多标签学习和分类浅谈--从语言通信的角度看

鲁晨光

1. 序言

作者以前研究语义信息论，最近研究了多标签分类基本方法，觉得，从语义通信的角度看， 结合已有的方法，应该能得到更简单更合理的方法。下面是最近研究的一个总结。

数学定义：

- 实例 instance——$x \in U=\{x1,x2,\ldots,xm\}$
- 标签 label——$y \in V=\{y1,y2,\ldots,yn\}$， 一个标签可能是由多个原子标签组成复合标签。
- 样本 $D=\{(x(t),y(t)\}, t=1,2,\ldots,N.\ x(t) \in U, y(t) \in V$， 样本分布 P(x,y)， 实例分布 P(x).
- 条件样本 Dj=是 D 中 y(t)=yj 的部分。相应的条件样本分布是 P(x|yj)。

本文使用交叉熵方法，即用样本分布代替样本序列。

参考文献：

- Wikipedia 的多标签学习介绍：https://en.wikipedia.org/wiki/Multi-label_classification
- 张敏灵和周志华的多标签学习文章：A Review on Multi-Label Learning Algorithms http://cse.seu.edu.cn/people/zhangml/files/TKDE%2713.pdf
- 张敏灵等人关于二元关联的文章：http://cse.seu.edu.cn/people/zhangml/files/FCS%2717.pdf
- 徐兆桂的一片中文介绍 http://lamda.nju.edu.cn/huangsj/dm11/files/xuzg.pdf
- 本文作者的有关研究：http://survivor99.com/lcg/CM/Recent.html 其中有《从贝叶斯推理到逻辑贝叶斯推理》后面简称《逻辑贝叶斯推理》。

2. 多标签学习和分类的区别和联系

假设一个狼孩回到人类社会，他要通过自然语言了解标签“男人”,“女人”,“小孩”,“成年人”,“年经人”,“中年人”,“老人”…语言含义和用法。了解含义就是得到标签的外延。掌握用法就能传递消息给别人。这是两个任务。可能发生在两个人身上：收信人（看标签者）和发信人（贴标签者）； 也可能一个人身兼二值。

流行的分类研究都假设一个人身兼二职， 自己学， 自己用（参看图 1）。所以，学习就是掌握分类器，分类就是使用分类器。学习是复杂的， 使用是简单的。所以，张敏灵和周志华的文章都是讲多标签学习， 学习中包含优化分类器。Wikipedia 中是这样定义的多标签分类的：Formally, multi-label classification is the problem of finding a model that maps inputs **x** to binary vectors **y** (assigning a value of 0 or 1 for each element (label) in **y**). 这里没有标签外延或语义学习，直接就是分类。

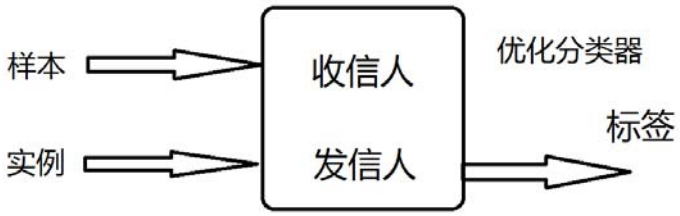

图 1 流行的学习方法中发信人收信人是同一人

但是从语言通信的角度看，我们要把发信人和收信人分开且，并且并列考虑（而不是一个人充当二职）。那么学习就是收信人获得标签外延或隶属函数，并不分类（并不划分实例空间），分类就是发信人划分实例空间，给每个实例或每个类别贴上最大内涵标签（通常是最小外延复合标签）。比如，狼孩学习“小孩”，“年轻人”，“大人”，“老人”，…的外延（几条隶属函数曲线），并不划分。70 岁的人同时被理解为“大人”和“老人”。但是，发信人要做划分，并且给实例选择一个外延较小的一个标签，比如“老年人”。不会说了“老年人”还要说他是“成年人”。

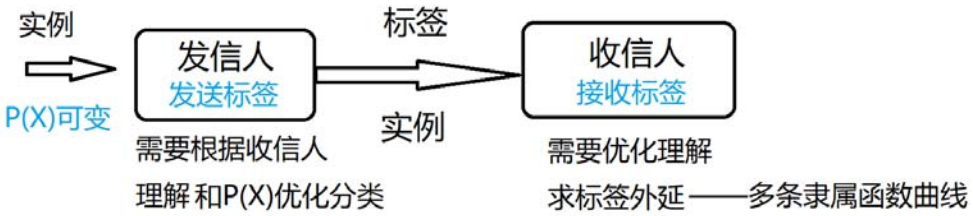

图 2 从语义通信的角度看，发信人和收信人并列，构成一个通信系统

分类贴标签时，对于可见实例，外延越小越好。但是对于不可见实例，可能预测不准，就有可能使用较大外延的标签，比如“是大人”或外延更模糊的标签“可能是老人”。天气预报有一般报“晴”，“小雨”或“中雨”…。如果测不准，就用更模糊标签“小到中雨”。外延小，一般信息量大，但是太小就容易出错，信息损失更大。所以折中考虑，就会使用比较模糊类别。对于可见实例分类，后面我们还要把外延最小准则改进为逻辑概率最小准则，这样就能和 Popper 理论即 Bar-Hillel-Carnap 定义的语义信息兼容。

要求学习和分类分开还有是一个原因：如果类别模糊，划分类别并不完全根据外延，还要根据实例的概率分布 P(X)。比方说：“年轻人”的外延可以用隶属函数表示，从 20-25 岁隶属度是 1，偏差越大，隶属度越小，50 岁大概是 0，划分界限要看 P(x)，你到部队去，30 岁以上肯定不能算年轻人；但是你要参加政协会议，40 岁甚至 50 岁还算是年轻人。为什么这样？语言是用来传递信息的，分类太少，信息少；分类太多成本又太高。所以传递年龄信息的标签大概“幼年”，“少年”，“青年”，“中年”，“老年”。粗分就是“小孩”和“大人”。人类寿命长了，分界点会自然改变，为的是传递更多信息。而信息准则更重视小概率事件——减少小概率事件的漏报。用法改变样本，样本改变外延（语义）；新的用法又会迁就已经形成的外延。两者相互促进。当然，给狮子老虎等动物分类，外延不是模糊的，不存在上述问题。天气预报语句给降雨量分类，用优良中差给学生考试分数分类，都需要根据 P(x) 改变分类边界。

3. 收信人如何求标签或概念外延？——多标签学习问题

假设我们知道普通人群不同年龄 x 的先验概率分布 $P(x)$（连续的），又知道成年人年龄概率分布 $P(x|y_1$ 是真的) (也是连续的，y_1 等于词或标签“成年人”)。 我们是否能得到合适的“成年人”的概念外延？换为不同人群(比如在学校，工厂， 部队)，其先验概率分布是 $P'(x)$, 我们能求出相应的后验分布 $P'(x|y_1$ 是真的)吗？ 我们能写出一个通用预测模型，用来求解这样的问题吗？

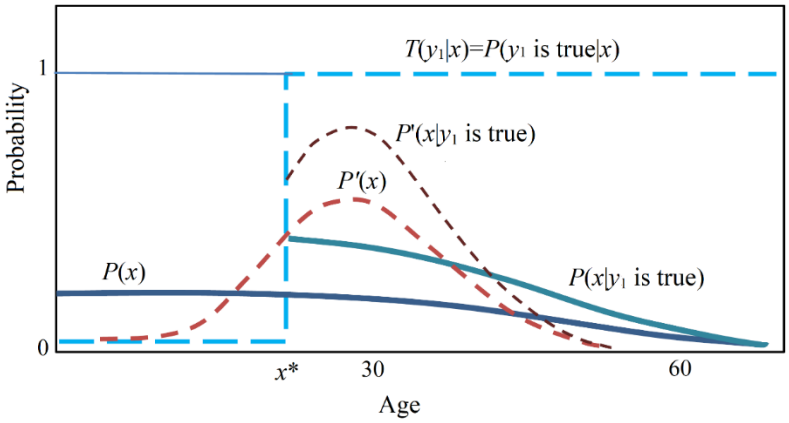

图 3 求 y_1=“成年人”的外延和后验分布 $P'(x|y_1$ is true).

已有的似然方法不能解决这个问题，因为似然函数不能加进先验知识，从一个人群得到的似然函数， 换一个人群就会失效。贝叶斯主义推理即 Bayesian Inference (BI)声称使用主观概率和逻辑概率，也考虑先验知识。用它可以吗？也不行！因为虽然 BI 用到贝叶斯定理（Bayes' theorem），但是实际上没有用到逻辑概率，因为逻辑概率不是归一化的(别处详谈)，而 BI 用到的先验 $P(\theta)$ (θ 是模型参数) 和后验 $P(\theta|x)$都是归一化的；2)BI 的先验是 θ 或 y(词或标签)的先验，它是主观的；而不是 x 的先验——它比较客观。我们上面例子是 x 的先验分布 P(x)变了，求相应的后验分布。

但是人脑求上面问题还是勉强可以的。 比如，我们划出先验和后验分布 P(x)和 P(x|y1 是真的)，就能知道“成年人”的外延——x 从哪里 x*开始就是成年人。然后我们把 P(x)中大于 x*的部分做归一化处理就行了。

但是类别模糊时， 比如没有法定“成年人”年龄规定的时候，如何计算“成年人”的外延，即类别{成年人}的隶属函数，或标签“成年人”的真值函数？人脑也不容易。

我发现了第三种贝叶斯定理， 用它可以解决上述外延问题， 有了外延， 还可以在 P(x)改变的情况下做新的预测。求外延公式是

$$T(\theta_j \mid x)=[P(x \mid \theta_j) / P(x)] / \max[P(x \mid \theta_j) / P(x)] \qquad (1)$$

其中 θj 表示一个模糊集合或预测模型，P(x|θj)=P(x|yj 是真的)。T(θj|x)是标签 yj 的真值函数或 θj 的隶属函数。做新的概率预测公式是：

$$P(x \mid \theta_j) = P(x)T(\theta_j \mid x) / \sum_i P(x_i)T(\theta_j \mid x_i) \qquad (2)$$

其中 P(x)是可以变的，也就是说， 隶属函数作为预测模型， 可以用于不同先验分布 P(x)的样本。我称上面两个公式为第三种贝叶斯定理。第一种是贝叶斯使用的，第二种是 Shannon 使用的。 现在我提供第三种。其证明详见：http://survivor99.com/lcg/CM/Homepage-

NewFrame.pdf （P.12）. 当样本足够大时， 求外延公式就变为：

$$T^*(\theta_j|X)= [P(X|y_j)/P(X)]/\max[P(X|y_j)/P(X)]= P(y_j|X)/\max[P(y_j|X)], j=1, 2, \ldots, n \quad （3）$$

它和转移概率函数称正比， 最大值是 1——就这么简单！

如果样本足够大， 能得到连续的分布 P(x,y)或 P(yj|X)(j=1,2,…,n), 直接用上面公式得到隶属函数。 如果样本不够大， 则用广义 Kullback-Leibler 公式优化隶属函数：

$$\begin{aligned} T^*(\theta_j \mid X) &= \arg\max_{T(\theta_j|X)} I(X;\theta_j) = \arg\max_{T(\theta_j|X)} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \\ &= \arg\max_{T(\theta_j|X)} [\log(1/T(\theta_j)) - \text{误差项}] \end{aligned} \quad （4）$$

T(θj)就是（2）中的分母，表示标签 yj 的逻辑概率。第一项就是 Bar-Hillel-Carnap 的信息，第二项反映误差。这是一个特殊的正则化误差准则， 误差是惩罚项。隶属函数覆盖范围小，逻辑概率就小， 潜在信息就多，但是相对误差就大了。所以要走“中庸之道”。也不难证明，语义信息准则等价雨大最大似然准则。

有了标签外延， 求他们之间蕴含关系就容易了。 如果总有 T（θj|x）≥T（θk|x）, 则 yk 蕴含 yj.

4. 发信人如何选择标签？——分类问题

有了标签外延， 分类就简单了——用最大语义信息准则。y_j 提供关于 x_i 的语义信息(量)被定义为对数标准(normalized)似然度：

$$I(x_i;\theta_j) = \log \frac{P(x_i \mid \theta_j)}{P(x_i)} = \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \quad (5)$$

其中用到公式（2）。这个公式就能反映 Popper 的思想：(先验)逻辑概率越小，并能经得起检验(后验逻辑概率越大) ，信息量就越大; 永真句在逻辑上不能被证伪，因而不含有信息。

如果实例可见， 分类函数就是

$$y_j^* = f(X) = \arg\max_{y_j} \log[T(\theta_j \mid X)/T(\theta_j)]$$

当类别清晰时， 隶属函数是 0 或 1， 上述分类函数就变为最小逻辑概率准则——不光外延小， 还要包含的实例概率小。 外延小就要求尽可能选择复合句， 比如“男青年”且是“成年人”。逻辑概率小的例子比如：“百岁以上老人”比用“一岁婴儿”外延大， 但是逻辑概率小， 潜在信息更多。

这个分类准则考虑到先验概率 P(x)的变化， 因而能解决类别不平衡问题。我们以标签 y1=“老年人”为例说明类别划分随实例的先验分布 $P(X)$变化(参看图 4)。其中设 X 表示年龄，标签 y_1=“老年人”。假设

$$T(\theta_1 \mid X) = \frac{1}{1+e^{-0.2(X-75)}}, \quad P(X) = 1 - \frac{1}{1+e^{-0.15(X-c)}} \quad (6)$$

图 4 显示了有关函数和优化的划分点 x^*。表 1 显示划分点随 X 的先验分布 $P(X)$中参数 c 的变化而变化。

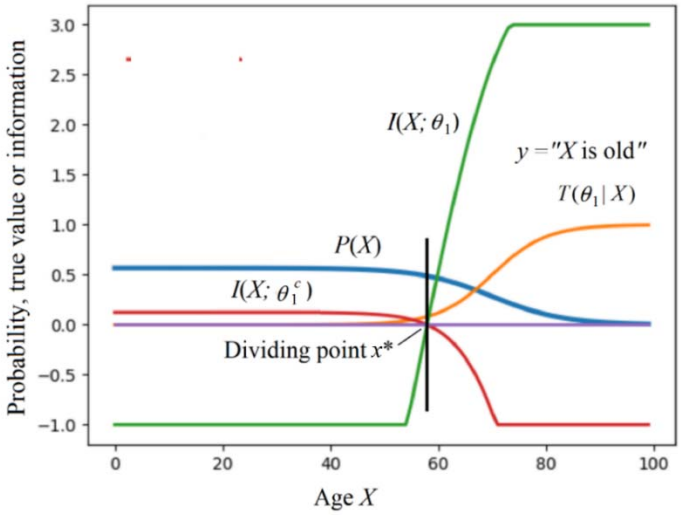

图 4 老年人分类划分点 x*. 它在老年人人口比例增加时右移。

表 1 "老年人" 分类器 $f(X)$ 随人口密度分布 $P(X)$中的参数 c 的变化

$P(X)$中参数c	人口密度下降区域	划分点(y_1=$f(X\geq x^*)$
50	40-60	49
60	50-70	55
70	60-80	58

注意：这时候"老年人"的外延并没有变化。划分点右移会影响新样本的 $P(x,y)$以及转移概率函数 $P(y_1|X)$，从而改变词的接收者理解的真值函数。这些变化也会反过来影响划分函数 $f(X)$。"老年人"的语义和使用规则就是这样进化的。没有人规定老年人的严格分界年龄在哪里，语言交流过程中会自动形成一个模糊分界，它随人口的年龄分布变化而变化。老年人多的群体，可能 70 岁才算老年人，50 岁算年轻人。以前人的寿命短，50 岁就算老年人了。"大雨"类似，雨水多的地区可能日降雨量 50mm 才算大雨；而雨水少的地区可能日降雨量 20mm 就算大雨了。

上述方法自然考虑了类别不平衡问题——因为信息准则更加重视减少小概率事件的漏报。用流行的方法， 分界点不随 P(x)改变（参看前面 wikipedia 的定义），类别不平衡问题是免不了的。

如果实例是不可见的(比如我们根据人的声音把人分成男女，或者根据西瓜的外观把西瓜分为好瓜和差瓜，我们只知道观察数据 $Z\in C$，则可用平均语义信息准则选择标签或划分观察数据空间 C 为 C1, C2,…， 即

$$y_j^*=h(Z)=\arg\max_{y_j}\sum_i P(x_i\mid Z\in C_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{7}$$

在预测不准时，模糊集合外延较大可以减小误差带来的信息损失， 所以用上式选出来的 y_j^*, 其逻辑概率 $T(\theta_j)$未必最小。这也是为什么天气预报经常报"小到中雨"。对于不可见实例分类，划分函数 $h(Z)$(确定每个 C_j)会改变 Shannon 信道和与之相匹配的语义信道，所以又要求重新划分… 因此需要迭代方法。参看鲁晨光的求最大似然估计的信道匹配算法 http://survivor99.com/lcg/CM/CM4tests.pdf 。

5. 流行的多标签学习方法存在的问题

流行的学习方法中用后验概率估计——信息论中称之为转移概率函数——$P(y_j|X;\theta)$ (j=1,2,…n)作为学习函数。因为归一化要求($\Sigma_j P(y_j|X;\theta)$ =1)，n>2 时， 很难构造一组转移概率函数。于是大家提出用多个二分类代替多分类，比如用 logistic 函数做二分类的转移概率函数。这也有多种。

一种方法是 One-vs-Rest，比如对于“年轻人”标签， 分成“年轻人”和“非年轻人”。然而，一个实例（24 岁）没有标注某个标签（“年轻人”），不等于它就是反例。比如一个样本中有两个例子： (25 岁, “年轻人”)和 (24 岁,“成年人”)。标注都是对的。按照 One-vs-Rest，对于“年轻人”，(24 岁,“成年人”)就是反例。这是不合理的。

另一种是二元关联（Binary relevance， 简称 BR）。这种方法用一个长度为 n 的二进制数字表示一个复合标签，1 表示“Yes”， 0 表示“No”。n 原子标签就有 2**n 种复合标签。这种方法不会出现上面问题， 但是又有新的问题：1）对样本要求太高， 从自然语言很难得到复合要求的样本。比如“非年轻人”,“非老年人”在自然语言中很少见。要把原来的 D 分成 n 个 Dj, 每次训练一个标签 yj，但是包括所有实例。2）存在标签相关性问题。比如，我们不会在说一个是“老年人”之外， 还说他是“成年人”或“非年轻人”。BR 方法提供太多标签， 不符合经济学要求，不符合自然语言习惯。

上面两种方法都很难解决类别不平衡问题。 因为标签学习的时候就已经分类了，选择标签的时候，尽管 P(x)不同， 也没法改变分类边界了。

本文前面方法可以做到：

1）使用隶属函数取代后验概率估计，就可以避免归一化难题；

2）训练一个标签的隶属函数的时候，把所有样例分为 3 种（正例，反例，不清楚例）。不考虑不清楚样例就可以了。不需要筛选样本。第三种贝叶斯定理就可以忽略不清楚样例。

3）因为学习的时候并不划分， 所以训练一个标签的隶属函数不一定要用反例（要用也可以， 参看英文文章）。因为 2）和 3）， 这种方法对样本没有特别要求， 能得到 P(x,y)或 P(y|x)就行。

4）求最小逻辑概率就已经考虑标签相关性，多余的标签自然就不要了。

5）始终使用和最大似然准则兼容的最大语义信息准则。但是也考虑到 P(x)变化后， 原来的似然函数失效——流行的方法都有这个问题。用隶属函数学习就能解决失效问题。

6. 关于优化准则和泛化能力

我们常常看到有人抱怨：用原来的样本正确率高，换一个样本就差了，泛化能力不行。原因在哪里？ 通常认为是因为过度拟合。 我以为更重要的原因是：1）P(x)会变， 用正确率准则，P(x)变了，总体正确率必然差了。2）我们需要根据新的 P(x)改变分类器。

先看正确率准则的问题。对于地震预测（或艾滋检测），永远报没有地震（没有病毒），准则率很高， 但是没有意义。如果总是根据今天有无雨预测明天有雨， 根据今天股市涨或跌预测股市明天涨或跌，正确率也比较高，但是也没啥意义。因为你不说别人也估计得差不多。好的预测要能提高相对正确率。最大似然准则和信息准则（包括语义信息准则）就是这样的准则。正则化误差平方（RLS）也是类似准则（参看《逻辑贝叶斯推理》）。

医学界还用似然比或置信水平（而不用正确率）评价检验本身的好坏。因为检验好坏在

于检验本身，在于信道，而和信源 P(x)无关。而正确率和 P(x)有关。机器学习也应该采用置信水平或确证度这样的评价（参看《逻辑贝叶斯推理》）。

关于正确率准则和信息准则的差别，打个比方。一个教师 A 重在提高全班学生学习总成绩（重视概率较大群体），另一个教师 B 重在提高每个学生的相对成绩。如果换一批学生，P(x)变了， 两人还用原来教学方案， A 教的班级的总成绩就可能下降很多。而 B 教的班级学习成绩会比较稳定。也就是说，B 的泛化能力强。要提高学生相对成绩，老师还需要根据 P(x)重新划分小组施教。多标签学习和分类是类似的，用信息准则可以提高隶属函数的泛化性能；根据 P(x)做多标签分类就是优化分组，以便提高平均“相对成绩”。

关于本文方法的进一步讨论见 《逻辑贝叶斯推理》 http://survivor99.com/lcg/CM/Recent.html

欢迎批评交流！lcguang@foxmail.com