	难以协调不同层次以及不同关系之间的权重.

向下的贪心算法存在一个重要的错误传播问题, 上层的分类错误会传递给下层节点, 并且这种错误不可修复.

因此一些学者利用层次结构信息以减小错误传播问题带来的影响. 2007 年, Decoro 等 [52] 对多个独立的二分类器结构进行 Bayes 聚合, 并将其应用到音乐题材分类上. 但是这些策略只是启发式的方法, 并没有完全利用结构信息对分类器进行更好地训练. 2009 年, Bennett 等 [53] 提出了一种级联策略, 将下层分类器的粗略输出作为上层分类器的附加特征以提高上层分类的准确率.

自顶而下的分类过程也可以看作是寻找最优路径的问题, 利用经典的机器学习分类器 (如支持向量机), 结合不同的路径选择策略进行分类. 2011 年, Gao 等 [18] 提出了一种剪枝策略, 即在每一层将那些不可能到达正确叶子节点的标记剔除掉, 以减小搜索空间. 2013 年, Sun 等 [54] 提出了一种类似于分支定界方法的最优路径计算方法. 在该工作基础上, 2016 年 Qu 等 [17] 则采用了近似的动态规划算法进行了改进. 这两个工作的共同点在于在每一层保留置信度最高的 N 条路径 ($N \geqslant 1$), 最终在叶子节点上挑选出一条最优路径. 寻找最优路径的策略通过每一层保留多个可能的路径, 使得错误传播问题得到改善. 但是与上一类策略类似, 它们也没有充分利用层次结构蕴含的结构信息来帮助分类器更好地改进错误传播问题.

为了解决上述问题, 一些学者通过最小化全局损失函数的方式对层次结构中的多个分类器进行联合优化 [55~59]. 1998 年, Mccallum 等 [60] 则是将分层收缩引入到朴素 Bayes 分类器中. 2005 年, Shahbaba 等 [61] 使用了多项式逻辑回归分类器. Wu 等 [62] 在 2016 年的工作和 Fan 等 [63] 在 2017 年的工作类似, 将树状结构中的每一层分类需求看作一个任务, 提出了基于多任务框架的深度学习模型对分层的标记结构进行建模. 另外, 层次结构中蕴含的丰富信息也可以用来提升分类器的性能. 2011 年, Zhou 等 [64] 引入了父子节点上的正交约束, 2015 年, Xie 等 [65] 在学习过程中加入了上层分类结果对下层结果的具有指导作用这一先验知识. Yang 等 [25] 在 2015 年的工作中通过增强层次结构中父子节点参数的相似性, 协同学习层次结构中的所有分类器. 他们针对树形结构, 利用 Bayes 理论提出了分层 Bayes logistic 回归模型; 对于图结构, 他们在经验风险最小化框架中加入了相邻节点参数相似的正则化先验, 提出了递归正则化模型. 另外, Zheng 等 [35] 在 2017 年的工作中认为错误传播问题可以通过学习到更具有判别性的度量来解决, 他们利用层次结构信息, 将父节点独有的度量信息加传递给子节点, 并将这种度量与同一父节点下兄弟节点之间的度量利用多任务学习框架进行学习. 在上述策

略中, 类别间层次结构信息的引入使得分类器的性能有所提高, 但是解决错误传播的问题还有继续改进的空间.

另外, 其他的分类任务如细粒度分类, 与分层分类也有着密切的联系. 细粒度分类目的是更好地区分下属层次的相似度较高的细分种类, 这些种类之间的差别非常细微, 例如北极燕鸥与里海燕鸥, 只在翅膀和脚的颜色等局部有差异. 大量的分类算法利用了人工标注信息来帮助区分这些细微的差别. 早期的工作主要关注学习更具有判别性的特征描述器 [66, 67], 而后来借助人工标注的局部区域和属性, 其判别能力大幅增强. 对于受到广泛关注的图像细粒度分类任务, 学者罗建豪和吴建鑫进行了归纳并综述了重要的学习策略 [68]. 在分层分类中, 经过高层的粗粒度类别的区分之后, 低层次的分类子任务中也常常需要对相似度较高的类别进行区分, 因此在分层分类中, 细粒度分类常常可以在低层节点甚至叶节点层次上的分类任务中进行应用.

7 停止机制设计

在大规模分层分类任务中, 一个样本通常从根节点开始, 经过每一层分类器的划分, 最终到达一个叶节点类别作为该样本的预测标签. 但是, 分层分类中的错误传播问题会严重影响分类的精度: 样本在上层分类的错误结果会传递到下层, 导致该样本发生错分. 尤其当信息不充分或者不确定性较强的时候, 错分很难避免. 如果损失一些信息量, 把样本停止在粒度稍粗的中间节点类别上, 可以提高分类的准确性, 减少甚至避免因为错分而产生的巨大风险. 例如, 一个病人腹部不适, 他有可能患有肠胃疾病, 但是也存在患有心脏疾病的可能. 如果医生鲁莽地诊断为肠胃疾病, 那么病人就会承担误诊的风险. 因此需要设计一个恰当的停止机制将样本停止在中间节点类别上以避免错分.

2001 年, Sun 等 [69] 通过对每个节点设置阈值首次引入分层分类的停止机制, 当样本的后验概率或置信度大于该阈值时继续向下层分类, 否则停止在当前节点上. 停止机制设计的工作可分为 4 类.

第 1 类工作通过保守预测来尽可能减少错分. 2008 和 2010 年的选择性预测模型 [70~72] 是一种当不确定性较强时拒绝进一步预测的方法, 2012 年 Deng 等在文献 [24] 中作为对比实验将其扩展到分层分类中. 这种方法会设置一个全局的阈值, 当分类器给出的最大后验概率或置信度不小于该阈值时, 会将样本分到该类别, 否则将样本分到根节点. 这样的策略虽然经常可以进行 “正确” 的预测, 但是过于保守的预测 (如预测到根节点) 丢失了过多的信息, 因此对决策的帮助有限.

为了克服第 1 类工作的缺陷, 第 2 类工作鼓励样本在层次结构中尽可能向下细分. 2000 年 D'Alessio 等 [73] 和 2004 年 Sun 等 [74] 降低了上层节点的阈值使得样本更容易进入到下层分类器中, 并通过优化 $F1$ 分数进行参数调节. 他们之间的不同在于, 文献 [74] 在每一层设置一个阈值, 优化的是宏 $F1$ (Macro $F1$) 分数; 而文献 [73] 给每一个节点都设定一个阈值, 优化的是真阳数 (true positives) 减去假阳数 (false positives). 尽管他们能够在预测中提供更多的信息, 但是当信息包含的不确定性强时容易发生错分.

第 3 类工作结合了以上两类工作的优点, 寻求细分程度与准确度的平衡. 2007 年, Ceci 等 [75] 在传统的阈值法停止机制中引入了度量预测类别和真实类别在树结构上距离的树诱导损失, 通过优化这种分层分类损失来确定阈值参数值. 2012 年, Deng 等 [24] 用信息增益来度量细分程度, 优化了分层准确率和信息增益的权衡. 他们在保证一个任意给定的分层准确率的基础上, 最大化了预测的信息增益, 从而实现了细分程度与准确度的平衡. 这一类工作解决了前两类工作的固有缺陷, 更适合应用在大规模分层分类任务中. 但是, 同前两类工作一样, 他们都忽略了在下层节点信息不充分和不确定性较强情况下的分类问题, 而这种情况往往容易发生错分.

表 2 不同停止机制策略比较

Table 2 Comparison between different stopping strategies

	特点	不足
保守策略	将样本更多停留在上层粗粒度节点上, 分层准确率较高.	为决策所提供的信息量较小.
精细化策略	将样本尽量划分到下层细粒度类别上, 分类结果更精确.	信息不充分和不确定性较强条件下误分率较高.
折中策略	根据不同场景的需求提供多样化的结果, 保证一定准确率的基础上进行更细粒度的分类.	忽略下层信息对当前决策的影响.
风险最小化策略	根据不同场景的需求, 考虑下层信息对当前决策影响, 提供风险最小的预测结果.	时间复杂度较高, 有时无法得到全局最优解.

针对这种情况, 第 4 类工作更多地考虑了不确定性和风险的问题. 2017 年, Wang 等 [76] 提出了分层分类中的保守风险和冒进风险. 如果为了追求正确而进行过于保守的预测, 会丢失掉大量的有益信息; 如果在信息不充分或不确定性较强的情况下还一味追求细分, 则有可能会因为错分而产生误分类风险. 他们提出了一个局部 Bayes 风险最小化的框架, 将预测过程分解为循环地在每一个节点比较这两种风险来确定停止还是继续向下传递的子过程. 与传统的 Bayes 风险框架中设定的全局损失函数不同, 他们分别利用信息熵和上层到下层损失的信息增益这样的不确定性指标来度量两种风险. 另外, 加权树诱导损失的指标在文中被提出用来度量对两种风险不同的重视程度. 该工作的不足在于由于树诱导损失和加权树诱导损失的非凸特性, 使用遗传算法作为优化方法, 因此得到的全局最优解比较耗时, 且在某些情况下只能得到局部最优解. 表 2 归纳了不同类策略停止机制的特点和不足.

8 总结与展望

大规模分类任务的分层学习发掘和利用了数据类别之间的层次结构关系, 能够高效地处理包含海量类别的大规模数据学习任务. 本文介绍了分层学习中的 5 个基本问题: 如何设计分层分类的评价指标、如何构建层次结构、如何利用层次结构信息进行分层特征选择、如何利用层次结构信息训练分类器以及如何设计分层分类的停止机制, 并论述了各问题中具有代表性的相关工作. 分层学习已经有了广泛的应用, 但是还有大量的问题有待解决.

首先, 多场景分层评价指标. 现有的分层评价指标多针对于单标记的强制叶节点停止问题 [22, 23], 极少考虑真实类别标签处于中间节点以及多标记分类任务的情况. 设计可应用于多种场景的分类指标是重要的研究方向之一.

其次, 已有层次结构调整. 对于给定的层次结构, 如何度量其与特征空间的不一致性以进行结构调整是一个重要的问题, 但是目前只有少数工作关注这个方向 [77~79], 其中大多是启发式的方法 [78, 79], 无法给出其有效性的理论证明和有效适用的场景. 利用数学理论对不同分类任务进行结构调整, 对分层分类的错误传播问题和实际应用有着极为重要的意义.

再次, 分层特征选择开销最小化. 少数分层特征选择的工作已经开始关注如何层次结构信息 [6], 但在海量类别下如何保证各节点选择不同的特征在整体上的总开销最小化是一个现实的问题. 设计出总体选择特征数少且同时具有强判别性的分层特征选择算法, 以更有效地解决海量类别下维数灾难问题, 是现实应用中一项不可忽视的关键技术.

最后, 停止机制高效求解策略设计. 现有的停止机制设计大多利用随机优化方法 [76] 或全局搜索

策略 [73, 75] 进行求解, 这些优化方法较为耗时, 在资源受限的情况下很难得到全局最优解. 如何设计出高效的求解策略解决资源受限下的停止机制算法变得越来越重要.

参考文献

1 Yen I E H, Huang X, Zhong K, et al. Pd-sparse: a primal and dual sparse approach to extreme multiclass and multilabel classification. In: Proceedings of the International Conference on Machine Learning, New York, 2016. 3069–3077

2 Hippel T V, Storrielombardi L J, Storrielombardi M C, et al. Automated classification of stellar spectra-I. Initial results with artificial neural networks. Mon Not Royal Astron Soc, 1994, 269: 97

3 Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vision, 2014, 115: 211–252

4 Powers D M W. Applications and explanations of zipf's law. In: Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, Sydney, 1998. 151–160

5 Farid M, Ilyas I F, Whang S E, et al. Lonlies: estimating property values for long tail entities. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, 2016. 1125–1128

6 Zhao H, Zhu P F, Wang P, et al. Hierarchical feature selection with recursive regularization. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, 2017. 3483–3489

7 Zhai Y, Ong Y S, Tsang I W. The emerging "big dimensionality". IEEE Comput Intell Mag, 2014, 9: 14–26

8 Collins A M, Quillian M R. Retrieval time from semantic memory. American Assoc Artif Intell, 1995, 8: 240–247

9 Pons F, Harris P L, Rosnay M D. Emotion comprehension between 3 and 11 years: developmental periods and hierarchical organization. Eur J Dev Psychol, 2004, 1: 127–152

10 Friedman N. Inferring cellular networks using probabilistic graphical models. Science, 2004, 303: 799–805

11 Bjorklund M, Taipale M, Varjosalo M, et al. Identification of pathways regulating cell size and cell-cycle progression by rnai. Nature, 2006, 439: 1009–1013

12 Hayes E C. The classification of social phenomena. American J Sociol, 1911, 17: 375–399

13 Li J Y, Fong S, Zhuang Y, et al. Hierarchical classification in text mining for sentiment analysis. In: Proceedings of the International Conference on Soft Computing and Machine Intelligence, New Delhi, 2015. 46–51

14 Jernite Y, Choromanska A, Sontag D. Simultaneous learning of trees and representations for extreme classification and density estimation. In: Proceedings of the International Conference on Machine Learning, Sydney, 2017

15 Yin B, Ambikairajah E, Chen F. Hierarchical language identification based on automatic language clustering. In: Procedings of the 8th Annual Conference of the International Speech Communication Association, Antwerp, 2007. 178–181

16 Oh H S, Myaeng S H. Utilizing global and path information with language modelling for hierarchical text classification. J Inf Sci, 2014, 40: 127–145

17 Qu Y Y, Li L, Shen F M, et al. Joint hierarchical category structure learning and large-scale image classification. IEEE Trans Image Process, 2016, 26: 4331–4346

18 Gao T, Koller D. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In: Proceedings of the International Conference on Computer Vision, Barcelona, 2011. 2072–2079

19 Silla C N, Freitas A A. A survey of hierarchical classification across different application domains. Data Mining Knowl Discov, 2011, 22: 31–72

20 Wu F H, Zhang J, Honavar V. Learning classifiers using hierarchically structured class taxonomies. In: Proceedings of the International Conference on Abstraction, Reformulation and Approximation. Berlin: Springer, 2005. 313–320

21 Esposito F, Malerba D, Tamma V, et al. Classical resemblance measures. In: Analysis of Symbolic Data. Berlin: Springer, 2000, 12: 139–152

22 Dekel O, Keshet J, Singer Y. Large margin hierarchical classification. In: Proceedings of the International Conference on Machine Learning, Banff, 2004. 27

23 Gaussier E, Paliouras G, Androutsopoulos I. Evaluation measures for hierarchical classification: a unified view and

novel approaches. Data Mining Knowl Discov, 2015, 29: 820–865
24 Deng J, Krause J, Berg A C, et al. Hedging your bets: optimizing accuracy-specificity trade-offs in large scale visual recognition. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Providence, 2012. 3450–3457
25 Ferrari V, Guillaumin M. Large-scale knowledge transfer for object localization in imagenet. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 3202–3209
26 Gopal S, Yang Y. Hierarchical bayesian inference and recursive regularization for large-scale classification. Acm Trans Knowl Discov Data, 2015, 9: 1–23
27 Deri L, Martinelli M, Sartiano D, et al. Large scale web-content classification. In: Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, 2016. 545–554
28 Griffin G, Perona P. Learning and using taxonomies for fast visual categorization. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Anchorage, 2008. 1–8
29 Bengio S, Weston J, Grangier D. Label embedding trees for large multi-class tasks. In: Proceedings of the International Conference on Neural Information Processing Systems, Vancouver, 2010. 163–171
30 Liu B, Sadeghi F, Tappen M, et al. Probabilistic label trees for efficient large scale image classification. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Portland, 2013. 843–850
31 Deng J, Satheesh S, Berg A C, et al. Fast and balanced: efficient label tree learning for large scale object recognition. In: Proceedings of the International Conference on Neural Information Processing Systems, Granada, 2011. 567–575
32 Zhou N, Fan J P. Jointly learning visually correlated dictionaries for large-scale visual recognition applications. IEEE Trans Pattern Anal Mach Intell, 2013, 36: 715–730
33 Lei H, Mei K Z, Zheng N N, et al. Learning group-based dictionaries for discriminative image representation. Pattern Recogn, 2014, 47: 899–913
34 Fan J P, Zhou N, Peng J Y, et al. Hierarchical learning of tree classifiers for large-scale plant species identification. IEEE Trans Image Process, 2015, 24: 4172–4184
35 Zheng Y, Fan J P, Zhang J, et al. Hierarchical learning of multi-task sparse metrics for large-scale image classification. Pattern Recogn, 2017, 67: 97–109
36 Hwang S J, Grauman K, Fei S. Semantic kernel forests from multiple taxonomies. In: Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, 2012. 1718–1726
37 Wang Y, Forsyth D. Large multi-class image categorization with ensembles of label trees. Int J Mol Medicine, 2013, 31: 1–6
38 Zhao S, Han Y H, Zou Q, et al. Hierarchical support vector machine based structural classification with fused hierarchies. Neurocomputing, 2016, 214: 86–92
39 Zhao S, Zou Q. Fusing multiple hierarchies for semantic hierarchical classification. Int J Mach Learn Comput, 2016
40 Deng J, Satheesh S, Berg A C, et al. Fast and balanced: efficient label tree learning for large scale object recognition. In: Proceedings of the International Conference on Neural Information Processing Systems, Granada, 2011
41 Jernite Y, Choromanska A, Sontag D. Simultaneous learning of trees and representations for extreme classification and density estimation. In: Proceedings of the International Conference on Machine Learning, Sydney, 2017
42 Zhou Y C, Hu Q H, Wang Y. Deep super-class learning for long-tail distributed image classification. Pattern Recogn, 2018
43 Tang B, Kay S, He H. Toward optimal feature selection in naive bayes for text categorization. IEEE Trans Knowl Data Eng, 2016, 28: 2508–2521
44 Zhang T, Ren P, Ge Y, et al. Learning proximity relations for feature selection. IEEE Trans Knowl Data Eng, 2016, 28: 1231–1244
45 Yang Y, Shen H T, Ma Z, et al. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In: Proceedings of the International Joint Conference on Artificial Intelligence, Barcelona, 2011. 1589–1594
46 Freeman C, Kulić D, Basir O. Feature-selected tree-based classification. IEEE Trans Cybernet, 2013, 43: 1990–2004
47 Freeman C, Kulić D, Basir O. Joint feature selection and hierarchical classifier design. In: Proceedings of the IEEE

International Conference on Systems, Man, and Cybernetics, Anchorage, 2011. 1728–1734
48 Grimaudo L, Mellia M, Baralis E. Hierarchical learning for fine grained Internet traffic classification. In: Proceedings of the International Wireless Communications and Mobile Computing Conference, Limassol, 2012. 463–468
49 Peng H C, Long F H, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell, 2005, 27: 1226–1238
50 Feng S H, Lang C Y, Feng J S, et al. Human facial age estimation by cost-sensitive label ranking and trace norm regularization. IEEE Trans Multimedia, 2017, 19: 136–148
51 Wen Y D, Zhang K P, Li Z F, et al. A discriminative feature learning approach for deep face recognition. In: Proceedings of the European Conference on Computer Vision, Amsterdam, 2016. 499–515
52 Decoro C, Barutcuoglu Z, Fiebrink R. Bayesian aggregation for hierarchical genre classification. In: Proceedings of the International Conference on Music Information Retrieval, Vienna, 2007. 77–80
53 Bennett P N, Nguyen N. Refined experts: improving classification in large taxonomies. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, 2009. 11–18
54 Sun M, Huang W, Savarese S. Find the best path: an efficient and accurate classifier for image hierarchies. In: Proceedings of the IEEE International Conference on Computer Vision, Sydney, 2014. 265–272
55 He H, Garcia E A. Learning from imbalanced data. IEEE Trans Knowl Data Eng, 2009, 21: 1263–1284
56 Ramanathan V, Li C, Deng J, et al. Learning semantic relationships for better action retrieval in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 1100–1109
57 Shuai B, Zuo Z, Wang B, et al. Dag-recurrent neural networks for scene labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 3620–3629
58 Wang H, Wu J J, Yuan S, et al. On characterizing scale effect of chinese mutual funds via text mining. Signal Process, 2016, 124: 266–278
59 Zhao B, Li F F, Xing E P. Large-scale category structure aware image categorization. In: Proceedings of the International Conference on Neural Information Processing Systems, Granada, 2011. 1251–1259
60 Mccallum A, Rosenfeld R, Mitchell T M, et al. Improving text classification by shrinkage in a hierarchy of classes. In: Proceedings of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 1998. 359–367
61 Shahbaba B, Neal R M. Improving classification when a class hierarchy is available using a hierarchy-based prior. Bayesian Anal, 2005, 2: 221–237
62 Wu H, Merler M, Uceda-Sosa R, et al. Learning to make better mistakes: semantics-aware visual food recognition. In: Proceedings of the ACM on Multimedia Conference, Amsterdam, 2016. 172–176
63 Fan J P, Zhao T Y, Kuang Z Z, et al. Hd-mtl: hierarchical deep multi-task learning for large-scale visual recognition. IEEE Trans Image Process, 2017, 26: 1923–1938
64 Zhou D, Xiao L, Wu M R. Hierarchical classification via orthogonal transfer. In: Proceedings of the International Conference on Machine Learning, Bellevue, 2011. 801–808
65 Xie S N, Yang T B, Wang X Y, et al. Hyper-class augmented and regularized deep learning for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 2645–2654
66 Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, Kerkyra, 1999
67 Nchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: theory and practice. Int J Comput Vision, 2013, 105: 222–245
68 Luo J H, Wu J X. A survey on fine-grained image categorization using deep convolutional features. Acta Autom Sin, 2017, 43: 1306–1318
69 Sun A, Lim E P. Hierarchical text classification and evaluation. In: Proceedings of the IEEE International Conference on Data Mining, San Jose, 2001. 521–528
70 Ran E Y, Wiener Y. On the foundations of noise-free selective classification. J Mach Learn Res, 2010, 11: 1605–1641

71 Yuan M, Wegkamp M. Classification methods with reject option based on convex risk minimization. J Mach Learn Res, 2010, 11: 111–130
72 Hanczar B, Dougherty E R. Classification with reject option in gene expression data. Bioinformatics, 2008, 24: 1889–1895
73 D'Alessio S, Murray K, Schiaffino R, et al. The effect of using hierarchical classifiers in text categorization. Content-Based Multimedia Inf Access, 2000, 1: 302–313
74 Sun A, Lim E P, Ng W K, et al. Blocking reduction strategies in hierarchical text classification. IEEE Trans Knowl Data Eng, 2004, 16: 1305–1308
75 Ceci M, Malerba D. Classifying web documents in a hierarchy of categories: a comprehensive study. J Intell Inf Syst, 2007, 28: 37–78
76 Wang Y, Hu Q H, Zhou Y C, et al. Local bayes risk minimization based stopping strategy for hierarchical classification. In: Proceedings of the IEEE International Conference on Data Mining, New Orleans, 2017
77 Babbar R, Partalas I, Gaussier E, et al. Learning taxonomy adaptation in large-scale classification. J Mach Learn Res, 2016, 17: 3350–3386
78 Naik A, Rangwala H. Inconsistent node flattening for improving top-down hierarchical classification. In: Proceedings of the IEEE International Conference on Data Science and Advanced Analytics, Montreal, 2016. 379–388
79 Naik A, Rangwala H. Filter based taxonomy modification for improving hierarchical classification. ArXiv: 1603.00772

Review on hierarchical learning methods for large-scale classification task

Qinghua HU[1*], Yu WANG[1], Yucan ZHOU[1], Hong ZHAO[1], Yuhua QIAN[2] & Jiye LIANG[2]

1. *School of Computer Science and Technology, Tianjin University, Tianjin* 300350*, China*;
2. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China*
* Corresponding author. E-mail: huqinghua@tju.edu.cn

Abstract Hierarchical classification is a task that uses hierarchy of categories in data. It can handle large-scale data. In recent years, significant research has emerged in this field, which is receiving increasingly more attention. In this paper, we first introduce the definition of hierarchical classification and thereafter review the important studies on several basic issues in large-scale hierarchical classification tasks based on different problem-solving strategies. First, we define the hierarchy formally and introduce some hierarchical evaluation metrics. Second, we explain how to construct the hierarchy, how to learn classifiers and perform feature selection using the information in the hierarchy, and how to design stopping strategies and introduce some representative studies on each issue. Finally, we summarize the features of large-scale hierarchical classification task and discuss the possible future work in this field.

Keywords hierarchical classification, hierarchy construction, hierarchical classifier learning, hierarchical stopping strategy, hierarchical feature selection, classification

Qinghua HU received his B.S., M.S., and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively. He was a post-doctoral fellow in the Department of Computing, Hong Kong Polytechnic University from 2009 to 2011. He is currently a full professor and the vice dean of School of Computer Science and Technology, Tianjin University, Tianjin, China. His current research interests include rough sets, granular computing, and data mining for classification and regression.

Jiye LIANG received his M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1990 and 2001, respectively. He is currently a professor in School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of the Ministry of Education, Shanxi University, Taiyuan, China. His current research interests include computational intelligence, rough set theory, and granular computing.

Yu WANG received his B.S. and M.E. degrees from Tianjin University in 2013 and 2016, respectively. He is currently a Ph.D. student in School of Computer Science and Technology in Tianjin University. His research interests include machine learning, data mining, hierarchical classification, uncertainty, large-scale data classification tasks, and uncertainty modeling.

Yuhua QIAN received his M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively. He is currently a professor of Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. He is actively pursuing research in pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence.