****

**毕业设计（论文）中期报告**

**题 目：面向DAG标签结构数据的分类方法设计**

**专 业 测控技术与仪器**

**学 生 李慧莹**

**学 号 1170100603**

**指导教师 乔家庆**

**日 期 2020年12月1日**

**哈尔滨工业大学教务处制**

# 课题主要研究内容及进度情况

## 课题主要研究内容

本课题主要内容为设计DAG标签结构数据分类算法，算法将针对求解最小化条件风险进行数学建模，主要包含三个方面，针对有向无环图分类的损失函数设计，将层级多标签分类转换为二元分类的二元分类器选择，以及为满足层级约束的算法设计。针对有向无环图结构，在层级多标签分类问题中，需要注意分类结果必须符合预先定义的层级约束关系，这种关系可以概括为：若一个样本属于一个特定类别，则它同时属于这个类别的所有祖先；若一个样本不属于一个特定类别，那么同时它一定不属于这个类别的所有子孙类别。故本课题旨在得到面向DAG标签结构数据的符合层级约束要求的分类结果。

## 课题进度情况

2020年11月30日至2021年12月4日 学习相关理论及背景，撰写开题报告

2020年12月5日至2021年1月10日 分析支持向量机等分类器基本原理

2021年1月11日至2021年3月7日 实现多标签分类方法的处理流程

2021年3月8日至2021年4月8日 进行对比实验，分析评估本文所述层级多标签分类方法，撰写中期报告

# 目前已完成的研究工作及结果

## 总体研究方案

针对层级多标签分类问题（Hierarchical Multi-Label Classification）设计的算法必须能够将样本数据标记为符合层次结构中的一个或多个路径。为此，算法必须对损失函数进行局部或全局优化，对应于目前已有的两类HMC分类算法，局部分类方法和全局分类方法。执行局部学习的算法试图发现在层次结构的特定区域中类别关系的特殊性，随后结合局部预测以生成最终分类。其思想是按照自顶向下的策略生成分类器层次结构，其中每个分类器负责预测特定节点或特定层次结构。而HMC的全局方法通常仅由单个分类器组成，能够将样本数据与其在整个层次结构中的对应类相关联。使用全局或局部方法都有优点和缺点。全局方法通常比局部方法计算量小，并且它们不会受到错误传播问题的影响，但它们无法从层次结构中捕获局部信息，最终会出现欠拟合的现象。局部方法由于依赖于分类器的级联，因此计算开销要大得多，但是它们更适合于从层次结构中提取信息。本文所述方法属于局部分类方法，将层级多标签分类问题转换成一组二分类问题，并将层级结构信息单独加入考虑，在保证架构灵活的同时，减小了训练工作量。

## 多标签基础分类

### 数据预处理

实验所用数据集来自生物学实验数据长链非编码 RNA (lncRNAs)，在lncRNA共表达网络中，拓扑结构相似的节点可能具有相似的功能，于是采用DCA策略提取lncRNAs的低维拓扑信息。目前，由于没有lncRNAs的公共GO注释。因此，根据已知的蛋白质GO注释，采用邻域计数法对一些lncRNAs进行注释。这种方法基于这样一个事实，即目标lncRNA可能与lncRNA蛋白结合网络中的直接相邻蛋白具有非常相似的功能。

由于当某些 GO 标签包含的样本过少时，没有足够的信息可以用于分类器的训练，故对节点进行筛选，筛选策略为若全部数据集中含某节点数小于阈值，则丢弃此节点，一方面可以减少分类节点数量，另一方面保证每个节点有一定数量的正样本可供分类训练使用。此外，由于原始节点分布在三大生物学功能(MF)Molecular Function, (CC)Cellular Component和(BP)Biological Process中，每一种生物学功能有不同的根节点，对应于三种独立的DAG层级结构，对筛选后的节点进行分析，仅有少量不属于BP功能的节点存在，且构成的层级结构规模较小、不完整，于是经二次筛选，留下属于BP功能的204个节点，构成了含根节点在内的12级层次结构。

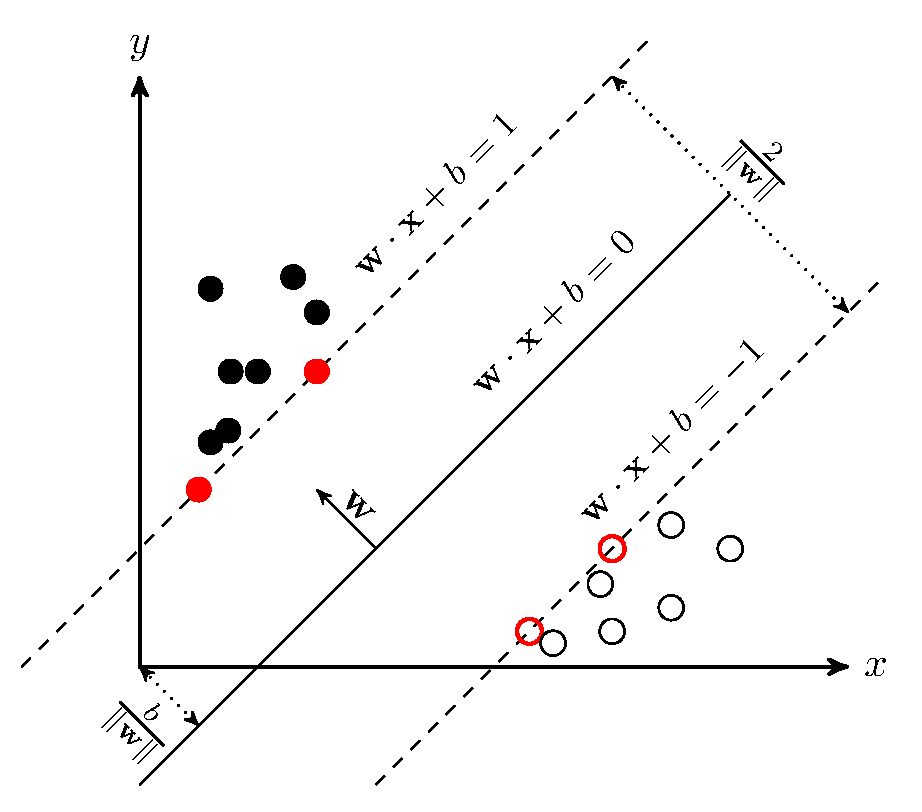
将全部样本数据集划分为5个子数据集，在每个数据集中按6：2：1划分训练集，验证集以及测试集，为每个样本数据进行符合GO层级结构的完全标注，数据集信息如下所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | 属性值数量 | 训练样本数量 | 验证样本数量 | 测试样本数量 | GO标签数量 |
| Dataset1 | 50 | 537 | 179 | 90 | 204 |
| Dataset2 | 50 | 537 | 179 | 90 | 204 |
| Dataset3 | 50 | 537 | 179 | 90 | 204 |
| Dataset4 | 50 | 537 | 179 | 90 | 204 |
| Dataset5 | 50 | 538 | 179 | 90 | 204 |

### SVM分类

实验中采用的基础分类器为SVM分类器，Support vector machine(SVM)具有坚实的统计学理论基础，它可以很好地应用于高维数据，避免维灾难问题，。SVM分类器既可以用于线性可分的数据，又可以通过映射扩展到非线性可分的数据上。SVM学习问题可以表示为凸优化问题，因此可以利用已知的有效算法发现目标函数的全局最小值。而其他的分类方法（如基于规则的分类器和人工神经网络）都采用一种基于贪心算法的策略来搜索假设空间，这种方法一般只能获得局部最优解。

SVM分类器模型如图所示。



线性SVM寻找具有最大边缘的超平面，如果与直线l平行的两直线间距离，是能区分开正类和负类的一组平行直线间的最大距离，那么直线l为最大边缘超平面。具有较大边缘的决策边界比具有较小边缘的边界具有更好的泛化误差，如果边缘较小，决策边界的轻微扰动都可能对分类产生较大影响。支持向量机根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力（或称泛化能力）。一个线性分类器的决策边界可以写成如下形式：

如果标记所有的正类的类标号为+1，所有负类的类标号为-1，则可以用以下方式预测任意测试样本的类标号y：

线性SVM的学习任务可以形式化地描述为以下被约束的优化问题

受限于1,i=1,2,…,N

对于非线性决策边界数据集，将数据从原来的坐标空间x变换到一个新的坐标空间(x)中，从而可以在变换后的坐标空间中使用一个线性的决策边界来划分样本。

非线性SVM的学习任务可以形式化地描述为以下被约束的优化问题

受限于1,i=1,2,…,N

同线性SVM相比，主要区别在，学习任务是在变换后的属性，而不是在原属性x上执行的。经拉格朗日函数变换后，为避免空间中向量对之间的点积计算导致的维灾难问题，引入核函数

变换后的空间中的点积就可以用原空间中的相似度函数函数表示，这样，非线性SVM问题在使用核函数后就转化为了线性分类问题，便于SVM进行分类操作。

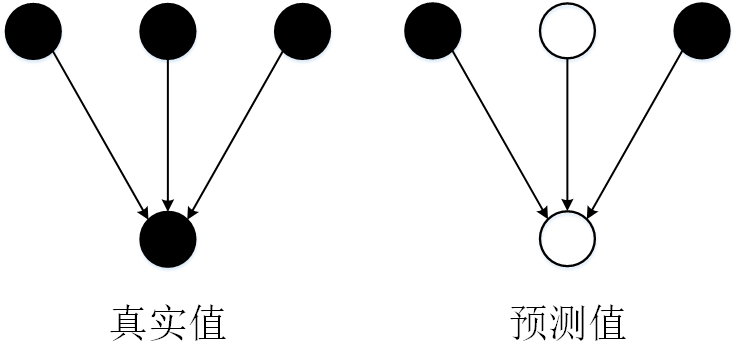
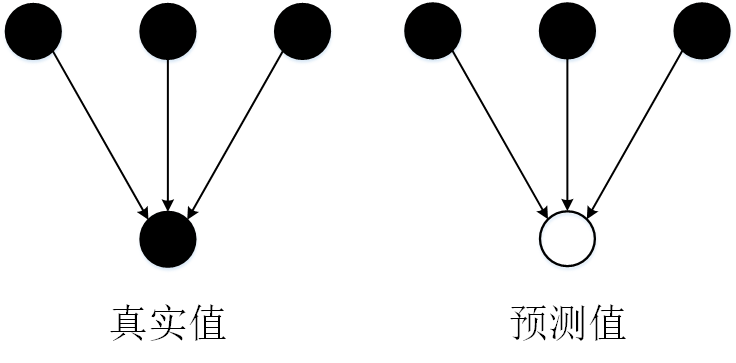
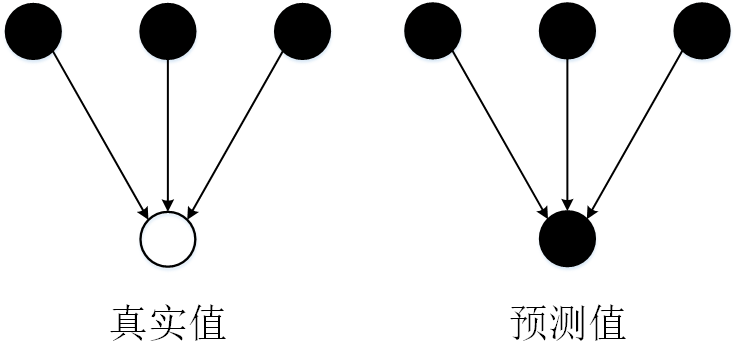
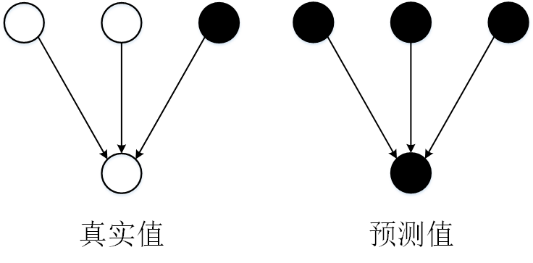
在实验中，支持向量机SVM使用LIBSVM软件包实现。LIBSVM是台湾大学林智仁教授开发的一个实现SVM算法的软件包，该软件运算速度快，提供源代码，便于研究者进行扩展和修改，并且可以集成到多个平台下。作者为SVM提供了很多默认的参数，这使得SVM的使用和调整参数的过程变得简单，因此是目前国内实现SVM比较流行的方案。实验中的核函数选择了RBF(Radial Basis Function)径向基函数，分类遵循如下步骤：转换数据至SVM格式，执行数据归一化，使用径向基核函数，用交叉验证找到最优参数C和，最后使用最优的参数进行整个数据集的训练，并进行测试。

## 层级约束实现

### 损失函数模型构建

将真实标签为y的样本x分为标签的损失，由损失函数给出，其中。此时，将样本采取决策，分为标签的条件风险R(为：R(=。根据基于最小风险的贝叶斯决策原理，对于样本x，使条件风险最小的标签即为样本x的预测标签，因此对样本x，标签的预测问题转换成为优化问题：。

在符合层级约束条件的情况下，在DAG层级结构中可能会发生的预测错误类型分为如下四种：

由此，层级多标签分类问题的损失函数可由四部分组成，DAGH 损失函数的具体公式为：

其中，均为权值常数，表示不同错误在损失函数中所占的权重。是节点 i 的误分类代价：

对于一个样本 x ，其采用 DAGH 损失函数的条件风险的具体形式为：

于是，采用 DAGH 损失函数，标签预测问题可以转化为如下优化问题：

### 损失函数优化问题

求取风险函数最小化的问题等价于如下优化问题：

其中函数定义为：

定义节点函数，对于某节点i有

其中

在函数 中，当节点i的子节点集合为空集时，该函数值为0；即当 child(i) =∅ 时，有 = 0。函数 的定义不包括根结点。

在引入了节点函数 的概念后，可以用 进行表示

对于一个有向无环图结构的层级多标签分类问题，当采用 DAGH 损失函数时，根据贝叶斯决策理论，对样本的分类问题可转化为如下优化问题：

此公式即为有向无环图结构的层级多标签分类问题的数学模型。当对公式描述的这一优化问题进行求解时，只需求得样本在各节点的后验概率即可。

### DAGlabel介绍

以上损失函数的构建需满足层级约束要求这一先决条件，即不会出现祖先节点预测值为0，而子孙节点预测值为1的情况。而在多个节点独立训练的过程中，此现象时有发生。为解决违反层级约束关系的问题，提出层级约束算法DAGlabel，来处理矩阵。算法思路如下：

将每个标签节点的值从大到小排序，然后选取值最大的节点，记为，查找的全部父节点，记录其中值小于0的个数，记为N。如果N等于0，则将与其值最小的父节点合成为一个超级节点，超级节点的值为被合成节点的平均值；若N大于等于1，则将与其值小于0的所有父节点合成为一个超级节点，超级节点的值为被合成节点的平均值。

其中有一种情形需要单独考虑：如图4-4所示的特殊情况，当子节点为3，且将与父节点1合成超级节点时，此时如果直接合成，新节点就会和节点2构成环路。所以，在这种情形下，子节点3与父节点1父节点2一起合并为超级节点。

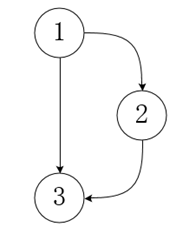


图 4-4 DAGlabel实例

直至最大的σ(i)值小于0，或所有节点均已经标记，DAGlabel算法结束。

该贪婪算法可以在不知道预测实例的最大标签数的情况下求出最优的分类结果，且分类结果满足层级约束要求。DAGLabel算法通过搜索层级结构中值最大节点，并将其与可能违反层级约束的父节点合成新节点，来逐步简化 DAG 并求出最优分类结果，通过将值小的父节点与子节点进行合并来提高父节点被选为正类或子节点选为负类的可能，经DAGlabel算法后的分类可满足层级约束条件。算法4-2中描述了 DAGLabel 算法的伪代码。

|  |
| --- |
| Input：  H：有向无环图层级结构；σ：待分类样本在各节点的σ值；  Output：  ：待分类样本的最终分类标签向量。  1:初始化变量，；初始化变量，。  2: **while TRUE do**  3: 找到 中具有最大值的节点  4: **if** **then**  5: **return**  6: **end if**  7: **if** 节点 的所有父节点均被标记 **then**  8: 将 存入 ，并将 从 中删除  9: **else**  10: 计算节点 的在集合 U 中父节点中 值小于零的个数 N  11: **if then**  12: 在 中找到 的 小于零的父节点集合  13: **if** 若存在节点集合 ϕ ，该集合中的节点既是 的父节点，  同时是 中节点的子节点 **then**  14: 将集合 ϕ 中节点与节点 和 中节点合成一个超级节点  15: **else**  16: 将节点 和集合 中节点合成一个超级节点  17: **end if**  18: **else**  19: 在 中找到 的具有最小 值的父节点  20: 将节点 和节点 合成一个超级节点  21: **end if**  22: 计算进行合并的各节点 值的平均值作为超级节点 的 值  23: 将超级节点 存入 中  24: **end if**  25: **end while** |

## 实验结果

### DAGlabel实验结果

将从SVM分类器中直接预测的结果与加入DAGlabel之后的结果进行对比，由于本分类问题面向层级多标签分类问题，故采用经典评价指标在层级分类领域的推广，微平均F值与宏平均F值作为评价指标。是第i个样本最详细的预测类别及所有祖先类别构成的集合，是第i个样本最详细的真实类别及其所有祖先类别构成的集合。下面给出微平均值和宏平均值指标的定义。

设一个数据集共包含m个样本，这些标签之间符合预定的层级结构关系，微平均形式下的精准率、召回率和值计算公式如下：

宏平均形式下的精准率、召回率和值的计算公式如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Macro.hm | | Micro.hm | |
| original | HMC-DAG-SVM | original | HMC-DAG-SVM |
| Dataset1 | 0.932334 | 0.929143 | 0.935129 | 0.931499 |
| Dataset2 | 0.91604 | 0.91556 | 0.919785 | 0.919566 |
| Dataset3 | 0.948706 | 0.949173 | 0.944818 | 0.945036 |
| Dataset4 | 0.927479 | 0.926701 | 0.932607 | 0.932605 |
| Dataset5 | 0.954988 | 0.955511 | 0.953178 | 0.953047 |

对比结果发现，HMC-DAG-SVM算法在微平均F值与宏平均F值方面均略逊于不加DAGlabel的结果，然而从层级多标签分类的实际意义出发，加入DAGlabel使得层级分类满足约束要求，避免了高层次的节点分类结果与其子节点分类结果发生冲突造成的分类无意义情况的出现。

### 对比实验

目前适用于有向无环图结构的分类算法较少，通过整理相关文献，将DAGlabel算法与目前较为流行的TPR，CLUS-HMC，R-SVM三种算法进行对比，下面简要介绍这三种算法。

TPR算法被称为真路径规则，整个算法过程整体上可分为两步，自底向上(bottom-up step)自顶向下(top-down step)，具体做法为，首先从下至上遍历整个层级结构，将下层节点的正类预测值传递给上层节点，使其对上层节点的判定产生影响。此过程结束后，再自根节点自上而下访问层级结构，将预测结果仍为负类的上层节点结果传递给相关的下层节点。

CLUS-HMC算法是一种全局算法，方法对层级结构的不同标签设置不同权重，生成归纳决策树一次性对全部标签进行分类，用加权的欧氏距离作为度量，并采用交叉验证的方式确定所需的层级参数。 FTest是停止标准，只有当FTest在某一级子集内部方差显著减少时，节点才会被分割。CLUS-HMC方法可以设置一组FTest并进行优化，在这种情况下，将选择最小的FTest以最小化所提供验证集上的RMSE度量。

HR－SVM算法是基于阈值可调支持向量机的局部分类方法，它在分类过程中考虑了不平衡数据集的影响，即节点越远离根节点，负类样本越多，正类样本越少，越不易被标记为正，R-SVM采用了潜在最佳阈值选择（potential best threshold selection）的策略，选择一组最重要的SVM阈值，然后采用最佳阈值估计（best threshold estimation）的方法计算出最佳阈值,应用到SVM上，它可以提升SVM对不平衡数据集的处理能力，在得到分类结果后，应用TPR算法对整体进行修正，保证分类结果满足层级约束要求。

### 算法评估

#####这里分析为什么DAGlabel效果不好，1/0情况，极端，越分越好，不好。

上述结果可以看出，DAGlabel算法的分类效果整体上比CLUS-HMC以及HR-SVM好，但仍略逊于TPR算法，考虑到TPR算法属于较为简单的层级约束，DAGlabel算法仍无法获得结果上的优势，在此对DAGlabel算法与TPR算法进行对比，并从结果出发，在原理上说明DAGlabel的局限性。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset  index | tochangelabels | DAG  change | TPR  change | DAG  work | TRP  work | DAG  right | TPR  right |
| 1 | 144 | 237 | 52 | 144 | 144 | 43 | 38 |
| 2 | 21 | 143 | 11 | 21 | 21 | 12 | 18 |
| 3 | 44 | 145 | 27 | 44 | 44 | 18 | 35 |
| 4 | 126 | 170 | 38 | 126 | 126 | 58 | 61 |
| 5 | 11 | 90 | 5 | 11 | 11 | 9 | 11 |

# 后期拟完成的研究工作及进度安排

## 后期拟完成的研究工作

根据已构建好的层级多标签分类模型，优化DAGlabel算法，使其在数据集上有更好的分类效果，并在原理上拥有可解释性。

## 进度安排

# 存在的困难与问题

# 如期完成全部论文工作的可能性

目前本课题进度与此前安排基本一致。剩余的工作主要是优化算法，尝试原理上给出解释，分析总结实验结果；撰写毕业论文。按照目前的进度，可以在剩余的时间完成毕业设计相关工作。