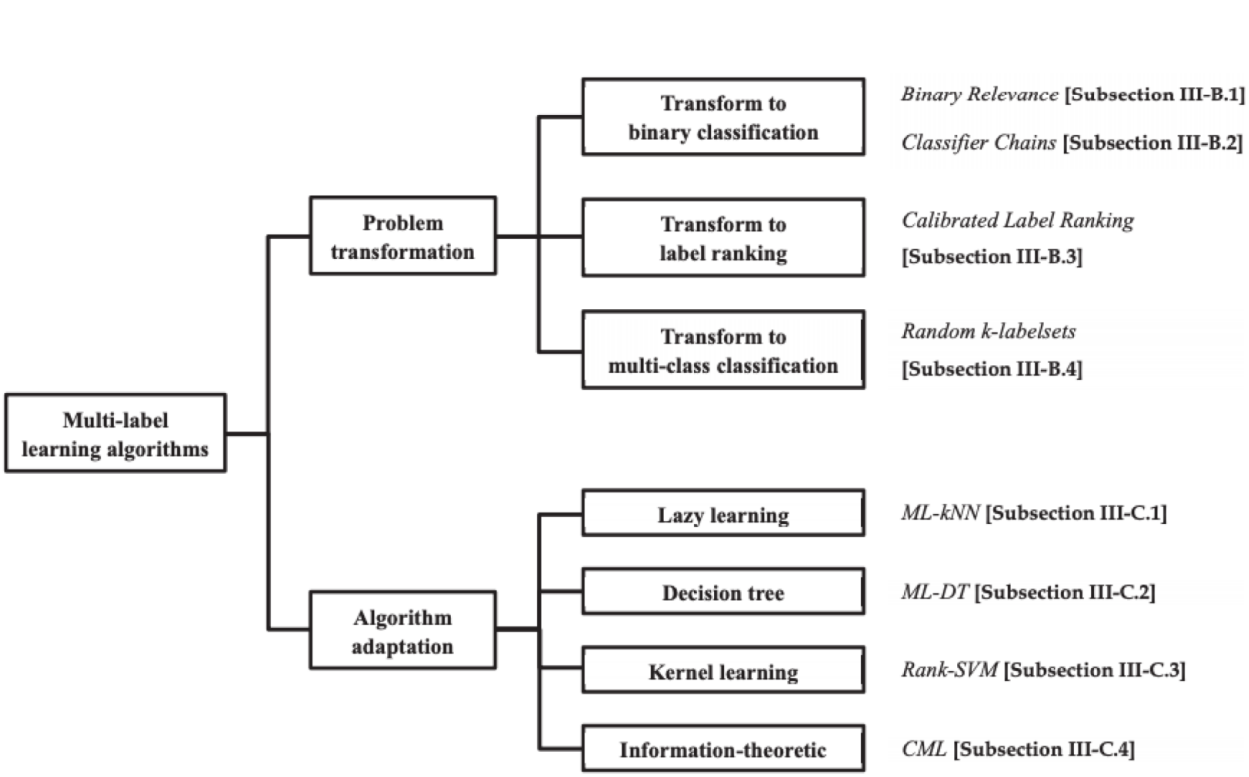
多标签/多分类学习

1 概念定义

多标签问题：各个标签是不互斥的（eg：电影标签是喜剧、动作片、国语片）

多分类问题：各个类别是互斥的（eg：学历是小学或初中或高中或大学），可知，多分类问题是包含于多标签问题中。

2 整体的逻辑框架如下：



参考：A review on multi-label learning algorithms --2014 TKDE(IEEE Transactions on Knowledge and Data Engineering)

https://www.researchgate.net/publication/263813673_A_Review_On_Multi-Label_Learning_Algorithms?enrichId=rgreq-945f5f0a8f6325d0f93a02d440de1366-XXX&enrichSource=Y292ZXJQYWdlOzI2MzgxMzY3MztBUzozMDE4MzQzMDEzOTQ5NTFAMTQ0ODk3NDMxMzMyMw%3D%3D&el=1_x_3&_esc=publicationCoverPdf

3 思路

无论多标签还是多分类，算法的思路都为可分为三类：

3.1 First-order strategy（一级策略）：

方法：one-vs-rest ：忽略和其它标签的相关性，比如把多标签分解成多个独立的二分类问题

优点：简单、高效。

缺点：忽略标签间相关性，结果可能不好

3.2 Second-order strategy（二级策略）：

方法：one-vs-one ：考虑标签之间的成对关联。例如：1、相关标签对（between relevant label and irrelevant label）；2、任何标签对

优点：考虑标签对相关性，泛化能力好。

缺点：只能表示标签对的相关性（one-vs-one）

3.3 High-order strategy（高级策略）：

方法：multi-vs-multi：考虑多个标签之间的关联，比如对每个标签考虑所有其它标签的影响

优点：模型考虑标签间关联

缺点：计算量大，不可扩展

4 多标签/多分类通用解决办法：

4.1 从问题/机制角度解决（problem transformation)：

4.1.1 二分类方法：

1. 多个二分类
2. 分类链

在这种情况下，第一个分类器只在输入数据上进行训练，然后每个分类器都在输入空间和链上的所有之前的分类器上进行训练。

X	y1
x1	0
x2	1
x3	0

Classifier 1

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

Classifier 2

X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

Classifier 3

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

Classifier 4

优点：考虑多个标签之间的关联，三级策略
缺点：算法好坏受链的顺序影响，可以通过随机序列解决；缺失了平行计算，因为需要链式调用

4.1.2 排序方法：

4.1.2.1 Calibrated Label Ranking

算法的基本思想是把多标签学习问题转为标签排序问题;
标签两两对比，得到一个排序list
优点：解决类间不均衡问题，考虑两两标签的关系，二级策略
缺点：复杂度高，分类器较多

4.1.3多分类方法：

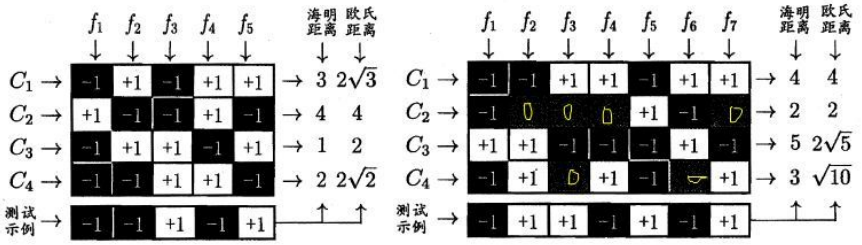
4.1.3.1 LP算法（label powerest）

根据输出的自然数映射回标签集
把 2q2^q2q ，个可能的标签集，映射成 2q2^q2q 个自然数。
缺点：泛化能力低，只能解决见过的类别；类别太大，低效

4.1.3.2 Random k-Labelsets 算法

随机划分长度为k的子集，收缩样本空间，构造n个子集，n个训练器做集成（投票方式）
优点：三级策略

4.1.3.3 纠错输出码（Error Correcting Output Codes, ECOC）



优点：三级策略

缺点：模型过于复杂，分类器过多

4.3 从改进算法角度解决(algorithm adaptation)：

4.3.1 Multi-Label k-Nearest Neighbor（ML-KNN）：

优点：简单，快；通过对K选择可具备丢噪音数据的健壮性
缺点：储存开销大，样本不均影响过大；一级策略

4.3.2 Multi-Label Decision Tree（ML-DT）

首先计算每个特征的信息增益IG，挑选IG最大的特征来划分样本为左右子集，递归下去，直到满足停止条件（例如叶子节点中子集样本数量为100）结束，对未知样本，沿根节点遍历一条路径到叶子节点，计算叶子节点样本子集中每个标签为0和1的概率，概率超过0.5的标签定为未知样本标签。
缺点：一级策略

4.3.3 Ranking Support Vector Machine（Rank-SVM）

优点：定义了”相关-不相关“标签对的超平面，考虑的两两标签关系，二阶策略

5. 可直接调用的多分类/多标签分类器：

参考：https://sklearn.apachecn.org/docs/0.21.3/13.html

5.1 只适用多分类问题：

- **4.1.1 1对1的多类分类器（从问题角度）（二级策略）：**
- `sklearn.svm.NuSVC`
- `sklearn.svm.SVC.`
- `sklearn.gaussian_process.GaussianProcessClassifier` (setting multi_class = “one_vs_one”)

- **4.1.2 1对多的多类分类器（从问题角度）（baseline）（一级策略）：**
- `sklearn.ensemble.GradientBoostingClassifier`

- sklearn.gaussian_process.GaussianProcessClassifier (setting multi_class = “one_vs_rest”)
- sklearn.svm.LinearSVC (setting multi_class=”ovr”)
- sklearn.linear_model.LogisticRegression (setting multi_class=”ovr”)
- sklearn.linear_model.LogisticRegressionCV (setting multi_class=”ovr”)
- sklearn.linear_model.SGDClassifier
- sklearn.linear_model.Perceptron
- sklearn.linear_model.PassiveAggressiveClassifier
- XGBOOST

5.2 多标签/多分类问题通用：

- **多标签/多分类通用的分类器（改进算法角度）：**
- sklearn.tree.DecisionTreeClassifier
- sklearn.tree.ExtraTreeClassifier
- sklearn.ensemble.ExtraTreesClassifier
- sklearn.neighbors.KNeighborsClassifier
- sklearn.neural_network.MLPClassifier
- sklearn.neighbors.RadiusNeighborsClassifier
- sklearn.ensemble.RandomForestClassifier
- sklearn.linear_model.RidgeClassifierCV