基于路径选择的层次多标签分类

张春焰,李 涛,刘 峥
(南京邮电大学 计算机学院,江苏 南京 210046)

摘 要:多标签分类为每一个实例分配多个标签,当这些标签存在一种预定义的层次化结构时,该机器学习任务称为层次多标签分类(HMC)。传统的分类问题(二分类和多标签分类)往往会忽略各标签之间的结构关系,而层次多标签分类充分考虑标签集之间的层次结构关系,并以此来提高分类的效果。层次多标签分类是输出结构化预测结果的分类任务,其中类标签被组织成某种预定义(树形或者有向无环图)的结构,并且一个实例可以属于多个类。在 HMC 中有基于全局标签集的分类方法和基于单个标签的局部分类方法。全局方法将整个问题作为一个整体来处理,但往往会随着数据集的增长而出现性能瓶颈,而局部方法将问题分解为基于单个标签的二分类方法,但未充分考虑层次结构信息,并且无法处理预测节点终止于层次标签树内节点的分类问题。在分类阶段,修剪掉概率较低的分支,达到预测标签不一定到达叶子节点的目的。基于路径选择的层次多标签分类充分考虑修剪后的层次标签树从根节点出发的所有可能路径,结合各节点的预测概率值和节点所在的层次来选出得分最高的标签路径。该方法和现有的层次多标签分类方法在三种不同的数据集上进行实验对比,结果表明该方法在处理层次较深且叶子节点稠密的层次结构时获得了较好的结果。

关键词:层次多标签分类;多标签学习;路径选择;层次分类;文本分类;层次标签树;剪枝
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2018)10-0037-07
doi:10.3969/j.issn.1673-629X.2018.10.008

Hierarchy Multi-label Classification Based on Path Selection

ZHANG Chun-yan, LI Tao, LIU Zheng
(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: Multi-label classification assigns more than one label for each instance when the labels are ordered in a predefined structure. The task is called hierarchical multi-label classification (HMC). Traditional classification problems (binary classification and multi-label classification) tend to ignore the structural relationship between the labels, and hierarchical multi-label classification takes full account of the hierarchical relationship between the label sets, thus improving the classification effect. HMC is a task of structured output prediction where the classes are organized into a hierarchy and an instance may belong to multiple classes. The hierarchy structure that organizes the set of classes can assume the form of a tree or of a directed acyclic graph (DAG). In HMC there are global and local approaches. Global approaches treat the problem as whole but tend to explode with large datasets. Local approaches divide the problem into local subproblems, but usually do not exploit the information of the hierarchy. The hierarchical multi-label classification based on path selection studies the problem that the classification label does not reach the leaf node of the label tree. In the classification phase, the branches with low probability to occur are pruned, performing non-mandatory leaf node prediction. This method evaluates each possible path from the root of the hierarchy, taking into account the prediction value and the level of the nodes, selecting one or more label paths whose score is above a threshold. It has been tested in three datasets with tree hierarchy structured hierarchies against a number of state-of-the-art methods. The experiment shows that this method can obtain superior results when dealing with deep and populated hierarchies.

Key words: hierarchical multi-label classification; multi-label learning; path selection; hierarchical classification; text classification; hierarchical label tree; pruning

收稿日期:2017-11-04 **修回日期**:2018-03-13 **网络出版时间**:2018-05-16
基金项目:2015 年教育部-中国移动科研基金项目(5-10);江苏省自然科学基金面上项目(BK20171447);江苏省高校自然科学研究面上项目(17JKB520024)
作者简介:张春焰(1992-),男,硕士,研究方向为数据挖掘;李 涛,通讯作者,博士,美国佛罗里达国际大学正教授,研究方向为数据挖掘、机器学习和信息检索及生物信息学等。

0 引 言

层次分类(hierarchical classification,HC)问题是分类问题的一个分支,在层次分类问题中类别是不相交的,而是以分层结果组织的。在该情况下,一个样本会与给定的类别标签及该标签的父标签相关联,而组织类的层次结构可以采用树或者有向无环图(directed acyclic graph,DAG)的形式。存在复杂 HC 问题的子集,其中每个样本可以与类层次结构中的多个路径相关联,即层次多标签分类(HMC)[1-3]。典型的 HMC 问题是文本分类[4-6]和生物信息学任务,如蛋白质功能检测[7-8]、图像分类[9-12]等。

HMC 算法一般通过优化局部或者全局的损失函数来选择层次标签树上一条或者多条路径以标示实例[13]。执行局部学习的算法尝试挖掘类层次结构的区域特征,然后将预测结果进行组合得到最终分类结果。而基于全局的方法往往由单个分类器组成,并且能够一次性找出实例相关联的类标签。传统的分类任务主要解决的是一个样本 e 只会对应于单个标签 $y \in L$ 的问题,其中 L 是标签的集合,标签的数量大于等于二,即通常所说的多类分类。然而,有些分类问题会更加复杂,因为一个样本可以对应多个标签。一个多标签数据集 D 由 N 个样本组成,每一个样本会对应一个标签集合 Y,其中 $Y \in L$。当这些标签之间存在某种预定义的结构(树形或者有向无环图)时,该任务称为层次多标签分类。树形结构与有向无环图的主要区别在于图形结构中一个节点可以有多于一个的父节点,为了简化问题,文中只考虑树形层次结构。

基于局部分类方法,在层次结构中的每一个标签节点训练一个分类器的基础上提出新的 HMC 方法,通过分解问题来达到层次多标签分类。对比之前的层次多标签分类方法,该方法的主要改进有以下几点:

· 为层次树的节点加权,使得每个节点标签分类错误的权重随着层次标签树层次的下降而衰减。

· 通过组合各节点的概率值和节点在路径中的位置,对所有可能的预测路径进行打分,从而选出最佳路径,即预测标签集。

· 通过在寻找最优的预测路径之前对层次树进行裁剪,解决预测路径不在层次树叶子节点终止的层次多标签分类任务,即最具体的类别并未到达叶子节点。对层次标签树进行裁剪可以抛弃那些在真实标签集中不大可能出现的分支,减少了计算量和潜在的错误。

1 相关工作

在 HMC 任务中,属于某个类的示例自动属于这个类的所有超类(层次约束)。这里有两种预测结果[1]:强制性叶节点预测,返回的是从根节点到叶子节点的完整路径;非强制性叶节点预测,预测出来的路径可能未到达叶子节点。

HMC 根据其用于解决分类问题的探测策略进行分类,其中比较常见的方法为:直接法、全局法、自顶向下。直接法的分类策略借鉴于传统的多标签分类,只能预测层次树中的叶子节点,预测出一个叶子节点则到达该叶子节点路径上的所有节点都被标记为正,该方法有一个显著的缺点是完全忽略了标签之间的层次关系。然而,这种简单的分类策略必须解决的是,需要训练一个分类器来区分一个庞大的目标标签集(所有叶子节点),并且没有利用标签集提供的层次关系。当然,这种方法只能预测层次树中的叶子节点,所以对于非强制性叶节点的数据集也就无能为力。全局方法对标签集中的所有类学习一个单一的模型来预测层次树中的所有类,例如,对于一个深度为 3 的二叉树,总共有 14 个非叶子节点和叶子节点,所以基于全局的分类方法需要训练一个针对 14 个标签的分类器。该分类算法生成的模型在一次运行期间考虑了类层次结构上所有的类标签。基于全局方法的主要限制是随着数据集的增大,模型会过于复杂并且训练模型会花费大量时间。该方面已有不少研究成果,Vens[13],Blockeel 和 Bruynooghe[14]等都是基于预测聚类树(PCT)的决策树归纳算法来解决 HMC 问题,后者根据层次树中的多个节点的距离来导出预测类向量与真实类向量的距离度量。

基于局部分类器的 HMC 通过挖掘节点在层次结构中的局部信息来考虑类标签的层次关系,该方法可以根据局部上下文信息的组织方式和本地分类器构建方式的不同加以区分。特别地,主要有三种利用局部信息的方式:为每一个节点建立一个分类器(LCN),为每一个父节点建立一个分类器(LCNP),为每一个层次建立一个分类器(LCL)。这三种局部层次分类算法在模型训练阶段存在显著差异,但是在预测阶段都是基于相似的自顶向下方法。在预测阶段,这种自顶向下的方法先预测该层次树的根节点的类,然后根据预测出来的类缩小在下一层次所需预测的类的数目,如此循环直至所有特殊的节点都被预测出来。该自顶向下模式的限制是上层的分类错误将在层次结构中向下传播。

LCN 方法为层次树中的所有节点都训练一个二分类器(除了根节点)。Bi 和 Kwok[15-16]提出了 HIROM 方法,该方法使用独立于局部模型的局部预测方法,同时使用贪心算法在层次标签树中搜索满足层次约束的结果标签集。采用贝叶斯决策理论,通过最小化条件风险来得出最优预测。该方法的局限性在于该优化指标在其他评估措施中不一定有效。LCNP 方法

为层次标签树中的所有父节点训练一个多标签分类器,以区分各个子节点。在预测阶段,该方法也可以结合上述自顶向下的预测方法。

2　路径选择

在机器学习领域,多标签分类[17-20]受到广泛关注,基于差别排名的方法已在文献[21]提出,同时通过标签之间的依赖性来优化分类结果,但是当标签分层次组织时,却很难发挥作用[2]。图 1 为一棵层次标签树,针对该任务,提出了基于路径选择的层次多标签分类(based on path seletion, BPS)。该方法通过探索层次树中标签节点和该节点的所有祖先节点的上下文关系来得到最优的分类结果,同时使用计算规则评估从根到叶或中间节点的每个可能路径,该计算规则考虑了预测标签节点所在的层次来计算各个路径的得分,并最终返回满足阈值条件和层次约束的最优节点集。

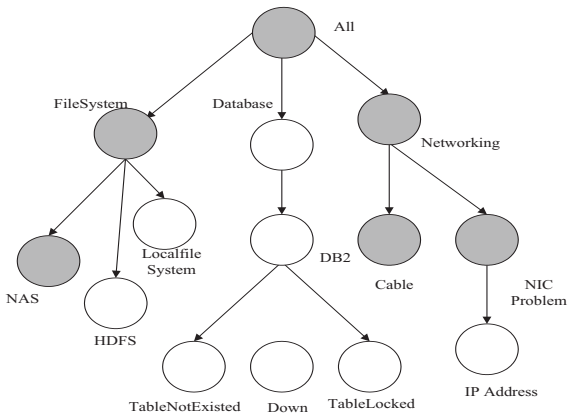

图 1　层次标签树

2.1　训练分类器

令 D 为具有 N 个实例的训练集, $E_e=(\boldsymbol{X}_e, Y_e)$,其中 $\boldsymbol{X}_e$ 为 d 维特征向量, $Y_e \in L$, $L=\{y_1, y_2, \cdots, y_{|l|}\}$,表示 $|l|$ 可能的标签或类组成的有限集合,即所有样本对应的标签集合。需要注意的是与 L 相比, Y_e 是一个较小的集合。 Y_e 可以用一个 0/1 向量来表示,即 $Y_e \in \{0,1\}^{|l|}$,当且仅当 $y_i \in Y_e$ 时, $y_i=1$,否则 $y_i=0$。基于路径选择的层次多标签分类算法,除了根节点外,在层次树中的所有其他节点都表示一个标签或者一个类,用 y_i 表示,其中 i 为层次树按层次遍历的次序。对于每一个非叶子节点 y_i ,训练一个标签多类分类器 C_i ,后面称之为基分类器。使用该方法对较大的层次结构具有很好的扩展性,只要返回与预测类相关联的概率值或其返回值可以转化为概率值的多类分类器都可以作为基分类器。对于多类分类器 C_i 所包含的预测类别标签有节点标签 y_i 的所有孩子节点标签(child(y_i)),上一个“unknown”标签代表那些不属于 child(y_i) 的样本。

把多类分类器 C_i 的训练样本分成两部分,其中一部分样本由 child(y_i) = 1 的实例组成,即这些实例对应的标签集都含有 y_i 的孩子节点标签(child(y_i) $\in Y_e$),用 $\mathrm{Tr}^+(C_i)$ 表示。在这部分训练样本集中,每一个样本都会打上对应的 child(y_i) 标签。另一部分样本由那些含有 y_i 的兄弟节点标签(sib(y_i))的不同样本组成,用 $\mathrm{Tr}^-(C_i)$ 表示。如果 y_i 没有兄弟节点,则在层次标签树中向上搜索,找到离 y_i 最近的含有兄弟节点的祖先节点(sib(pa(y_i))),这些兄弟节点由 y_i 父节点的所有子节点的集合除去节点 y_i 的子集构成。 $\mathrm{Tr}^-(C_i)$ 对应样本集的标签不会含有 y_i 的孩子标签,把这部分样本标记为“unknown”标签。同时考虑到训练数据的平衡性,需要对这部分数据集进行欠采样,欠采样的数量与每个 y_i 孩子节点对应训练样本的平均值成正比。图 2 描绘了构造标签节点 y_5 局部分类器 C_5 的训练集实例的过程。数据集 $\mathrm{Tr}^+(C_5)$ 由满足条件 child(y_5) = $\{y_6, y_7\}$ = 1 的样本组成,这些样本都标记为对应的 child(y_5) 标签。而 $\mathrm{Tr}^-(C_5)$ 由 sib(y_5) = $\{y_8\}$ 的样本组成,该数据集的样本都标记为“unknown”。同时需要对该数据集进行欠采样,从而保证该样本集的数量与训练集中 child(y_i) 的平均样本数成比例。

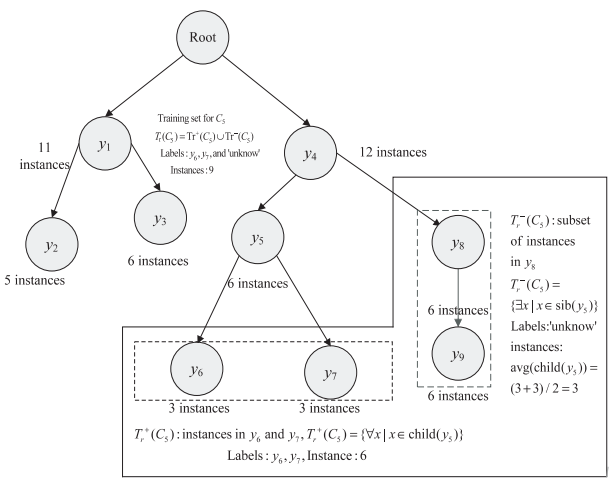

图 2　基分类器

2.2　裁剪层次标签树

某些情况下根据现有的信息不能很有把握地估计一个样本在层次标签树底部的标签,为了让预测的标签节点路径在未到叶子节点前终止,需要对层次标签树进行裁剪。裁剪可以以自下而上或者自上而下的方式进行,这取决于层次结构如何遍历修剪,同时可以在分类阶段之前执行,或者首先修剪分类阶段所选择的路径,同时裁剪可以根据不同的条件进行。文献[2]中进行了几个实验评估裁剪的最优策略,根据该结果选用自顶向下测试标签节点的每条后代分支进行裁

剪。为了对一个新的实例进行分类,实例的标签集对应于层次标签树的一条路径或者多条路径。为了选出对应的路径,需要为每条可能的路径计算出相应的概率值。对层次树裁剪需要在计算各路径得分之前进行,裁剪路径的条件是节点最可能的子节点的概率值小于“unkown”标签的概率。该方法的理由是,当一个分类器预测的最可能的类不是该分类器对应标签节点的孩子节点时,就有很大的把握对该路径进行剪枝。

裁剪过程由算法 1 进行描述,该过程发生在获得针对给定样本 x_i 的情况下层次标签树中每个节点的概率值。从根节点开始,将节点 y_i 最可能的孩子节点的概率值(maxConfidence)与节点未知的概率值(unknown)相继进行比较。如果 maxConfidence 大于 unknown,则该过程在节点 y_i 的各子节点上迭代进行,否则节点 y_i 的子孙节点都将剪去,如果一个节点没有兄弟节点,则在层次标签树中向上搜索其祖先节点的兄弟节点。为了在修剪后的层次标签树中搜索出最优的一条或者多条路径,需要为层次标签树中的路径计算相应的得分。

2.3 计算层次标签树路径得分

该合并规则将路径上每个局部分类器的预测结果合并为一个分值,并且考虑标签节点在层次结构中的层次,以确定该节点在所有分值中的权重。错误的分类发生在层次标签树的顶部的代价往往比发生在层次标签树的底部大,在层次树的顶部有更多的训练样本并且类别之间也有更大的差异,对分类的贡献也就更大。

为了达到该目的,节点 y_i 的权重定义如式 1。

$$w(y_i) = 1 - \frac{\text{level}(y_i)}{\text{maxlevel} + 1} \tag{1}$$

其中, $\text{level}(y_i)$ 表示节点标签 y_i 在层次标签树的层次,即该节点的父节点层次加 1; maxlevel 表示层次标签树中最长路径的长度。式 1 定义的节点权重可以随着节点层次的降低而线性衰减,从而保证权重沿层次分布均匀。使得较低层次节点的权重既不会快速下降为 0[12],也不会衰减得太慢[13]。

式 2 定义了每条路径的得分计算方法,它是沿着路径节点上预测概率的加权和。

$$\text{score}_k = \sum_{i=1}^{q} w(y_i) * p(y_i \mid x_e) \tag{2}$$

其中, q 为路径中的节点数; y_i 为路径中的第 i 个节点; $p(y_i \mid x_e)$ 为实例 x_e 在节点 y_i 被局部分类器预测为真的概率;下标 k 表示第 k 条路径,则 score_k 为第 k 条路径的得分。

图 3 描述了计算路径得分的执行过程。首先计算概率和权重,然后利用式 2 合并成一个分值,选定相应的阈值 σ ,分值大于 σ 的路径都作为最终预测返回。阈值 σ 可以根据测试数据集进行训练得出,返回路径上除了根节点以外的节点标签组成的集合就是预测样本对应的预测标签集。假设训练得到阈值 σ 为 0.5,在图 3 的层次标签树各路径的得分中有两条路径得分大于 0.5,故返回的集合为 $\{y_1, y_2, y_6, y_7, y_{10}\}$ 。

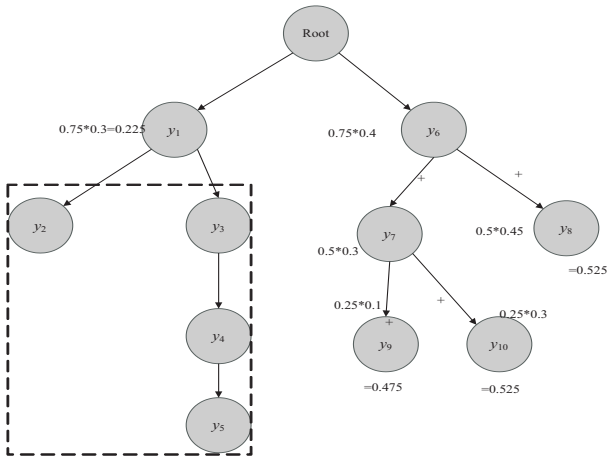

图 3 路径得分计算示例

3 实验结果与分析

实验使用了三种数据集,其中两种来自功能基因组学领域[6,8],一种来自图像分类领域[22-23],见表 1,其中 $|L|$ 为类别标签的个数, A 为特征维度, N 为样本个数, D 为层次标签树的高度。

表 1 实验数据

Dataset	$\|L\|$	A	N	D
cellcyle_FUN	36	78	2 339	4
spo_GO	53	81	1 685	11
caltech-256	319	12 000	7 680	6

算法描述如下:

算法 1:Prune 裁剪层次标签树算法, $\text{LC}(y_i)$ 表示在节点 y_i 运行局部分类器

输入:confidences(一个由在位置 j 的标签节点 y_j 的概率值 $p(y_i \mid x_e)$ 组成的数组), y_i (层次标签树的根节点), C (由层次标签树中除叶子节点外每个标签节点的基分类器组成的集合)

输出:裁剪之后的层次标签树

if y_i is not a left and has not been visited before then
if all pa(y_i) have been visited then
mark y_i visited
maxConfidence = 0
totalConfidence = 0
for all $y_j \in \text{child}(y_i)$ do
totalConfidence = toalConfidence+confidences[j]
if confidences[j]>maxConfidence then
maxConfidence = confidences[j]
end if

end for
Unknown = 1 − totalConfidences
if maxConfidence>unknown then
for all $y_j \in child(y_i)$ do
$LC(y_i)$
end for
else
Prune descendants of y_i
end if
else
continue with another node
end if
end if

3.1　评估指标

在 HMC 的情况下,评估指标的定义并不像二分类那么直观,因为预测可以是部分正确的,对于该问题已提出相关针对多标签分类[24]和 HMC[25]的特殊的评估指标。

令 $X=(x_0,x_1,\cdots,x_{d-1})$ 是来自 d 维输入特征空间 Y 的一条样本, $\boldsymbol{Y}=(y_0,y_1,\cdots,y_{L-1})$ 为 L 维输出类标签向量,其中 $y_i \in \{0,1\}$。多标签分类为每条样本 X 分配一个多标签向量 $\boldsymbol{y}$,当样本 X 属于第 i 个类时, $y_i=1$,否则 $y_i=0$。用 y_i 表示真实的标签集,而 $\hat{y_i}$ 表示预测的标签集。

Full−Match:式 3 表示测试集中预测的标签路径完全正确的样本占全部样本的比例,即预测的标签集和真实的标签集完全相同。

$$\text{FullMatch}=\frac{1}{N}\sum_{i=1}^{N}1_{Y_i=\hat{Y}_i} \tag{3}$$

Accuracy[26]:式 4 表示预测标签集和真实标签集交集的大小与并集的大小的比例。

$$\text{Accuracy}=\frac{1}{N}\sum_{i=1}^{N}\frac{|\boldsymbol{Y}_i\cap\hat{Y}_i|}{|\boldsymbol{Y}_i\cup\hat{Y}_i|} \tag{4}$$

F1−measure[27]:对于多标签分类,F1 − measure 定义为式 6,但需要重新定义查准率和查全率。查准率:真实标签集中有多少比例的标签被正确预测,即 $\frac{z_i\cap\hat{z}_i}{\hat{z}_i}$。查全率:预测的标签集中正确预测的标签所占的比例,即 $\frac{z_i\cap\hat{z}_i}{z_i}$。

$$F_1=\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}} \tag{5}$$

用向量 $\boldsymbol{Z}$ 替换向量 $\boldsymbol{Y}_i$。对于多标签分类,有多种方法对 F1−measure 进行平均,主要取以下两种方法:

F1−macroD 是对样本个数取平均:

$$\text{F1-macroD}=\frac{1}{N}\sum_{i=1}^{N}F_1(z_i,\hat{z_i}) \tag{6}$$

其中, $z_i\equiv\boldsymbol{Y}_i$, N 为样本个数。

F1−macroL 是对标签个数取平均:

$$\text{F1-macroL}=\frac{1}{|L|}\sum_{i=1}^{N}F_1(z_i,\hat{z_i}) \tag{7}$$

其中, $z_i\equiv[y_i^1,\cdots,y_i^N]$, N 为样本个数, $|L|$ 为标签个数。

F1−macroL 指标对每一个标签节点的分类效果进行评估,而 F1−macroD 指标是对每一个样本的分类效果进行评估。

3.2　基分类器

该节设计了一个实验来分析不同基分类器对基于路径选择的层次多标签分类算法性能的影响,并选择最适合该数据集的基分类器对算法进行测试。使用十折交叉验证对三份数据集进行实验,根据 3.1 节介绍的四种不同的层次度量方法来比较该层次多标签分类算法的性能。使用四种常见的方法作为基分类器:支持向量机(SVM),核函数使用多项式核函数;C4.5,用于剪枝的置信度设置为 0.35,每个叶子的最小实例数设置为 3;朴素贝叶斯(NB);随机森林(RF),生成十棵树。

实验结果见表 2。

表 2　不同基分类器的性能度量　　%

base classifier	C4.5	SVM	NB	RF
Accuracy	21.04	18.34	25.05	29.45
Full−Match	8.10	12.32	16.50	21.26
F1−macroD	30.10	23.45	30.45	36.76
F1−macroL	3.50	3.67	14.45	14.23

可以观察到,随机森林在所有评估指标中表现最好,其次是朴素贝叶斯。因为随机森林能够处理比较大的不平衡数据,并且能够有效处理存在噪声的数据,所选择的验证数据都存在这些特征。因此,选用随机森林作为其余实验的基分类器。

3.3　与多类分类和多标签分类的比较

在该实验中,将文中方法与两个未考虑层次结构的替代方法进行对比。

(1)多类分类(multi−class classification, MC):一种仅预测层次结构叶子节点的多类分类器。

(2)链式分类器(classification chains, CC):为层次标签树中每一个标签节点分别建立一个二分类器,并且各个分类器根据父子关系构成链式结构的多标签分类方法。在预测阶段,它根据链式结构将先前分类器的输出结果作为附加属性,然后在层次标签树中选择最优子树标签集作为最终结果输出[28-29]。

实验结果见图 4。

measure	BPS	MC	CC
FUN datasets			
Accuracy	21.23	24.34	15.21
Full-match	17.45	17.64	7.89
F1-macro D	25.67	26.89	17.78
F1-macro L	14.23	15.45	9.78
GO datasets			
Accuracy	38.56	38.45	38.34
Full-match	23.96	21.93	9.56
F1-macro D	49.67	46.67	51.45
F1-macro L	14.56	15.56	9.56
measure	BPS	MC	
Caltech datasets			
Accuracy	26.56	18.45	
Full-match	10.23	6.34	
F1-macro D	35.45	24.45	
F1-macro L	12.45	8.67	

图 4 三种分类方法在数据集上的表现

所有数据都是在各个数据集上相应指标的平均值,同时记录了对应的标准差。从图中可以看出,尽管 MC 方法在大多数指标下都是较优的,并且有一些显著优越的结果,但是 BPS 在层次标签树较浅或者叶子节点较少的数据集中具有竞争力(Fun 数据集层次结构为 4 层,大约 15 个叶子节点)。可以看到该方法在 Accuracy 和 Full-match 两个指标有较好的结果,这些指标对预测结果中出现的错误进行了度量。这意味着文中方法如果基分类器输出的概率值过低,会对层次标签树进行裁剪,而 MC 方法返回完整的路径可能是不正确的。对于层次标签树高度较高但标签节点不是很稠密的数据集(GO 数据集层次结构为 11 层,大约 24 个叶子节点),大多数情况下,文中方法比 MC 方法能获得更好的结果。

在基因组数据集中,文中方法和 MC 方法的性能差异很小,是因为在叶子节点较少的层次结构中,MC 方法往往能返回良好的结果。然而,从 Caltech 数据集的实验结果可以看到,当层次标签树中有较多的叶子节点时就很难用单一的分类器来区分不同的路径。Caltech 数据集含有 256 和 84 个叶子节点,五个度量指标数据都表明笔者的方法要优于 MC 方法。故在大型层次结构中文中方法要优于 MC 方法。同时基于路径选择的层次多标签分类(BPS)对于多标签链分类(CC)也是很有竞争力的,除了在 Full-match 指标下 BPS 要明显优于 CC,其他指标都有相似的结果,但是 BPS 计算性能要优于 CC。

4 结束语

提出了一种新颖的基于路径选择的层次多标签分类方法,可以用于解决标签节点路径在层次标签树内节点终止的数据集。BPS 为层次标签树中每一个内节点训练一个多类分类器,同时用含有该节点兄弟节点标签的数据集构成未知标签样本,用于层次树的裁剪。该方法在层次标签树中选择路径得分超过一定阈值的一条或多条路径,其中路径得分是结合路径上对应标签节点基分类器的预测值和该节点在层次标签树中的层次赋予不同的权重计算得到。为了使预测的标签路径在层次标签树的内节点终止,需要根据各标签节点基分类器对应的子节点概率值进行层次标签树的裁剪,从而消除可能性较低的分支。使用三份不同的数据对基于路径选择的层次多标签分类方法进行验证,同时与现有方法进行比较,结果表明该方法均能取得较好的结果,并且在层次较深且叶子节点较多的层次结构表现更优。

参考文献:

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

[2] CESA-BIANCHI N, GENTILE C, TIRONI A, et al. Incremental algorithms for hierarchical classification[J]. Journal of Machine Learning Research, 2006, 7: 31-54.

[3] MAYNE A, PERRY R. Hierarchically classifying documents with multiple labels[C]//IEEE symposium on computational intelligence and data mining. Nashville, TN, USA: IEEE, 2009: 133-139.

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

[5] SCHIETGAT L, VENS C, STRUYF J, et al. Predicting gene function using hierarchical multi-label decision tree ensembles[J]. BMC Bioinformatics, 2010, 11: 2-7.

[6] RUEPP A, ZOLLNER A, MAIER D, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes[J]. Nucleic Acids Research, 2004, 32(18): 5539-5545.

[7] OTERO F E B, FREITAS A A, JOHNSON C G. A hierarchical multi-label classification ant colony algorithm for protein function prediction[J]. Memetic Computing, 2010, 2(3): 165-181.

[8] VALENTINI G. True path rule hierarchical ensembles for genome-wide gene function prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(3): 832-847.

[9] DIMITROVSKI I, KOCEV D, LOSKOVSKA S, et al. Hierarchical classification of diatom images using ensembles of predictive clustering trees[J]. Ecological Informatics, 2012, 7(1): 19-29.

[10] 陈自洁. 多标签分类问题的图结构描述及若干学习算法的研究[D]. 广州: 华南理工大学, 2015.

[11] 焦 阳. 基于主动学习的多标签图像分类方法研究[D]. 苏州: 苏州大学, 2015.

[12] BI Wei, KWOK J T. Hierarchical multilabel classification with minimum bayes risk[C]//12th international conference on data mining. Brussels, Belgium: IEEE, 2012: 101-110.

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

[14] BLOCKEEL H, BRUYNOOGHE M, DŽEROSKI S, et al. Hierarchical multi-classification[C]//Workshop notes of the KDD'02 workshop on multi-relational data mining. [s. l.]: [s. n.], 2002: 21-35.

[15] ALVES R T, DELGADO M R, FREITAS A A. Multi-label hierarchical classification of protein functions with artificial immune systems[C]//Brazilian symposium on bioinformatics. Berlin: Springer, 2008: 1-12.

[16] TSOUMAKAS G, KATAKIS I, VLAHAVAS I. Mining multi-label data[M]//Data mining and knowledge discovery handbook. Boston, MA: Springer, 2009: 667-685.

[17] 徐有正,黄　刚. 多标签图像的识别分类处理算法[J]. 计算机时代,2017(10):9-12.

[18] 郑晓雪,张大方,刁祖龙. 基于用户身份特征的多标签分类算法[J]. 计算机应用,2017,37(6):1697-1701.

[19] 姚小慧,孙国强. 基于局部和全局一致性的多标签分类算法[J]. 电子科技,2017,30(3):4-7.

[20] 王　梅. 基于多标签学习的图像语义自动标注研究[D]. 上海:复旦大学,2008.

[21] SHALEV-SHWARTZ S, SINGER Y. Efficient learning of label ranking by soft projections onto polyhedra[J]. Journal of Machine Learning Research, 2006, 7: 1567-1599.

[22] LI Feifei, FERGUS R, PERONA P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories[J]. Computer vision and Image understanding, 2007, 106(1): 59-70.

[23] VAN GEMERT J C, GEUSEBROEK J M, VEENMAN C J, et al. Kernel codebooks for scene categorization[C]//European conference on computer vision. Berlin: Springer, 2008: 696-709.

[24] READ J. Scalable multi-label classification[D]. New Zealand: University of Waikato, 2010.

[25] KOSMOPOULOS A, PARTALAS I, GAUSSIER E, et al. Evaluation measures for hierarchical classification: a unified view and novel approaches[J]. Data Mining and Knowledge Discovery, 2015, 29(3): 820-865.

[26] GODBOLE S, SARAWAGI S. Discriminative methods for multi-labeled classification[J]. Advances in Knowledge Discovery and Data Mining, 2004, 23(4): 22-30.

[27] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International Journal of Data Warehousing and Mining, 2006, 3(3): 34-56.

[28] HANG Minling, ZHOU Zhihua. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[29] 王　进,王　鸿,夏翠萍,等. 基于 Spark 的组合分类器链多标签分类方法[J]. 中国科学技术大学学报,2017,47(4):350-357.

(上接第 36 页)

recognition via sparse spatio-temporal features[C]//Joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance. Beijing, China: IEEE, 2006: 65-72.

[15] 王　博,赵君喜. 基于时空兴趣点的人体行为识别方法研究[D]. 南京:南京邮电大学,2014.

[16] SCOVANNER P, ALI S, SHAH M. A 3-dimensional sift descriptor and its application to action recognition[C]//Proceedings of the 15th international conference on multimedia. Augsburg, Germany: ACM, 2007: 357-360.

[17] JI Xiaofei, WU Qianqian, JU Zhaojie, et al. Study of human action recognition based on improved spatio-temporal features[J]. International Journal of Automation and Computing, 2014, 11(5): 500-509.

[18] 张世超,陈恩庆. 基于 Kinect 深度图像的动作识别[D]. 郑州:郑州大学,2016.

[19] HAERING N, VENETIANER P, LIPTON A. The evolution of video surveillance: an overview[J]. Machine Vision and Applications, 2008, 19(5-6): 279-290.

[20] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE computer society conference on computer vision and pattern recognition. San Diego, CA, USA: IEEE, 2005: 886-893.

[21] 姬晓飞,左鑫孟. 基于关键帧特征库统计特征的双人交互行为识别[J]. 计算机应用,2016,36(8):2287-2291.

[22] 姬晓飞,周　路,李一波. 基于 AdaBoost 算法特征提取的人体动作识别方法[J]. 沈阳航空航天大学学报,2014,31(2):65-69.

[23] JI Yanli, YE Guo, CHENG Hong. Interactive body part contrast mining for human interaction recognition[C]//IEEE international conference on multimedia and expo workshops. Chengdu, China: IEEE, 2014: 1-6.

[24] SONG Sijie, LAN Cuiling, XING Junliang, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[J]. IEEE Intelligent Systems, 2016, 31(2): 45-53.

[25] LIN Liang, WANG Keze, ZUO Wangmeng, et al. A deep structured model with radius - margin bound for 3D human activity recognition[J]. International Journal of Computer Vision, 2016, 118(2): 256-273.