

文章编号: 1000-5013(2008)03-0370-05

一种数据规约的近似挖掘方法的实现

喻小光, 陈维斌, 陈荣鑫

(华侨大学 信息科学与工程学院, 福建 泉州 362021)

摘要: 讨论基于数据规约的近似挖掘技术, 在数据预处理阶段对海量数据集进行数据规约. 近似数据挖掘的工作流程包括任务定义、数据准备与预处理、数据挖掘建模、结果的解释与评估、模型发布与应用 5 个阶段. 同时, 提出使用属性选择和实例选择方法实现近似挖掘的方案, 并对该方案进行挖掘效率和结果模型准确性的分析评估. 该方案能满足对企业级大数据集进行高效挖掘的需要.

关键词: 近似挖掘; 数据规约; 属性选择; 实例选择

中图分类号: TP 311. 12

文献标识码: A

数据量的迅速增长是数据挖掘技术发展的驱动因素, 然而, 对海量数据的数据挖掘也是对数据挖掘研究的一大挑战. 例如, 简单的穷举法、经验分析法对于小数据集可以工作得很好, 而对 TB 级的海量数据则无能为力. 目前, 有许多针对海量数据挖掘的解决方案, 其中近似挖掘方法能得良好的挖掘效果. 本文在数据归约方法基础上, 研究近似挖掘技术, 提出针对企业级海量数据挖掘的解决方案.

1 近似挖掘方法

1.1 近似挖掘的定义

近似挖掘是指在数据挖掘过程中, 不直接使用待挖掘的原始数据集, 而使用原始数据集的子集作为挖掘对象, 以获得近似挖掘效果的方法. 直接对海量数据集进行挖掘需要较长时间, 且未必能获得满意的结果. 其挖掘结果可能是一个复杂的模型, 尽管准确度较高, 但可解释性却大大降低. 数据归约技术可以用来获得数据集的简化表示 (简称近似子集), 并且该近似子集的信息表达能力非常接近于原数据集, 规模却小得多. 对经过归约预处理后的数据集进行挖掘, 可以产生相近的分析结果, 而效率却大大提高. 数据归约技术是近似挖掘的实现基础.

近似挖掘的基本思路是, 在数据预处理阶段对海量数据集进行数据归约. 数据规约一般可以采用属性选择法和实例选择法, 也可以二者结合使用.

1.2 近似挖掘的理论基础

实现近似挖掘的核心问题是, 如何判断保留哪些数据才能保证原始数据集中的有效信息不丢失或不严重丢失. 相关性理论为解决这个问题提供了理论基础. 近似挖掘要从原始的挖掘对象数据集中, 获得与挖掘目标相关性最好的部分作为替代的挖掘对象数据集, 所以本质是进行相关性分析.

从机器学习的角度来说, 相关性指与目标概念的关联性. 定义如下^[1]: 假设在实例空间中, 存在 A 和 B 两个样本, 对应于目标概念 C , 如果 A 与 B 有区别, 即 $C(A) \neq C(B)$, 是因为它们的属性 X_i 取不同的值, 那么, 属性 X_i 和目标概念 C 是相关的.

挖掘对象的复杂度包含属性和实例两方面的复杂度. 属性复杂度包括属性的数目和类型、属性取值的数目和范围、属性间的关联性等; 实例复杂度包括实例数目、实例分布等. 挖掘对象的数据质量直接影响挖掘的效果, 去除无关和弱相关的部分, 既降低了挖掘对象的复杂度, 又提高了挖掘对象的数据质量.

收稿日期: 2007-09-22

作者简介: 喻小光 (1976-), 男, 讲师, 主要从事数据库技术及应用与网络安全的研究. E-mail: xiaoyu@hqu.edu.cn.

基金项目: 福建省青年科技人才创新基金项目 (2002J011); 华侨大学科研基金资助项目 (04HZR17)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.3 近似挖掘的工作流程

近似数据挖掘的工作流程包括任务定义、数据准备与预处理、数据挖掘建模、结果的解释与评估、模型发布与应用 5 个阶段^[2-3], 如图 1 所示. 其中, 任务定义决定数据挖掘要发现何种知识, 是整个分析过程中的第 1 个阶段, 也是最重要的一个阶段. 数据准备与预处理包括数据整理、数据集成、数据选择、数据变换等步骤. 产生近似子集就在这个阶段进行, 也是近似挖掘方法与其他挖掘方法的最大区别. 数据挖掘建模指根据挖掘的任务或目的, 如分类、聚类、关联规则等, 选择具体的实现算法进行建模. 结果解释与评估指完成建模后, 采用可视化和用户易于理解的知识表示方式来表达挖掘结果. 模型发布与应用指发布通过评估的模型, 为用户提供分析决策服务.

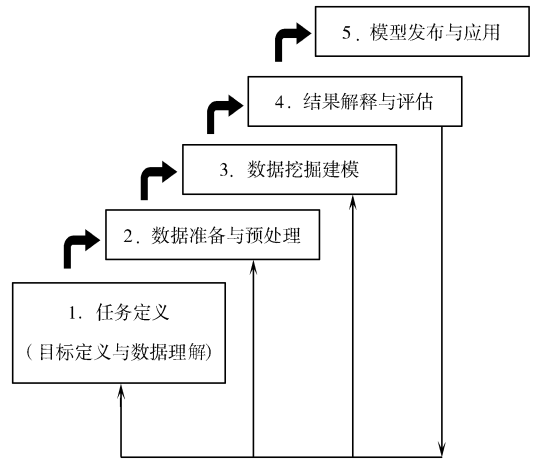


图 1 近似数据挖掘过程

Fig. 1 Approximate data mining process

2 近似挖掘的实现

近似挖掘的实现按上文提及的 5 个步骤依次进行. 在数据准备与预处理中, 近似挖掘采用与其他挖掘方法不同的策略, 数据规约有属性选择和实例选择两类方法, 以下分别讨论两种方法的实现.

2.1 属性选择

属性选择是根据某些指标选择一个优化属性子集的过程. 一个优化的子集可以是属性数目最小的子集, 也可以是具有最佳预测准确率的子集.

评估指标的选取与属性选择的目的有关, 可以是分类精度不会再减少, 也可以是由选择的属性产生的类分布尽可能地接近原始数据集的类分布. 属性选择可以看成是在属性子集状态空间的搜索过程. 属性选择方法由属性评估方法和搜索方法两部分组成. 本文讨论的属性类型主要有连续型和离散型两种.

2.1.1 属性选择的工作方式 根据属性选择与实际的挖掘算法之间的配合关系, 可以将属性选择的工作方式分为嵌入型、过滤型和封装型^[1,4]. 嵌入型属性选择完全融入具体的挖掘算法, 是挖掘算法的组成部分^[4]; 过滤型属性选择是独立于具体挖掘算法进行的属性选择. 封装型属性选择也在具体挖掘算法之外执行属性选择, 不过它通过调用具体挖掘算法获得挖掘准确率, 并以此作为属性选择的评估标准^[5]. 从整个挖掘过程看, 过滤型的属性选择可以当作预处理步骤, 简单易用, 而且产生的中间结果可以为后续的各种挖掘算法所利用.

2.1.2 评估方法 根据是否考虑到非类别属性之间的相关性, 把评估方法分为两类^[4], 基于单属性的选择和基于属性子集的选择. 前者仅考虑单个属性与类别属性的相关性, 属于该类型的方法有基于信息增益的属性排序法、基于实例的属性排序 Relief 法、主成分分析法等; 后者还要考虑属性子集内各个属性之间的相关性, 属于该类型的方法有基于相关的属性选择法 (CFS)^[6]、基于一致性的子集选择法^[7]等. CFS 是一种基于属性子集的评估方法. 算法的核心是一个启发式的子集评估算子, 用于测试单个属性对类别预测的贡献度以及属性之间的相关性. 评估算子把较高的权重赋予一些属性子集, 这些子集内的属性与类别属性有强相关性, 而本身内部属性之间又有较弱的相关性. CFS 方法事先会对连续属性加以离散化处理, 适用于各种类型属性的选择. 下文的属性选择实验就采用这种评估方法.

2.1.3 搜索方法 对于基于属性子集选择的评估方法, 通常需要结合某种属性空间的搜索方法, 完成属性选择, 比如采用启发式搜索技术^[8]. 实践表明这种技术可以提供良好性能, 但无法保证获得最优解. “贪心”启发算法^[9]是常用的搜索方法. 根据搜索的起始状态的不同, 使用贪心启发搜索方法可以分为向前选择算法和后向选择算法^[8]. 前者从最小容量的模型开始, 不断增加模型变量; 后者则从全模型开始, 以相反的方式工作. 当模型变量很大时, 往往采用向前选择算法. 下文的属性选择实例使用的是向前选择搜索策略.

2.1.4 实例效果评测 通过对 C4.5 算法建模的时间复杂度进行分析, 了解到属性选择对建模时间的作用, 为了进一步评估属性选择对决策树分类挖掘的时间、准确度和树的规模的影响, 用实验加以测试

实验硬件环境: 主频 2.8 GHz 的奔腾 4 处理器、512 MB 内存; 软件环境: Windows 2000 Server 操作系统和 Sun Application Server PE8.0EJB 应用平台. 分类挖掘使用 C4.5 算法, 采用基于误差的剪枝方法, 设置剪枝的置信度为 0.25, 用子树替代策略, 每个实验采用 10 个 10-折交叉验证获得准确率数据的平均值. 实验数据集来自 UCI 机器学习数据库^[10], 如表 1 所示. 表中, ϕ 为准确率, t 为建模时间, 属性选择采用 CFS 的评估方法, 结合向前选择搜索策略. 表 1 列出了未进行属性选择和进行属性选择两种情况下的挖掘结果. 需要说明的是, 进行属性选择的建模时间已经包括了属性选择和决策树生成与剪枝这两部分时间.

从建模结果可以发现, 对于较大的数据集(如 letter 和 adult)建模时间降低了, 达到提高挖掘效率的效果. 从准确率比较结果得知, 属性选择后的挖掘准确度虽有下降, 但不明显, 说明是可行的. 树的规模反映了模型的复杂度, 也影响了模型的可解释性. 尤其是 adult 数据集, 由于仅选择了相关性最强的 5 个属性进行建模, 大大降低了树的规模, 从而提高了模型的可理解性.

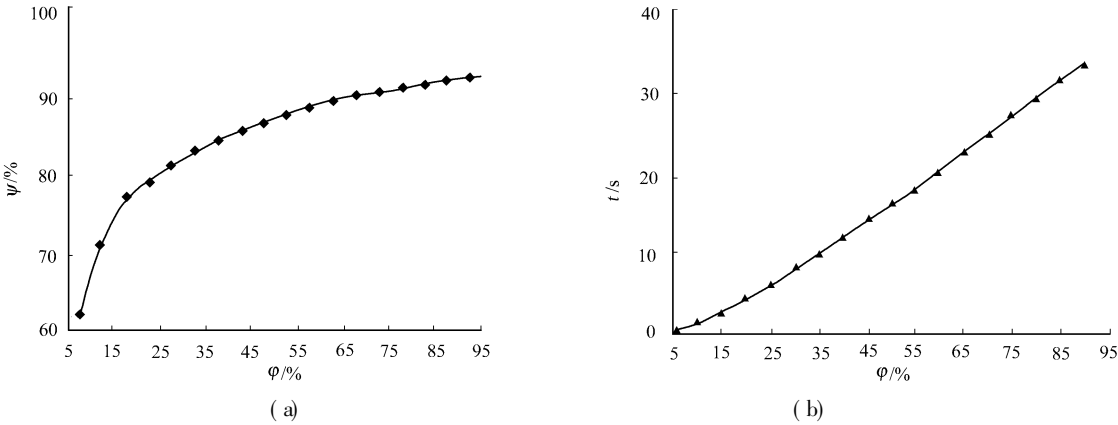
表 1 实验数据集与试验结果比较

数据集	实验数据					试验结果					
	实例 个数	属性 个数	类别 个数	连续型属 性比例/%	离散型属 性比例/%	未属性选择			属性选择		
						ϕ /%	树的规模	t /s	ϕ /%	树的规模	t /s
labor	57	16	2	50.0	50.0	73.7	5	0.01	80.9	6	0.01
soybean	683	35	19	0	100.0	91.5	93	0.22	90.4	90	0.20
mushroom	8 124	22	2	0	100.0	100.0	30	0.26	98.5	10	0.21
letter	20 000	16	26	100.0	0	92.5	2 451	38.53	87.3	2 423	30.80
adult	32 561	14	2	42.9	57.1	89.8	2 310	81.32	85.7	115	11.97

2.2 实例选择

实例选择的目的是, 使用部分数据记录替代原来的所有数据记录进行挖掘, 以便降低挖掘的时间和资源代价, 获得高效的挖掘性能. 实例选择主要是通过对挖掘对象数据集进行采样实现的. 采样是指从海量数据集中选择子集的过程. 目前已经有各种各样的采样方法^[1], 比如简单随机采样、等距采样、起始顺序采样、聚类采样和分层采样等.

本文仍然采用 C4.5 算法, 参数设置与属性选择时的设置相同, 结合最简单的随机采样技术, 对多个数据集进行测试, 试验环境与属性选择测试环境相同. 为了便于比较, 把各个参数在不同采样比例时的实际数值转换成与采样率达 100% (即不进行实例选择, 使用全部数据集作为训练对象) 时的数值比例. 可以发现, 建模时间(t)、产生树的规模(n)与采样率(ϕ)的增长基本上呈线性关系, 而挖掘模型的准确率(ψ)则与采样率成非线性关系, 如图 2 所示. 对于 letter 数据集, 当采样率低于 20% 时, 随着采样率的提高, 准确率有显著的提高, 采样率在 20% 时, 准确率就可达 80%; 随着采样率的进一步提高, 准确率的增加不明显. 这个结果在其他数据集的实验中也可以观察到. 因此可以得出的结论: 在挖掘实践中, 很多情况下完全可以通过实例选择, 使用很少量的数据进行训练, 在可接受的准确率范围内, 获得良好的时间性能和挖掘结果. 显然, 实例选择是近似挖掘的有效手段.



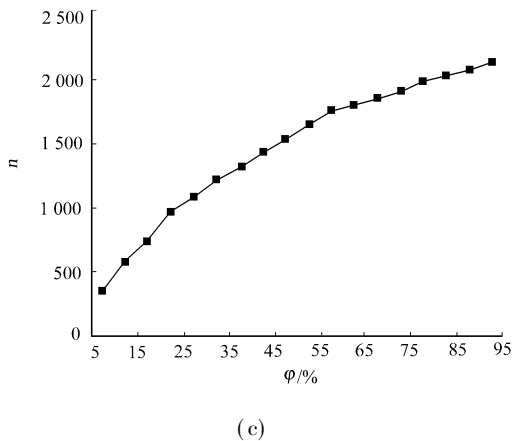


图 2 实例选择效果评测

Fig. 2 Effect evaluating of instance selection

3 应用实例

在银行贷款过程中, 风险管理是一项重要的内容. 一项贷款的审批, 必须考虑到客户的信用等级及偿还能力等因素. 考虑到与日俱增的业务量, 高效地进行客户资格评估, 自动完成申请审批成为银行的迫切需要. 这就要求银行建立一个分析模型, 如基于贷款资格的客户分类模型, 它能根据申请人的各种自然信息和社会信息来推断出其偿还能力, 从而决定是否批准其贷款申请. 然后, 应用该分类模型, 对申请贷款的客户进行资格评估. 使用该银行所有贷款申请和偿还纪录作为挖掘数据集. 每个记录包含教育背景, 婚姻状况, 家庭关系, 资产盈余, 资产损失、工作性质、工作单位、税务状况等指标. 分析人员的建模有 4 个步骤.

- (1) 任务定义. 获取一个基于贷款资格的客户分类模型.
- (2) 数据预处理. 根据挖掘数据集的实际情况, 采用属性选择和实例选择相结合的方法对原始数据集取近似子集. 属性选择使用 CFS 评估方法和向前选择搜索方法. 实例选择采用随机采样, 采样率设为 30% .
- (3) 数据挖掘建模. 本例采用基于 C4. 5 算法的分类挖掘, 并选用 10-折交叉验证法. 挖掘任务设置的界面如图 3 所示. 界面来源于数据分析助理系统(DAA). 分析人员完成挖掘任务的定义后“发送任务”, 将该任务文件提交服务器处理. 待服务器处理完毕后, 点击“获取结果”按钮(图 4), 可以在界面的结果栏浏览挖掘的结果. 挖掘结果包括文本形式的运行结果说明、图形化的测试结果视图和图形化的分类决策树(对应 C4. 5 算法结果) 视图.



图 3 任务定义界面

Fig. 3 Task definition interface

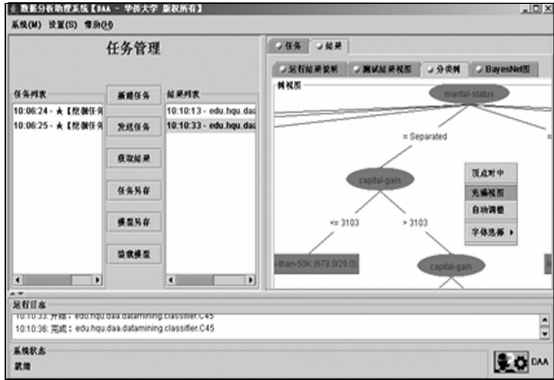


图 4 获取结果界面

Fig. 4 Result obtaining interface

- (4) 结果解释与评估. 获取分类模型后, 由银行业务分析人员对模型进行评估, 如模型正确, 则投入使用, 否则, 通知挖掘人员调整相关参数, 重新进行挖掘.

模型通过评估后,即可发布使用.系统根据客户提交的个人信息,结合基于贷款资格的客户分类模型进行分类预测.如,王五提交的个人信息:教育背景(Masters),婚姻状况(Married+ Civ-Spouse), 家庭关系(Ownr Child),资产盈余(0),资产损失(0),年龄(30),工作性质(College Instructor)、工作单位: (Huaqiao University)、税务状况(Well)等.由于预测结果是王五有还款能力,因此他通过了贷款审批.

4 结束语

近似挖掘技术有两个优点:(1)在保证模型性能的前提下,改善了挖掘效率.(2)提高了挖掘算法的适用性.这使得许多原本伸缩性能不好,不适合大数据集挖掘的挖掘算法能在企业级数据挖掘中得以使用.因此,近似挖掘技术使得高效挖掘算法对企业级海量数据进行挖掘成为可能.

参考文献:

- [1] BIUM L, LANGLEY P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence, 1997, 97(1-2): 245-271.
- [2] 邵峰晶, 于忠清. 数据挖掘原理与算法[M]. 北京: 中国水利水电出版社, 2003.
- [3] DUNHAM H. Data mining course[M]. Beijing: Tsinghua University Press, 2003.
- [4] HALL M A, HOLMES G. Bench marking attributes selection techniques for discrete class data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(3): 1-16.
- [5] KOHAVI R, JOHN H. Wrappers for Feature Subset Selection[J]. Artificial Intelligence, 1997, 97(1-2): 273-324.
- [6] HALL A. Correlation based feature selection for machine learning[D]. New Hamilton: University of Waikato, 1998.
- [7] DASH M, LIU H, MOTODA H. Consistency based feature selection[C] // Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence. Berlin: Springer Verlag, 2000: 98-109.
- [8] HAND D, MANNILA H, SMYTH P. Data mining principle[M]. Beijing: Publishing House of Mechanics Industry, 2003.
- [9] GREINER R. Probabilistic hill climbing: Theory and applications[C] // Proceedings of the Ninth Canadian Conference on Artificial Intelligence, 1992.
- [10] University of California. UCI machine learning databases[DB/OL]. [2006-09-20]. <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult>

Research and Realization of Approximate Data Mining Based on Data Reduction

YU Xiao-guang, CHEN Wei-bin, CHEN Rong-xin

(College of Information Science and Engineering, Huaqiao University, Quanzhou 362021, China)

Abstract: Data reduction based approximate data mining technique in which data reduction for massive data set was done in data pretreatment phase has been discussed. Approximate data mining work flow includes 5 phases, such as task definition, data preparing and pretreatment, data mining modeling, results explaining and evaluating and model publication. At the same time, the solution using attribute selection and instance selection to realize the approximation mining is brought out, and the mining efficiency and result model veracity are analyzed and evaluated. The solution can satisfy the need of mining on enterprise level massive data set.

Keywords: approximately mining; data reduction; attribute selection; instances selection

(责任编辑: 黄仲一 英文审校: 吴逢铁)