数据挖掘算法研究与综述

邹志文， 朱金伟
(江苏大学 计算机学院，江苏 镇江 212013)

摘 要：数据挖掘方法结合了机器学习、模式识别、统计学、数据库和人工智能等众多领域的知识，是解决从大量信息中获取有用知识、提供决策支持的有效途径，具有广泛的应用前景。以关联、分类、聚类归类，对当前数据挖掘的多种方法进行了研究，并指出其现存的问题。这些方法都有局限性，多方法融合、有机组合互补将成为数据挖掘的发展趋势。

关键词：数据挖掘；分类算法；关联分析；分类分析；聚类分析
中图法分类号：TP301.6 **文献标识码**：A **文章编号**：1000-7024 (2005) 09-2304-04

Research and summary of data mining algorithms

ZOU Zhi-wen, ZHU Jin-wei
(College of Computer, Jiangsu University, Zhenjiang 212013, China)

Abstract: Data Mining integrates with knowledge of numerous fields such as machine learning, pattern recognition, statistics, database and artificial intelligence. It is an effective approach to fetch useful information from large database and offer decision support. There is a broad application foreground of data mining. Many latest methods range by association, classification and clustering in data mining was researched, and their remaining problems were discussed. As a whole, all these algorithms have their own limitations, and organically combining several methods will be the development trend for data mining.

Key words: data mining; classification algorithm; association analysis; classification analysis; clustering analysis

1 引 言

随着信息化的到来，各类数据急剧膨胀，面对海量的存储数据，如何从中发现有价值的信息或知识是一项非常艰巨的任务。数据挖掘就是为了满足这种要求而迅速发展起来的。数据挖掘是从大量的、不完全的、有噪声的、模糊的和随机的数据中提取隐含在其中的、人们事先不知道的，但又是潜在有用的信息和知识的过程[1]。数据挖掘以机器学习、模式识别、统计学、数据库和人工智能等众多学科为基础，是目前国际上数据库和信息决策系统最前沿的研究方向之一，已引起了学术界和工业界的广泛关注。与此同时，各种数据挖掘算法纷纷出现，本文就目前有一定影响力的算法按基于关联、分类、聚类分别进行分析、评述，并指出了这一领域可能的发展方向。

2 关联分析

R.Agrawal 等人首先提出了关联规则挖掘问题。关联规则是数据挖掘研究的主要模式之一，侧重于确定数据中不同领域之间的关系，找出满足给定条件下的多个域间的依赖关系。关联规则挖掘对象一般是大型数据库(Transactional Database)，该规则一般表示式为：$A_1 \wedge A_2 \wedge \cdots A_m => B_1 \wedge B_2 \wedge \cdots B_m$，其中，$A_k$ $(k=1,2,\cdots,m)$，$B_j(j=1,2,\cdots,n)$是数据库中的数据项，且有 Support(A=>B)=P(A∪B)，Confidence(A=>B)=P(A|B) 数据项之间的关联，即根据一个事务中某些数据项的出现可以导出另一些数据项在同一事务中的出现。在关联规则挖掘法的研究中，算法的效率是核心问题，如何提高算法的效率是所要解决的关键。

2.1 Apriori 算法

在关联分析中经典算法是 R. Agrawal 等人提出的 Apriori 算法[2]，这是一种很有影响力的挖掘关联规则频繁项集的算法，它探查逐级挖掘 Apriori 性质：频繁项集的所有非空子集都必须是频繁的。根据频繁 k-项集，形成频繁 (k+1)-项集候选，并扫描数据库 1 次，完成第 k 次迭代 (k>1)，找出完整的频繁 (k+1)-项集 L_{k+1}。

Apriori 算法的优点是简单易懂；但同时也存在以下两方面的不足：①当事务数据库中的频繁 1-项集的数目$|L_1|$比较大时，由频繁 1-项集产生的候选 2-项集 C_2 就非常大，C_2 由 $C^2_{|L1|}$ 个 2-项集组成；②为了由 C_k 产生 L_k，需要重复扫描数据库中的事务并计算候选项集 C_k 中每个候选项集支持度，因而当事务数据库中的事务个数很大时，扫描数据库的开销将变得很大。

收稿日期：2005-03-12。
基金项目：国家 863 高技术研究发展基金项目 (2002AA412020)。
作者简介：邹志文 (1968-)，男，江西抚州人，硕士，讲师，研究方向为 webgis 和数据挖掘； 朱金伟 (1981-)，男，浙江永康人，硕士，研究方向为数据挖掘。

2.2 AprioriTid 算法

为了提高 Apriori 算法的有效性，目前已经提出了许多 Apriori 变形，旨在提高原算法的效率，在文献[2]中提出了 AprioriTid 算法。

从 AprioriTid 算法寻找频繁项集的思路中，可知道该算法的优点：即仅在第 1 次扫描时用事务数据库 D 计算候选频繁项集的支持度，其它各次扫描用其上一次扫描生成的候选事务数据库 D'来计算候选频繁项集的支持度。如此将减少对数据库的扫描次数，在一定情况下能迅速削减候选频繁项集。

即使进行了优化，但是 AprioriTid 方法一些固有的缺陷还是无法克服：①可能产生大量的候选集。当长度为 1 的频集有 10000 个的时候，长度为 2 的候选集个数将会超过 10M。还有就是如果要生成一个很长的规则的时候，要产生的中间元素也是巨大量的；②可能需要重复扫描数据库，通过模式匹配检查一个很大的候选集；③无法对稀有信息进行分析。

2.3 FP-growth 算法

为了解决这些问题文献[3]中采用了一种叫 FP-growth 的方法。它采取了分治策略：首先，构造频繁模式树 FP-树，根据事务数据库及设定的最小支持度阀值，将包含频繁项集的数据库压缩到 FP-树上；其次，在 FP-树上进行频繁模式的挖掘，FP-树的挖掘进行如下，由长度为 1 的频繁模式（初始后缀模式）开始，构造它的条件模式基（FP-树中和后缀模式一起出现的前缀路径集），然后构造条件模式基的 FP-树，即模式树的增长，并递归地在 FP-树上进行挖掘。

这种方法对于挖掘长的和短的频繁模式，都是有效和可以伸缩的，并比 Apriori 方法快了 1 个数量级；但是当数据库很大时，构造基于内存的 FP-tree 不太现实。

3 分　类

分类分析是通过分析训练集中的数据，为每个类别做出准确的描述或建立分析模型或挖掘出分类规则，以便以后用这个分类规则对其它数据库中的记录进行分类的方法。

3.1 决策树法

决策树归纳学习算法以其易于提取显式规则、计算量相对较小、可以显示重要的决策属性和较高的分类准确率等优点而得到广泛的应用。决策树根据不同的特征，以树型结构表示分类或决策集合，产生规则和发现规律。它的主要优点是描述简单，分类速度快，特别适合大规模的数据处理。

3.1.1 ID3 算法

Quinlan 在文献[4]中提出了著名的 ID3 算法，借用信息论中的互信息(信息增益)作为单一属性分辨能力的度量，试图减少树的平均深度，忽略了叶子数目的研究。ID3 的 1 个优点是：它的建树时间和任务的困难度(如样本集样本个数，每个样本的属性个数，研究概念的复杂程度即决策树的节点数)呈线性递增关系，计算量相对较小。但存在的主要问题有：①互信息的计算依赖于属性取值的数目较多的特征，而属性取值较多的属性不一定最优；②ID3 是非递增学习算法；③抗噪性差，训练例子中正例和反例较难控制。

Schimmer 和 Fisher 在文献[5]中设计了 ID4 递增式算法，通过修改 ID3 算法，在每个可能的决策树结点创建一系列表，每个表由未检测属性值及其示例组成，当处理新例时，每个属性值的正例和反例递增计量。

在 ID4 的基础上，Utgoff 在文献[6]中提出了 ID5 算法，它抛弃了旧的检测属性下面的子树，从下面选择属性构造树。

3.1.2 C4.5 算法

文献[7]提出了 C4.5 算法，对类 ID3 算法进行了改进，提高了算法的效率。尽管如此，C4.5 算法仍然有如下的缺点：首先，在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效；其次，C4.5 只适合于能够驻留于内存的数据集使用，当训练集大得无法在内存容纳时程序无法运行。

3.1.3 SLIQ 算法

文献[8]提出了 SLIQ 算法，使用 gini 指标(gini index)代替信息量(Information)，对数据集包含 n 个类的数据集 S，gini(S)定义为：$gini(S)=1-\sum p_j * p_j$，p_j 是 S 中第 j 类数据的频率，gini 越小，Informtion Gain 越大。

由于算法采用了“预排序”和“广度优先”这两种技术使得该算法能够处理比 C4.5 所能处理的大得多的训练集，因此在一定程度上具有良好的随记录个数和属性个数增长的可扩展性。然而它仍然存在如下缺点：①由于需要将类别列表存放于内存，而类别列表的长度与训练集的长度是相同的，这就在一定程度上限制了可以处理的数据集的大小；②由于采用了预排序技术，而排序算法的复杂度本身并不是与记录个数成线性关系，因此使得 SLIQ 算法不可能达到随记录数目增长的线性可扩展性。

3.1.4 SPRINT 算法

为了减少需要驻留于内存的数据量，文献[9]提出了 SPRINT 算法，进一步改进了决策树算法实现时的数据结构，去掉在 SLIQ 中需要驻留于内存的类别列表，将它的类别列合并到每个属性列表中。

其优点是：在寻找每个结点的最优分裂标准时变得相对简单一些。但是其缺点是：对非分裂属性的属性列表进行分裂变得很困难。解决的办法是对分裂属性进行分裂时用哈希表记录下每个记录属于哪个孩子结点，若内存能够容纳下整个哈希表，其它属性列表的分裂只需参照该哈希表即可。

3.1.5 RainForest 算法框架

在过去的研究提出的多种决策树算法中，到目前为止还没有一种算法在任何数据集合下生成决策树的质量方面能超过所有其它的算法。文献[10]提出了 RainForest 算法框架，该框架关注于提高决策树算法的伸缩性，该框架可运用于大多数决策树算法(例如 Sprint 和 SLIQ)，使算法获得的结果与将全部的数据放置于内存所得到的结果一致，但是在运行时可以使用较少的内存。生成的决策树的质量取决于具体的决策树算法，与本框架无关。因此，在内存一定的情况下，可以更好地满足算法的需求。

3.2 Bayes 分类算法

Bayes 分类算法是利用概率统计知识进行分类的算法，主要利用 Bayes 定理来预测 1 个未知类别的样本属于各个类别的可能性，选择其中可能性最大的 1 个类别作为该样本的最终类别。由于贝叶斯定理的成立本身需要一个很强的独立性假设前提，而此假设在实际情况中经常是不成立的，因而其分类准确性就会下降。较有代表性的算法是 NB 算法[11]，还有降

低独立性假设的 TAN(tree augmented Bayes network)算法[11]。

3.3 CBA（Classification Based on Association）算法

CBA 算法[12]是基于关联规则发现方法的分类算法。该算法分两个步骤构造分类器：第 1 步，发现所有的右部为类别的类别关联规则(classificationassociationrules，简称 CAR)；第 2 步，从已发现的 CAR 中选择高优先度的规则来覆盖训练集。

CBA 算法的优点是：其分类准确度较高，因为它发现的规则相对较全面。但是，当最小支持度被设为 0 时，产生的频繁集有时多得在内存无法容纳，从而会使程序无法继续运行。

3.4 MIND（Mining in Database）算法

MIND 算法[13]是采用数据库中用户定义的函数(user-defined function，简称 UDF)来实现发现分类规则的算法。该算法的优点是：通过采用 UDF 实现决策树的构造过程使得分类算法易于与数据库系统集成。该算法的缺点是：算法采用 UDF 完成主要的计算任务，而 UDF 一般是由用户利用高级语言实现的，无法使用数据库系统提供的查询处理机制，无法利用查询优化方法，且 UDF 的编写和维护相当复杂。另外 MIND 中用 SQL 语句实现的那部分功能本身就是比较简单的操作，而采用 SQL 实现的方法却显得相当复杂。

3.5 神经网络

神经网络是由大量的简单神经元，通过极其丰富和完善的连接而构成的自适应非线性动态系统，并具有分布存储、联想记忆、大规模并行处理、自组织、自学习、自适应等功能。在数据挖掘领域，主要采用前向神经网络[14]提取分类规则。

其最大的缺点是“黑箱”性，人们难以理解网络的学习和决策过程。通常有两种解决方案：①建立一个基于规则的系统辅助；②直接从训练好的网络中提取规则。

3.6 粗集理论

粗集理论[15]的特点是不需要预先给定某些特征或属性的数量描述，如统计学中的概率分布、模糊集理论中的隶属度或隶属函数等，而是直接从给定问题出发，通过不可分辨关系和不可分辨类确定问题的近似域，从而找出问题中的内在规律。粗集理论同模糊集、神经网络、证据理论等其它理论均成为不确定性计算的一个重要分支。在数据挖掘领域，粗集方法广泛应用于不精确、不确定、不完全的信息的分类和知识获取。

粗集的数学基础是集合论，难以直接处理连续的属性。而现实决策表中连续属性是普遍存在的。因此连续属性的离散化是制约粗集理论实用化的难点之一。目前比较有代表性的监督离散化方法有以下几种：①Holte 提出了一种贪婪的单规则离散器(one rule discretizer)方法；②统计检验方法；③信息熵方法等。

以上几种方法各有特点，但都存在 1 个不足：每个属性的离散化过程是相互独立的，忽略了属性之间的关联，从而使得离散的结果中含有冗余或不合理的分割点。

3.7 遗传算法

遗传算法是模拟生物进化过程，利用复制(选择)、交叉(重组)和变异(突变)3 个基本算子优化求解的技术。遗传算法类似统计学，模型的形式必须预先确定，在算法实施的过程中，首先对求解的问题进行编码，产生初始群体，然后计算个体的适应度，再进行染色体的复制、交换、突变等操作，优胜劣汰，适者生存，直到最佳方案出现为止。

在数据挖掘领域，遗传算法的作用表现在以下几个方面：①和神经网络、粗集等技术的结合。如用遗传算法和 BP 算法结合训练神经网络，然后从网络提取规则。实践证明这是一种有效的方法；②分类系统的设计。遗传算法用于分类器始于 20 世纪 80 年代初。90 年代后，遗传算法用于分类系统的理论得到广泛的研究和应用。

遗传算法具有计算简单、优化效果好的特点，它在处理组合优化问题方面也有一定的优势。但还存在以下问题：算法较复杂，收敛于局部极小的过早收敛等难题未得到彻底解决。

4 聚类分析

聚类分析与分类不同，聚类分析处理的数据对象的类是未知的。聚类分析就是将对象集合分组为由类似的对象组成的多个簇的过程。

4.1 Ipartitioning method（划分方法）

给定 1 个 N 个对象或者元组的数据库，1 个划分方法构建数据的 K 个划分，每 1 个划分表示 1 个聚簇，并且 K<N。

经典算法有 K-MEAN（K 平均值）[16]、K-MEDOIDS（K 中心点）[17]，而且这些算法已经被加入到许多统计分析软件包或系统中，例如 SAS、SPSS。

4.2 hierarchical method（层次方法）

层次方法对给定数据对象集合进行层次的分解。根据分解的形成不同，层次方法可以分为凝聚的层次方法和分裂的层次方法。

层次方法存在着缺陷：一旦 1 个步骤完成，它就不能被撤消，这样它就不能更正错误的决定。有两种方法可改进层次聚类的结果：第 1，使用 CURE 和变色龙方法中的做法，在每个层次划分时，仔细分析对象之间的联接；第 2，综合层次凝聚和迭代的重定位方法，首先用凝聚的自底向上的分析算法，然后进行迭代的重定位来改进结果，BIRTH 中用的就是这种算法。

4.3 grid-based method（基于网格的方法）

这种方法采用一个多分辨率的网格数据结构。将空间量化为有限数目的单元，这些单元形成了网格结构，所有聚类分析都在网格上进行。这种方法主要优点是：处理速度快，它的处理时间仅依赖于量化空间中每一维上的单元数目，却独立于数据的数目。常用的算法有 STING[18]、WAVECLUSTER[19]和 CLIQUE[20]。

4.4 其它基于模型的聚类分析方法

主要有统计学和神经网络方面的方法。

5 结 论

随着数据量的日益积累以及数据库种类的多样化，数据挖掘的应用前景相当广阔。本文对各类算法进行了分析、比较和总结。总而言之，各种数据挖掘方法作用范围有限，都有局限性，因此采用单一方法难以得到决策所需的各种知识。但它们的有机组合具有互补性，多方法融合将成为数据挖掘算法的发展趋势。

参考文献：

[1] Han J, Kambr M. Data mining:Concepts and techniques[M].

Beijing Higher Education Press, 2001.1-3.
[2] Agrawal R, Srikant R.Fast algorithm for mining association rules in large databases[A]. The International Conference on Very Large Data Bases [C]. 1994.487-499.
[3] Han J, Pei J, Yin Y. Ming frequent patterns without candidate generation[A]. In Proc.2000ACM-SIGMOD Int. Conf. Management of Data (SIGMOD '00)[C]. Dallas, TX, 2000.1-12.
[4] Quinlan J R. Induction of decision trees [J]. Machine Learning, 1986,(1): 81-106.
[5] Schlimmer J C, Fisher D. A case study of incremental concept induction[A]. In Proceedings of AAAI-86[C]. 1986.
[6] Utgoff P E. ID5: An incremental ID3 [A]. In Proceedings of ICML-88[C]. San Mateo, CA,1988.
[7] Quinlan J R. C4.5: Programs for machine learning [M]. San Mateo,California: Morgan Kaufmann, 1993.
[8] Mehta M, Agrawal R, Rissanen J. SLIQ: A fast scalable classifier for data mining[A].Lecture notes in computer sciProc of the 5th int conf on extending database Tech[C]. Avignon, France, 1996.18-33.
[9] Shafer J C, Agrawal R, Mehta M. SPRINT:A scalable parallel classifier for data mining[A].Proc of the 22nd lnt conf on very-large databases[C]. Mumbai(Bombay), India, 1996.
[10] Gehrke J, Ramakrishnan R, Ganti V. Rainforest a framework for fast decision tree construction of large datasets[A]. In VLDB[C]. 1998.
[11] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifier [J]. Machine L earning, 1997, 29(1): 131-163.
[12] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining[A]. Proc of the 4th int conf on knowle-dge discovery and dataMining[C]. NY, USA:AAAIPress, 1998.80-86.
[13] WANG M, Iyer B, Vitter J S. Scalable mining for classification rules in relational databases[A]. Eaglestone B, DesaiBC, SHAO Jianhua. Proc of the 1998 Int database eng and appl symp[C]. Cardiff, Wales, UK:IEEEComputer Society, 1998.58-67.
[14] Wang Li qiang,Tang Chang jie.Dataminingon Web[J]. Computer Applications,1998,18(10):912.
[15] 李永敏,朱善君.基于粗糙理论的数据挖掘模型[J].清华大学学报(自然科学版),1999,39(1):110-113.
[16] MacQueen J. Some methods for classification and analysis of multivariate observations[A].Proc 5th berkeley symp.math statist[C]. Prob, 1967-01.
[17] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis[M]. John Wiley and Sons, 1990.
[18] Wei Wang, Jiong Yang, Richard Muntz. STING: A statistical information grid approach to spatial data mining[A]. Twenty-third international conference on very large data bases[C], 1997.
[19] Sheikholeslami G, Chatterjee S, Zhang A. Wave cluster: a multi-resolution clustering approach for very large spatial databases [A]. Proc. Int. Conf. on very large data bases[C]. New York, NY, 1998.428-439.
[20] Agrawal R, Gehrke J, Gunopulos D. Automatic subspace clustering of high dimensional data for data mining applications[A]. Proc. ACM SIGMOD int. conf. on management of data[C]. Seattle, WA, 1998.94-105.

(上接第 2281 页)

3.3 SDG 自动建模策略

由于实际系统都有比较完备的系统工艺流程图 (Process Flow Diagram 简称 PFD)。通过工艺流程图,业主、建筑设计者、操作人员、全权评价人员和其他人员可以快速了解系统信息如设备位置容量、流量、进出口流体温度、泵的数据、系统压力等及系统价值、潜在问题等。

SDG 自动建模策略的提出是为了后续可以建立自动建模软件平台,以及和已经有的 SDG 推理软件连接以改善目前计算机在工程领域中使用缺乏连贯性的弱点,大力提高计算机的使用效率。通过考虑实效性、可扩展性、精确度,拟定采用基于流程的方法,辅助专家经验,使用面向对象的策略和模块化思想进行 SDG 自动建模策略研究。

4 结 论

综上所述,UML 的特色和 SDG (HAZOP) 模型的自身特点决定了 UML 的很多理念对研究 SDG (HAZOP) 模型自动建立策略有不少借鉴价值。

UML 严谨的定义、面向对象的方法、图形化的表示、软件内部的无缝连接、增量式开发思路、UML 本身的扩展机制对于 SDG 自动建模软件平台的分析设计都是比较理想的选择。使用 UML 方法分析设计该软件平台,可以使该平台从静态结构、动态行为、实现构造、模型组织管理、可扩展性等方面有优秀的表现。对于 SDG (HAZOP) 模型自动建立领域的空白来讲,UML 方法的引入无论是为 SDG (HAZOP) 模型的自动建模策略研究还是为 SDG 自动建模软件平台的后续开发都树立了良好的开端。

参考文献:

[1] 张贝克,夏涛,吴重光.集成化 SDG 建模、推理与信息处理软件平台[J].系统仿真学报,2003,15(10):1360-1363.
[2] 姚淑珍, 唐发根. UML 参考手册 [M]. 北京: 机械工业出版社, 2001.
[3] 李安峰,夏涛,张贝克,等.化工过程 SDG 建模方法[J].系统仿真学报,2003,15(10):1364-1368.
[4] Kletz T Hazop. Identifying and assessing process industry hazards[M] . Third Edition. Rugby ,Warwickshire CV21 3HQ ,UK: Institution of Chemical Engineers, 1992.
[5] Lapp S A, Powers G J. Computer-aided synthesis of fault trees [J]. IEEE Trans Reliability, 1977, 26(4): 2-13.
[6] Hiranmayee Vedam,VenkatVenkatasubramanian.Signed digraph based multiple fault diagnosis [J]. Computers Chem, 1997,21: 655-660.