第34卷 第13期 计 算 机 工 程 2008年7月
Vol.34 No.13 Computer Engineering July 2008
·软件技术与数据库· 文章编号：1000—3428(2008)13—0049—02 文献标识码：A 中图分类号：TP311

朴素贝叶斯分类器一阶扩展的注记

徐光美，杨炳儒，钱 榕
(北京科技大学信息工程学院，北京 100083)

摘 要：众多研究者致力于将朴素贝叶斯方法与原有的ILP系统结合，形成各种各样的多关系朴素贝叶斯分类器(MRNBC)。该文提出形成朴素贝叶斯分类器的一阶扩展的一般方法。现实中关系数据库广泛存在，可以直接作用于数据库表，而无须转换表示形式的MRNBC则是研究的重点，该方法主要基于关系数据库理论，分析了进行一阶扩展的关键问题。

关键词：多关系数据挖掘；朴素贝叶斯；分类；归纳逻辑程序设计；关系数据库

Notes on First-order Upgrading to Naïve Bayesian Classifier

XU Guang-mei, YANG Bing-ru, QIAN Rong
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083)

【Abstract】 Many researchers focus on the combination of ILP system together with the naive Bayesian classifier. And various Multi-Relational Naïve Bayesian Classifiers(MRNBC) have been proposed. This paper describes a methodology based on ILP for upgrading naïve Bayesian classifiers to first-order logic. There are almost relational databases everywhere in real-life, so this paper also aims at solving how to shape MRNBC based on relational databases theory. The upgrading can deal with data in multiple tables directly and transformation is not needed.

【Key words】 Multi-Relational Data Mining(MRDM); Naïve Bayesian; classification; Inductive Logic Programming(ILP); relational database

1 概述

朴素贝叶斯分类器(Naïve Bayesian Classifier, NBC)是最简单的分类模型之一，它训练简单且易于理解。研究表明，尽管在实际中朴素贝叶斯假定并不成立，但NBC仍取得了良好的分类效果[1]。随着多关系数据挖掘的兴起，各种各样的多关系分类器被提出，解决多关系分类问题的研究方向主要有以下几种：(1)命题化方法，即使用平滑技术或者特征构建方法将多关系数据转化到单个关系中，然后使用传统的方法对其分类，但是大多数命题化方法是启发式的，因此，不完备的数据转换会丢失很多有用信息；(2)将现有的分类算法扩展使其可以处理多关系数据，如ILP(Inductive Logic Programming)方法：FOIL, Progol, Golem, TILDE；概率方法：PRMs, RBNs, SRMs；(3)通过模式识别或者关联规则挖掘分类关系数据的方法，如频繁模式挖掘方法[2]。

多关系朴素贝叶斯分类器(Multi-relational Naïve Bayesian Classifier, MRNBC)是将统计学习和关系学习结合处理分类问题的最简单的方法。近年来，已经提出了多种多关系朴素贝叶斯分类器，其中大多数方法基于ILP技术形成，部分研究则旨在将朴素贝叶斯与关系数据库理论结合形成MRNBC。本文在分析已有研究成果的基础上，探讨了扩展朴素贝叶斯分类器到一阶情况下的一般方法。

2 基于ILP技术的一阶扩展

基于ILP技术形成的关系学习器使用一阶逻辑的一个子集作为假设的描述语言，提升了假设的描述能力，因此，可以更好地解释数据。但是就分类而言，过去的工作却显示属性值学习器经常比关系学习器有更好的分类效果，即使关系学习器可以提供更丰富的背景知识。原因可能是属性值学习器面对更小的假设空间，可以更彻底地搜索空间，而且命题学习器可以考虑一些额外的信息，这些信息对训练数据集合上的假设进行了概率分析。

为使现有ILP系统可以进行有效的分类，很多研究致力于将其和朴素贝叶斯方法结合。已有研究成果有：基于ILP-R的朴素贝叶斯分类方法(简称为P-K方法)[3]、1BC方法[1]、1BC2方法[4]、Mr-SBC方法[5]和nFOIL方法[6]等。本文分析了这些不同方法之间的共性，形成了基于ILP技术建立MRNBC的一般方法。

(1)将数据库中的元组转化为Prolog事实。基于ILP技术形成的MRNBC，都是在驻留主存的Prolog事实基础上进行系统实施。现实中事实是关系数据库中的元组，所以需要一些预处理步骤转换数据形式：首先要把关系中的每一个元组都表示成知识库的形式，即描述事实的子句，Prolog把关系名自动转换为谓词，并把数据库的每一条记录都转换为不含有任何变量的子句。

(2)选择合适的标准ILP系统静态或动态的产生一阶规则或一阶条件(发现全局最优的模型结构)。P-K方法中使用ILP-R系统学习假设，假设由若干个子句构成，因此，假设可以看成是分类规则的集合；nFOIL选择FOIL作为一阶规则的学习器，不同之处是它将FOIL和朴素贝叶斯学习模式紧密结合，使用概率函数指导规则产生；Mr-SBC则作为MURENA系统的一个模块存在，它也首先产生一阶规则；1BC和1BC2在Tertius一阶描述学习器下实施，不同的是1BC使用Tertius产生一阶条件，而1BC2并不产生一阶条件，而

基金项目：国家自然科学基金资助项目(60675030)
作者简介：徐光美(1977－)，女，博士研究生，主研方向：多关系数据挖掘，概率逻辑学习；杨炳儒，教授；钱 榕，高级工程师、博士
收稿日期：2007-08-20 **E-mail：**xgm_xgm@126.com

是定义列表、多集和子集分布。

产生一阶规则过程中的关键问题之一是效率问题。理论上ILP系统可以产生任意长度的一阶规则，这就需要一些限定性知识进行剪枝，ILP系统一般都事先指定规则产生的长度，并通过数据库间的主键外键关系控制规则产生的方向。此外ILP方法需要在许多关系中评价规则，这些关系中每一个都可以由若干个连接路径连接到目标关系。因此，在搜索好的规则过程中通常需要构建很多连接关系，这在时间和空间上都是无效的。已经存在一些研究旨在提高ILP算法的效率和可伸缩性，如评价同时共享相同的前缀的查询包的方法以及CrossMine[7]中的元组ID传播方法等。

(3)计算一阶规则的条件概率(发现最优的参数)。在 P-K 方法中，条件概率由 *m* 估计计算得到，实验中 *m* 取 2。在 1BC 中则采用 Laplace 估计方法计算条件概率。每个一阶规则都由若干文字构成，而且多个规则中会有共享的文字。P-K 方法、1BC、1BC2 以及 nFOIL 中不区分共享文字和一般文字，需要多次计算共享文字的频率。Mr-SBC 则基于将一阶分类规则和朴素贝叶斯分类结合的思想，使得共享文字的概率计算同其余文字的概率计算分开，而且 Mr-SBC 允许存储路径概率。此外，1BC 和 1BC2 进行频率计数的主要不同在于：前者的统计计数基于个体，而后者则基于出现在给定上下文环境中的所有相关的对象。

(4)根据朴素贝叶斯公式计算规则集的联合概率，并对测试数据进行分类。这些方法扩展了朴素贝叶斯的独立假定，它们假定给定类别前提下，构成假设的子句是条件独立的。P-K 方法中关键在于引入了覆盖因子的概念，每一个规则相应于每一个测试数据都有一个覆盖因子，覆盖因子表示了测试数据满足规则集的情况，从而进行分类。nFOIL 中采用 Grossman 和 Domingos 提出的一种混合方法，即参数(概率值)根据最大似然进行估计(看作最大条件似然的一种近似)，而最大条件似然(MCL)则仍旧作为选择特征的评分函数。1BC 和 1BC2 在分类时的主要区别在于：前者中每个个体中出现的属性仅被考虑一次，而后者则考虑个体中每次出现的属性；后者不考虑在测试数据中没有出现的属性的概率，而前者则考虑那些不满足的非确定特征的概率。Mr-SBC 与 1BC 则使用相同的知识表示方法，在分类时也类似。

3 基于关系数据库技术的一阶扩展

上述基于ILP技术的MRNBC方法都需要将数据库中元组转化为Prolog事实，这个过程不仅会增加系统复杂性，还会丢失有用的模型知识。因此，研究者开始考虑将朴素贝叶斯方法直接与关系数据库技术结合以形成MRNBC，H. Liu和 X. Yin等提出了一种不基于ILP的MRNBC(称为H-X方法)[2]，它并不产生由谓词构成的规则，而是直接对数据库表进行操作，直接计算属性的类条件概率信息，然后用朴素贝叶斯公式进行分类，而且它们提出了一种称为“ Cutting off ”的“ Wrapper ”剪枝方法对表剪枝，以提高分类准确率。本文对朴素贝叶斯与关系数据库理论结合的方法做了深入研究，并提出了朴素贝叶斯与关系数据库理论相结合形成MRNBC的一般步骤：

(1)根据 E-R 图或其变体指导表连接，并选择时空代价较小的表连接方法。本步骤中关键问题有：

1)表连接的依据选择，如可以依据主键-外键关系连接表。H-X方法中定义了语义关系图的概念[2]，它可以由关系数据库的E-R图变换得到，并根据该图连接表。

2)表连接的方法(主要考虑时空代价)。ILP方法中都需要单独对每个谓词进行评价，为此需要多次连接数据库表，降低计算代价的方法之一是首先将目标关系和其他关系连接，接着在连接以后的关系上评价每一个谓词，这样许多谓词就可以同时计算而仅需要扫描结果关系一次，但是不断寻找好的谓词并建立规则的过程，需要以不同的方式连接很多关系，因此时空代价仍然很大。为降低数据库表之间连接时的时空代价，Xiaoxin Yin和Jiawei Han在CrossMine中提出了元组ID传播方法，避免了数据库表间的直接物理连接，而是采用虚拟的连接方式[7]。元组ID传播的基本思想是：沿着一定的传播路径，依次传播目标元组IDs和相应的类标签到其他非目标关系中。H-X方法中采用元组ID传播方法对表进行虚拟连接，大大降低了时空代价。

3)连接环的处理。可以通过控制连接深度避免多次重复连接，也可以利用背景知识或其他的操作进行拆环。

(2)根据领域选择合适的频率计数方法，计算属性的类条件概率值。

频率计数的基本单位可以是个体也可以是每个属性(参照 1BC2 中个体概念)，对于数据库中存在的一对多等复杂关系可以引入合适的聚合函数，也可以定义确定的分布函数以确定属性值的类条件概率。根据不同的领域，可以选择不同的计数方法。

(3)根据相应的频率技术方法构建多关系朴素贝叶斯分类公式。

根据(2)中的频率计数方法确定合适的分类公式，其中朴素贝叶斯独立假定也被扩展到一阶情况下，即假设不同关系中属性间的类条件概率也是相互独立的。

(4)选择合适属性约简方法对属性剪枝，提高分类性能。

为进一步提高分类准确率，可以采用各种剪枝方法进行属性约简，H-X 方法中提出了一种称为“ Cutting off ”剪枝方法对表进行剪枝，而且实验证明，这种剪枝可以很大程度上提高分类准确率。也可以选择其他以属性为单位的剪枝方法，如互信息方法。

选择剪枝方法的标准有两个一个是效率，另一个是分类准确率，因为本身多关系域的时空复杂度较高，所以寻求效率更高、更有效的的剪枝方法显得尤为重要。

4 结束语

实验证明将朴素贝叶斯分类器扩展到一阶情况下，可以有效地提高 ILP 系统的分类准确率，而基于关系数据库理论的 MRNBC 在采用一定的剪枝方法后也可以有效改进分类效果。此外所有的方法在分类时都扩展了朴素贝叶斯独立假定，但是，在很多情况下，关系间属性的类条件概率是相互关联的，这种独立性的破坏对分类的影响情况是下一步研究的方向之一，也可以考虑将更复杂的模型扩展到一阶情况下，如树增广的朴素贝叶斯模型等。

此外，目前已有的 MRNBC 中，大多首先对连续属性进行简单的离散化，然后计数，下一步可以考虑引入合适的概率分布函数，直接估计连续属性的概率。效率问题始终是多关系方法能否在实际中成功应用的关键，为此可以继续研究更有效的一阶规则产生方法和更合理的剪枝方法，降低系统时空代价。

（下转第 53 页）