符号机器学习

张文生 研究员 首席教授

中国科学院自动化研究所

中科院大学人工智能学院

2019年11月14日

符号机器学习

- 最早的符号机器学习是基于文法归纳的方法，其动机是为了处理语言信息，这个动机被证明在技术上是不可行
- 符号机器学习的研究主要聚焦在定义在符号域上的结构化数据的约简上

- **两个步骤：**

1. 将观测数据映射到一个符号空间，“词”的处理
2. 将定义在符号空间的观测约简(归纳)为更为简洁的表示，“语句”的处理

- **两种策略：**

1. 将两个步骤分开
2. 将第一个步骤嵌入在第二个步骤之中(例如，C4.5)

泛化能力

1. 对观测中的噪音，使用统计或其他方法将其映射到符号域上，从而过滤噪音
2. 把观测数据约简为简洁的规则形式，以提高规则概括问题世界的能力

符号机器学习的目标

- 符号机器学习只与规则的描述长度有关
- 对于符号机器学习，由于描述长度与泛化能力有关，因此，一般以描述长度最短为目标

符号机器学习有竞争力吗?

- 一方面，符号机器学习动态地将观测数据映射到符号域，因此，其本质与符号域中符号的概念化无关
- 另一方面，符号机器学习强调符号在实际世界中的对应，其泛化能力没有竞争力

符号机器学习的特点：可阅读

由于符号机器学习需要将观测数据映射到一个符号域上，因此，约简后的规则比较简洁，那么规则集合是可以被人阅读的

两类任务

- 第一类任务，与传统的基于统计的机器学习任务相同，强调泛化能力，需要与其他机器学习方法竞争
- 第二类任务，与传统的数据分析任务相同，机器学习的目标是解读数据，强调数据的可阅读性，这类学习方法需要与传统数据分析方法竞争

数据分析与机器学习的区别

符号机器学习	符号数据分析
假设所有用户需求：相同	假设所有用户需求：不同
评测标准：泛化能力	评测标准：约简能力
目标：“最小”	目标：需求意义下的相对最小
算法设计：近似“最小”(NP)	算法设计：精确解答

信息系统

- 给定一个属性集合A，数据集合U中的任一次观察是根据属性A获得，A称为一个信息系统
- 每次的观察称为一个对象，数据集合U称为对象集合

$A=\{a,b,c,e\}\cup\{d\}$
$U=\{1,2,3,4\}$

对象	a	b	c	e	d
1	0	0	0	0	0
2	1	0	1	1	0
3	0	1	1	0	1
4	0	0	1	0	1

其中 d 是决策属性。

信息系统的类型

- 如果属性集合A可以分为两个部分C与D，其中C是条件，D是分类的标记，则这个信息系统称为有决策的信息系统，记为<U, C∪D>；否则称为无决策信息系统，记为<U, A>
- 对<U, A>，加入一个满足条件Card(U)=Card(U/D)的伪决策属性D，则<U, A>可以化为<U, C∪D>的形式
- 总之，只需考虑<U, C∪D>

等价关系

- 任何符号机器学习需要事先定义一个等价关系

符号机器学习中的等价关系

- 由于符号机器学习处理定义在符号域上的数据，因此，一般可以使用最简单的等价关系，称为不可区分关系
- 不可区分关系的定义为：x等价于y 当且仅当 a(x)=a(y)
- 在息系统属性被考虑为等价关系

单个属性对对象集合的划分

- 令a是一个属性（等价关系），U是对象的集合
- 对象集U关于属性 a的划分为：$U/a=\{E_1, \ldots E_k\}$
  - 其中E_j满足条件： $E_i \neq \varnothing$,且当$i \neq j$, i, j=1,2,…,k $E_i \cap E_j=\varnothing$, $\cup E_j=U$
- E_j（j=1,2,…,k）称为对象集U关于属性 a的等价类。这些等价类也称为对象集U关于属性a的初等范畴

属性子集对对象集的划分

- 令$C=\{c_1, c_2, \ldots, c_k\}$是定义在对象集合U上的等价关系的集合
- $R \subseteq C$是C的一个子集
- U/R是U关于R的划分，$U/R=\{E_1, \ldots E_k\}$，$E_i \neq \varnothing$，当$i \neq j$, $i, j=1,2,\ldots,k$时，$E_i \cap E_j=\varnothing$, $\cup E_j=U$
- 等价类$E_1, \ldots E_k$称为R的基础范畴

基础范畴的计算

- 令R={a, b}， U/a={G_1, G_2, …, G_p}， U/b={H_1, H_2, …, H_q
- U/R={$G_j \cap H_i$ | $G_j \in U/a$， $H_i \in U/b$}
- 基础范畴为： $G_j \cap H_i$ ， $j \le p, i \le q$
- 对U的任意一个子集R，$U/R=\cap_{\forall a \in R}(U/a)$
  - U/R是R中所有属性关于U的划分的笛卡尔交集

积木世界的例子

- 积木世界：

$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$

- 属性集合：

{颜色(c)，形状(s)，体积(v)}

根据颜色 c

红	x1, x3, x7
蓝	x2, x4
黄	x5, x6, x8

根据形状 s

园	x1, x5
方	x2, x6
三角	x3, x4, x7, x8

根据体积 v

大	x2, x7, x8
小	x1, x3, x4, x5, x6

- $U/c=\{\{x_1, x_3, x_7\}, \{x_2, x_4\}, \{x_5, x_6, x_8\}\}$
- $U/s=\{\{x_1, x_5\}, \{x_2, x_6\}, \{x_3, x_4, x_7, x_8\}\}$

U/c或U/s中的等价类称为U关于c或s的**初等范畴**

- $U/\{c, s\}=\{\{x_1\}, \varnothing, \{x_3, x_7\}, \varnothing, \{x_2\}, \{x_4\}, \{x_5\},\{x_6\},\{x_8\}\}$

U/{c，s}中的等价类称为U关于{c,s}的**基础范畴**

符号机器学习的典型方案

- 以约简为基础的符号机器学习发生在上个世纪七十年代末到八十年代中
  - 最有代表性的研究为AQ11、ID3与Reduct理论
- 尽管这类研究还取得了重要的进展，主要集中在数据由离散量向连续量过渡

符号机器学习原理的表述

- 存在两类不同但等价的表述方法：

1. 以命题逻辑为基础的表述方法(AQ11)

2. 以等价关系为基础的表述方法(ID3与reduct理论)

命题

- **命题1** 令F是命题逻辑公式，$F \wedge \sim F = \square$ 且 $F \vee \sim F = \blacksquare$，其中 为永假式
- **解释：**

  如果两个命题逻辑公式构成合取互补对，逻辑公式永假

  如果两个命题逻辑公式构成析取互补对，逻辑公式永真
- **命题2** 令G是命题逻辑公式，$G \vee \square = G$ 且 $G \wedge \blacksquare = G$，其中$\blacksquare$为永真式
- **解释：**

  在逻辑公式G上析取一个永假式，逻辑公式的的真值与G相同

  在逻辑公式G上合取一个永真式，逻辑公式的的真值与G相同

属性-值的逻辑表示

- 一个属性-值可以表示为一个带等号的命题原子，$[a=v]_x$，读为对象x关于属性a的的值为v
- 如果另一个对象y的类别标记不同于x，且其属性a的值不是v，记为$[a\neq v]_y$
- 计算两个对象的逻辑公式，可以删除下标

- $[a=v] \wedge [a \neq v] = \square$ (永假)
- 对一个析取范式，上式是一个子句
- 这个子句可以从范式中删除，不影响范式的真值

逻辑演算方法的解释

- 假设x与y是两个有不同决策标号的对象
  - 其中一个对象的逻辑表示必须取“非”
- 如果两个对象关于属性a具有相同值，根据上述讨论，必然存在一个包含互补对的子句：

  $[a=v]\wedge\sim[a=v]=[a=v]\wedge[a\neq v]=\square$，从而这个子句可以删除
- 从逻辑公式中删除一个子句，不改变真值，说明这个子句对于决策没有提供任何信息

逻辑演算与等价类的关系

- 对U/a，关于属性a具有相同值的对象，一定在同一个等价类中
- 对U/a，被删除的子句一定是a的值相同的对象
- 这就是逻辑演算方法与等价关系划分之间的关系

关于AQ11一个例子——“两组人”

- 条件属性集合：{身材(s)，发色(h)，眼睛(e)}
- 条件属性的值域：
  - s：{高(1)，矮(0)}
  - h：{金(0)，黑(1)，红(2)}
  - e：{蓝(0)，黑(1)，灰(2)}
- 决策属性： d
- 决策属性值域：
  - d：{0, 1}

- 对上述例子，考虑决策属性为0的为正对象，其他为反对象

正对象	身材 s	发色 h	眼睛 e	D		反对象	身材 s	发色 h	眼睛 e	D
1	0	0	0	0		5	1	0	1	1
2	1	2	0	0		6	0	1	0	1
3	1	0	0	0		7	1	1	0	1
4	0	0	2	0		8	1	1	2	1
						9	0	0	1	1

将对象表示为合取式：

对象1表示为：$[s=0]\wedge[h=0]\wedge[e=0]$

对象5表示：$\sim([s=1]\wedge[h=0]\wedge[e=1])=[s\neq1]\vee[h\neq0]\vee[e\neq1]$

- 令g_k表示一个对象。
- 对象集合表示为：$(g_1 \wedge g_2 \wedge g_3 \wedge g_4) \wedge (\sim g_5 \wedge \sim g_6 \wedge \sim g_7 \wedge \sim g_8 \wedge \sim g_9)$
- 一个正对象表示为：$g_1 \wedge (\sim g_5 \wedge \sim g_6 \wedge \sim g_7 \wedge \sim g_8 \wedge \sim g_9)$

$$
\begin{aligned}
&= [s=0] \wedge [h=0] \wedge [e=0] \\
&\quad \wedge ([s \neq 1] \vee [h \neq 0] \vee [e \neq 1]) \\
&\quad \wedge ([s \neq 0] \vee [h \neq 1] \vee [e \neq 0]) \\
&\quad \wedge ([s \neq 1] \vee [h \neq 1] \vee [e \neq 0]) \\
&\quad \wedge ([s \neq 1] \vee [h \neq 1] \vee [e \neq 2]) \\
&\quad \wedge ([s \neq 0] \vee [h \neq 0] \vee [e \neq 1])
\end{aligned}
$$

- 学习结果为：

$[h\neq1]\wedge[e\neq1]=([h=0]\vee[h=2])\wedge([e=0]\vee[e=2])$

- 对象1的归纳规则：

如果[发色=金/红]∧[眼睛=蓝/灰]，属于第一类人

AQ11

- 给定信息系统，C是条件属性集合
  - 对象：$g=\wedge_{a\in C}[a=v]$
- 根据决策属性，将对象集合划分为两类
- 对象集合：$(g_1\wedge g_2\wedge\ldots\wedge g_n)\wedge(\sim q_1\wedge\sim q_2\wedge\ldots\sim q_m)$
- 计算一个正对象：$g_1\wedge(\sim q_1\wedge\sim q_2\wedge\ldots\wedge\sim q_m)$

$$=(g_1\wedge\sim q_1)\wedge(g_1\wedge\sim q_2)\wedge\ldots\wedge(g_1\wedge\sim q_m)$$

ID算法

- ID算法来源于CLS
- ID算法与CLS的区别是在算法设计中使用了窗口技术，以解决对象集合庞大的问题
- 在属性选择策略与树结构方面，则继承了CLS的考虑
- CLS可以归纳在ID算法

CLS算法

- 信息系统<U, C∪D>， a∈C， D={d}。 U/d={X_1,…,X_n}
- f(<U, C∪D>)是属性选择策略。根据f(<U, C∪D>)选择a作为决策树的根节点

(1) 对U，如果U⊆X_p，属性a构成决策树(a是叶节点)，否则

(2) U/a={E_1,…,E_k}，对所有E_j，分别令U=E_j，根据f(<U, C-{a}∪D>)选择b，重复(1)，建立以b为根节点的子树

f(<U, C∪D>)的定义

- 一般采用信息熵。令$U/a=\{E_1,\ldots,E_k\}$

$$P_a = -\sum[p_j \log(p_j) + (1-p_j)\log(1-p_j)]$$

其中： $p_j=Card(E_j)/Card(U)$， $1-p_j=Card(U-E_j)/Card(U)$

- **说明：**

1. 使用信息熵的动机：事物的信息熵，表示事物的有序程度
2. 在机器学习中使用信息熵是为了获得关于事物的有序表示

ID3算法

- 从假设空间中搜索一个拟合训练样例的假设
- 决策树采用自顶向下的贪婪搜索策略
- 树的构造从判定“哪一个属性做根节点最合适”开始
- 每一步，属性选择都以信息增益为的度量标准

感谢同学们听课

欢迎讨论与交流