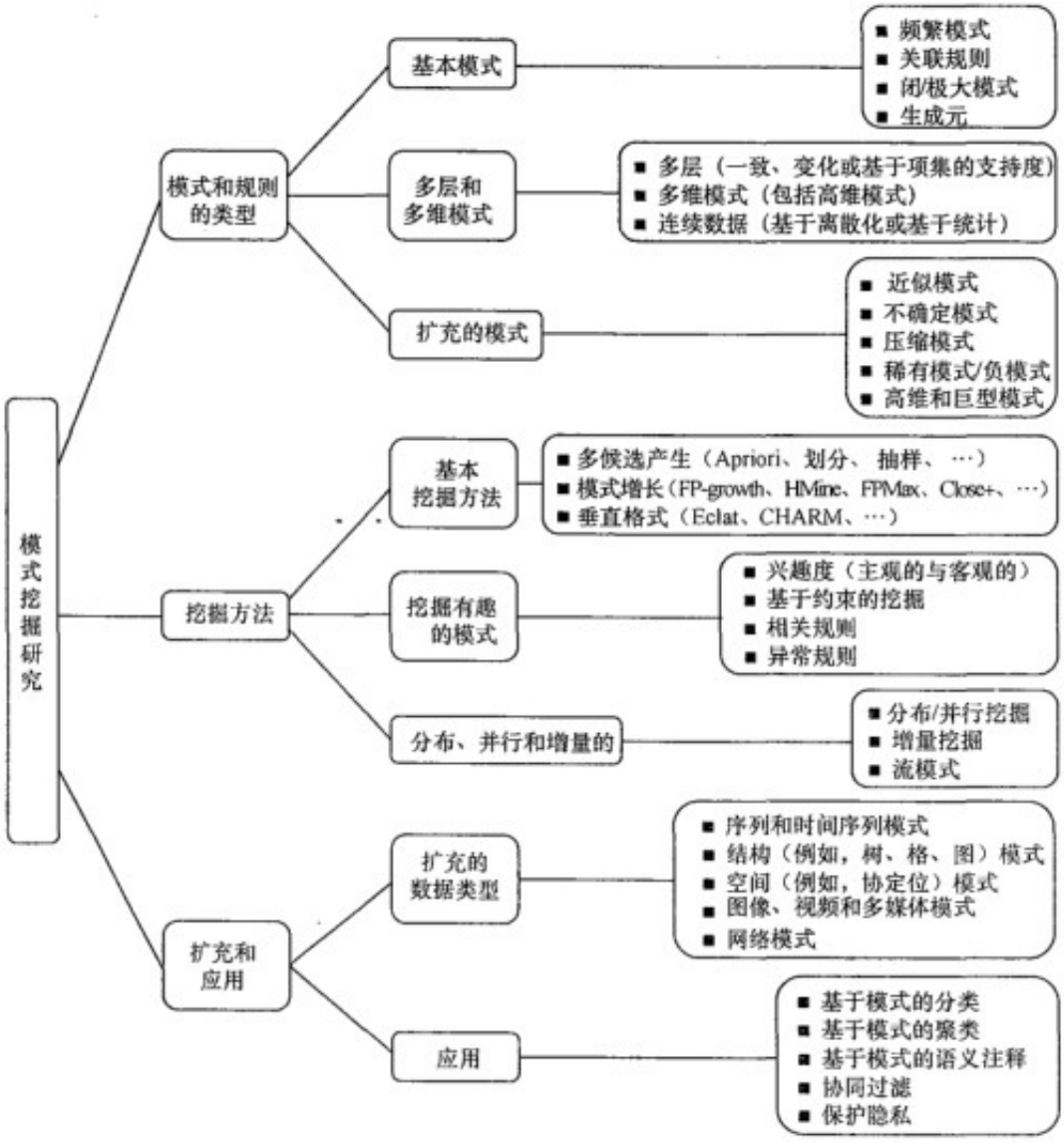
第七章 高级模式挖掘

笔记本： 数据挖掘：概念与技术

创建时间： 2017/12/20 15:44 **更新时间：** 2017/12/27 11:25

作者： Passero

7.1模式挖掘：一个路线图



7.2多层、多维空间中的模式挖掘

挖掘多层关联规则

在多个抽象层的数据上挖掘产生的关联规则称为多层关联规则。

[
---具有一致支持度的多层挖掘（一致支持度）
---具有递减支持度的多层挖掘（递减支持度）
---使用基于项或基于分组的最小支持度（基于分组的支持度）
]

挖掘多维关联规则

---规则中每个不同的谓词称作维。
---涉及单个维或谓词的关联规则称为单维或维内关联规则。
[e.g.
buys(X, "digital camera") => buys(X, "HP printer")
->因为包含单个不同谓词（例如，buys）的多次出现（即谓词在规则中出现的次数超过一次）。]
---涉及两个或多个维或谓词的关联规则称作多维关联规则。
[e.g.
age(X, "20...29") ∧ occupation(X, "student") => buys(X, "laptop")
->因为涉及了三个谓词，每个谓词在规则中仅出现一次。并且，我们称它具有不重复谓词。具有不重复谓词的关联规则称作维间关联规则。当包含某些谓词的多次出现时，这种规则称作混合维关联规则。如age(X, "20...29") ∧buys(X, "laptop") => buys(X, "HP printer")]

[review:
---标称属性：标称属性的值是事物的名称，标称属性具有有限多个可能值，值之间无序
---量化属性：量化属性是数值的，并在值之间具有一个隐序
]

挖掘多维关联规则的技术可以分为两种基本方法：
---使用预先定义的概念分层对量化属性离散化（这种方法叫做使用量化属性的静态离散化挖掘多维关联规则）
[e.g. 可以使用income的概念分层，用区间值，如"0..20k" "21..30k" "30..40k"等来替换属性原来的值]
---根据数据分布将量化属性离散化或聚类到箱（由这种方法挖掘的关联规则称为（动态）量化关联规则）

k-谓词集是包含k个合取谓词的集合。
e.g.谓词集{age，occupation，buys}是一个3-谓词集。

挖掘量化关联规则
（1）数据立方体方法
（2）基于聚类的方法
（3）揭示异常行为的统计学方法

挖掘稀有模式和负模式

definition：如果项集X和Y都是频繁的，但很少一起出现（sup(X∪Y) < sup(X) × sup(Y)），则项集X和Y是负相关的，并且模式X∪Y是负相关模式。如果sup(X∪Y) 《 sup(X) × sup(Y)，则X和Y是强负相关的，并且模式X∪Y是强负相关模式。
该定义可扩展到包括k-项集的模式，其中k>2。
attention: 这个定义的一个问题是，它不是零不变的，即它的值可能错误地被零事务影响，其中零事务是不包含被考察项集的任何项的事务。
该度量不是零不变的：

定义 7.2：如果 X 和 Y 是强负相关的，则

$$sup(X \cup \overline{Y}) \times sup(\overline{X} \cup Y) \gg sup(X \cup Y) \times sup(\overline{X} \cup \overline{Y})$$

该定义没有前两个定义的零不变问题：

定义 7.3：假设项集 X 和 Y 都是频繁的，即 $sup(X) \geqslant min\_sup$，$sup(Y) \geqslant min\_sup$，其中 min_sup 是最小支持度阈值。如果 $(P(X|Y)+P(Y|X))/2 < \varepsilon$，其中 ε 是负模式阈值，则 $X \cup Y$ 是**负相关模式**。

7.3基于约束的频繁模式挖掘

- **知识类型约束**：指定待挖掘的知识类型，如关联、相关、分类或聚类。
- **数据约束**：指定任务相关的数据集。
- **维/层约束**：指定挖掘中所使用的数据维（或属性）、抽象层，或概念分层结构的层次。
- **兴趣度约束**：指定规则兴趣度的统计度量阈值，如支持度、置信度和相关性。
- **规则约束**：指定要挖掘的规则形式或条件。这种约束可以用元规则（规则模板）表示，如可以出现在规则前件或后件中谓词的最大或最小个数，或属性、属性值和聚集之间的联系。

关联规则的元规则制导挖掘
一般而言，元规则形成一个关于用户感兴趣探查或证实的假定。然后，挖掘系统可以寻找与给定元规则相匹配的规则。

基于约束的模式产生：模式空间剪枝和数据空间剪枝

7.4挖掘高维数据和巨型模式

7.5挖掘压缩或近似模式

诵过模式聚类挖掘压缩模式

我们可以使用闭模式之间的距离度量。设 P_1 和 P_2 是两个闭模式，它们的支持事务集分别为 $T(P_1)$ 和 $T(P_2)$。P_1 和 P_2 的模式距离（pattern distance）$Pat\_Dist(P_1, P_2)$ 定义为

$$Pat\_Dist(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|} \quad (7.14)$$

模式距离是一种定义在事务集合上的有效距离度量（metric）。注意，正如我们所期望的，它包含了模式的支持度信息。

提取感知冗余的top-k模式

挖掘 top-k 个最频繁模式是一种减少挖掘返回的模式数量的策略。然而，在许多情况下，频繁模式不是相互独立的，而常常是集中在一些小区域内。这有点像在全世界找出 20 个居住中心，结果可能是集中在少数几个国家而不是均匀地分布在全球的城市。大部分用户更愿意得到 k 个最有趣的模式，它们不仅是显著的，而且是相互独立的，并且是很少有冗余的。不仅具有高显著性，而且具有低冗余的 k 个代表模式的小集合称为感知冗余的 top-k 模式（redundancy-aware top-k patterns）。

7.6模式探索与应用

频繁模式的语义注解

模式挖掘的应用