DOI:CNKI:50-1165/N.20110707.1744.011

数据挖掘中分类算法综述*

李玲俐
(广东司法警官职业学院 信息管理系,广州 510520)

摘要:对分类算法中需要解决的关键问题进行了分析;综述了不同分类算法的思想和特性,决策树分类算法能够很好地处理噪声数据,但只对规模较小训练样本集有效;贝叶斯分类算法精度高、速度快,错误率低,但分类不够准确;传统的基于关联规则算法分类准确率高,但容易受硬件内存的制约;支持向量机算法分类准确率高、复杂度低,但速度慢。针对各种分类算法的缺陷,结合其优点,论述了当前一些速度更快、准确率更高、能实现更好分类效果的新算法,如多决策树综合技术、基于先验信息和信息增益的混合分类算法,基于粗糙集和遗传算法的神经网络分类算法等;对数据挖掘分类算法作了展望,提出今后的研究重点。

关键词:数据挖掘;分类;综述
中图分类号:TP391 文献标志码:A 文章编号:1672-6693(2011)04-0044-04

数据挖掘是从海量数据中获取有用知识和价值的过程,是数据库技术自然演化的结果。数据挖掘已广泛应用于零售、金融、保险、医疗、通讯等行业,并展现出了其强大的知识发现的能力。在数据挖掘的研究与应用中,分类(Classification)算法一直受学术界的关注,它是一种有监督的学习,通过对已知类别训练集的分析,从中发现分类规则,以此预测新数据的类别[1]。数据分类算法中,为建立模型而被分析的数据元组组成的数据集合称为训练数据集,训练数据集中的单个样本(或元组)称为训练样本。分类算法是将一个未知样本分到几个已存在类的过程[2],主要包含两个步骤:第 1 步,根据类标号已知的训练数据集,训练并构建一个模型,用于描述预定的数据类集或概念集;第 2 步,使用所获得的模型,对将来或未知的对象进行分类。

1 分类算法中的关键问题分析

不同的分类算法有不同的特性,完成不同的任务。目前很多分类算法被机器学习、专家系统、统计学和神经生物学等的研究者从不同角度提出,判断不同分类算法的好坏可以由准确率、速度、健壮性、可伸缩性、可解释性等几个标准来衡量[3]。

另外,分类算法的效果通常和数据的特点有关,有的数据有空缺值,有的噪声大,有的分部稀疏,有的属性是连续的,有的则是离散或混合的[4]。经典的分类算法都有在不同的领域取得成功,比如决策树分类算法用于医疗诊断、金融分析、评估贷款申请的信用风险等广阔领域;支持向量机分类算法应用于模式识别、基因分析、文本分类、语音识别、回归分析等领域;由于对噪声数据具有很好的承受能力,神经网络广泛应用在字符识别、分子生物学、语音识别和人脸识别等领域。但每种分类算法都存在优缺点,加上数据的多样性以及实际问题的复杂性,使到目前为止,没有哪一种算法优于其他分类算法。例如,尚未有一种分类算法在任何数据集合下生成决策树的质量方面超过其他算法;神经网络是基于经验风险最小化原则的学习算法,本身存在一些固有的缺陷,而这些缺陷在 SVM 算法中可以得到很好解决。所以,如何寻找合适的分类算法是实际应用中亟待解决的问题。

2 数据挖掘的主要分类算法

数据挖掘的分类算法有多种,本文重点描述决策树、贝叶斯、基于关联规则、支持向量机等分类算法的特性及其新发展。

2.1 决策树分类算法

决策树分类算法也称为贪心算法,采用自顶向下的分治方式构造,它从一组无次序、无规则的事例

* 收稿日期:2011-05-16 网络出版时间:2011-07-07 17:44:00
作者简介:李玲俐,女,讲师,硕士,研究方向为数据挖掘与模式识别。
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20110707.1744.201104.44_011.html

中推理出决策树表示形式的分类规则[3]，是以实例为基础的归纳学习方法。决策树分类算法对噪声数据有很好的健壮性，能够学习析取表达式，是最为广泛使用的分类算法之一。

决策树的每个内部节点（非叶节点）表示在一个属性上的测试，每个分枝代表一个测试输出，每个叶节点代表类或类分布，树的顶层节点是根节点。决策树算法通过将样本的属性值与决策树相比较，来对未知样本进行分类，其生成过程如图 1 所示。

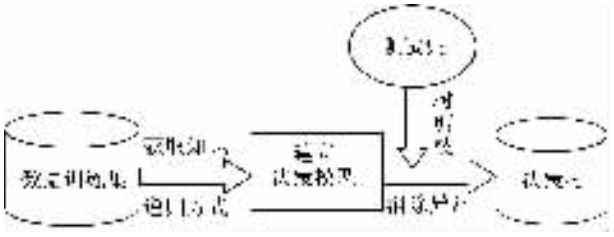

图 1　决策树的生成过程

首先根据训练数据集来构建决策树，建立决策树模型，这实际上是一个从数据中获取知识，进行机器学习的过程。树代表训练样本的单个根节点开始，使用分类属性（如果是量化属性，则需要进行离散化），递归地通过选择相应的测试属性来划分样本，一旦一个属性出现在一个节点上，就不在该节点的任何后代上出现，测试属性是根据某种启发信息或者是统计信息来进行选择（如信息增益）。第二个阶段是树剪枝，树剪枝试图检测和剪去训练数据中的噪声和孤立点，尽量消除模型中的异常。剪枝后的树变小、复杂度降低，在正确地对独立检验数据分类时效果更快更好。

ID3、C4.5 算法是最早的决策树分类算法，但只是对规模较小训练样本集有效。针对 ID3 算法构造决策树复杂、分类效率不高的问题，文献[5]采用加权分类粗糙度作为节点选择属性的启发函数，提出基于粗糙集理论的决策树构造算法，无论在规模或是分类效率上均优于 ID3 算法。Olaru. R 提出了一种基于模糊方法的软决策树算法，极大地提高了树的正确率和归纳能力。王熙照教授等研究者为处理多类问题，采用基于层次分解的方法产生多层决策树；针对 C4.5 算法的不足，提出新的决策树算法解决归纳学习的判决精度问题[6]。还有一种多决策树综合技术，先将数据集分成多个子数据集，然后将生成的多个不同的决策树综合起来，生成最终的、最稳定的决策树。

2.2 贝叶斯分类算法

贝叶斯（Beyes）分类算法基于概率统计学的贝叶斯定理，是一种在先验概率与类条件概率已知的情况下，预测类成员关系可能性的模式分类算法，如计算一个给定样本属于一个特定类的概率，并选定其中概率最大的一个类别作为该样本的最终类别。假设每个训练样本用一个 n 维特征向量 $X = \{x_1, x_2, \cdots, x_n\}$ 表示，分别描述 n 个属性 $A_1, A_2, \cdots, A_n$ 对样本的测量。将训练样本集分为 m 类，记为 $\theta_1, \theta_2, \cdots, \theta_i, \cdots, \theta_m$。贝叶斯原理通常用下面的公式来表示[7]。

$$P(\theta_i \mid X) = \frac{P(X \mid \theta_i) P(\theta_i)}{\sum_{j=1}^{n} P(X \mid \theta_j) P(\theta_j)}$$

其中，X 表示观测数据样本，θ_j 为某种假设，$P(\theta_i)$ 是 θ_i 的先验概率，$(i, j = 1, 2, \cdots, n)$ $P(X \mid \theta_i)$ 是条件概率，先验概率对条件概率加权平均后，得到条件 X 下，θ_i 的后验概率 $P(\theta_i \mid X)$。上述是朴素贝叶斯分类的工作过程，也是贝叶斯分类算法的判决准则。

贝叶斯分类算法的关键是使用概率表示各种形式的不确定性。对于大型数据集，从理论上讲，精确度高，运算速度快，具有最小的错误率，是贝叶斯算法的最大优点，但实际情况下，因其假定的不准确性，导致缺乏可用的数据，就需要足够大的样本。针对该缺陷，出现了一些降低独立性假设的贝叶斯改进分类算法，如半朴素贝叶斯算法、压缩候选的贝叶斯信念网络构造算法、TAN 算法等。贝叶斯分类算法还可以用来对不直接使用贝叶斯定理的其他分类算法提供理论判据[3,8]。文献[9]基于聚类分析思想，提出一种合理性、可信度都优于朴素贝叶斯缺损数据的修补算法。利用贝叶斯和决策树分类算法的优点，文献[10]提出将贝叶斯的先验信息法与决策树分类的信息增益法相结合的混合分类算法，在处理不一致或者不完整数据时，比单纯使用贝叶斯或决策树进行的分类运算速度更快，准确率更高。

2.3 基于关联规则分类算法

针对贝叶斯分类算法需要大样本量的缺点，研究者提出了基于关联规则（Classification based on association rule，CBA）的分类算法。CBA 算法通过发现样本集中的关联规则来构造分类器，其经典算法 Apriori[1]，通过 3 个步骤来构造分类器[10]，基于规则的分类器使用‘ if … then … ’来分类记录，其优先考虑置信度，迭代检索出数据集中所有的支持度不低于用户设定阈值的项集。基于关联规则分类算法集分类器构造与属性相关分析于一体，发现的规则

相对较全面且分类准确度较高,是一种很有潜力的分类算法。

传统的关联规则分类算法 Apriori 容易受到硬件内存的制约,时间代价高昂,针对该不足,研究者提出一种能够用于对等网模型 Kademlia 的分布式关联规则挖掘算法[11],该改进的 Apriori 算法通过对频繁项集阈值的设置,减少中间候选项集的数量,降低算法复杂度,提高算法执行效率;能解决传统关联分类算法中存在的冗余和冲突规则问题的基于有效规则提取的关联分类算法[12];优先考虑短规则分类的关联分类算法[13];针对网络入侵检测事务流日志数据库的关联规则挖掘改进算法[14],解决了当前主流关联规则算法应用到入侵检测过程中存在的多遍扫描、大量无效规则和算法复杂度过高等问题。

2.4 支持向量机分类算法

支持向量机(Support vector machine,SVM)分类是基于结构风险最小化准则的机器学习算法,使用数学方法和优化技术,具有优良的性能指标。SVM 算法用于数据预处理、样本化等 KDD 的过程,可以提高学习机的泛化能力。SVM 算法选择和保存有用的训练数据即支持矢量,该算法先自动找出对分类有较好区分能力的支持矢量,然后构造出分类器来最大化类与类的间隔,因此有较好的适应能力和较高的分准率;借助 SVM,类所属方法的分类准确度得到了很大提高并且时间复杂度得到了降低,大型数据库中小样本的训练数据的计算复杂度也得到了降低。从理论上讲,SVM 算法解决了在神经网络算法中无法避免的局部最小化问题。

但是,处理大规模数据集时,SVM 速度慢,往往需要较长的训练时间,针对该问题,文献[15]提出了一种缩减数据集以提高训练速度的算法,保证分类准确率,并有效地提高分类速度;文献[16]提出一种能提高分类正确率、速度以及使用样本的规模,并能增强 SVM 泛化能力的 BS-SVM 算法;基于 SVM 的优越性及其在声音信号分类中的广泛应用,采用 SVM 的识别算法,通过害虫产生的声音来识别害虫的种类[17];为提高音频特征的向量精度,文献[18]提出了一种基于小波变换和 SVM 的音频特征提取和分类的算法;文献[19]提出一种基于分类和回归树的学习算法,对不同特点的客户采用不同的手机银行模型,以提高其安全性能。

2.5 其他分类算法

除上述分类算法,常用的还有粗糙集、遗传算法、神经网络等分类算法。粗糙集算法以发现不准确数据或噪声数据内的结构联系,其知识表示是产生式规则。遗传算法基于生物进化思想,通过模拟自然进化过程搜索最优解,是现代智能计算中的关键技术之一。神经网络是一组连接的 I/O 单元,其中每个连接都与一个权重相关联[20]。神经网络分类中最流行的算法是 BP(Back propagation)算法、Hopfield 算法和后向传播分类算法。目前,研究者将神经网络算法与遗传算法、粗糙集算法、粒子群优化算法、蚁群算法相结合,如将粗糙集理论应用到 CBA 算法中,以提高分类关联规则的生成效率和准确度[21];文献[22]提出一种将粗糙集和遗传算法相结合神经网络模型,最大程度简化网络训练样本,优化神经网络结构,提高系统的学习效率和精度。算法的结合使用极大推动了数据挖掘分类技术的应用。

3 研究展望

本文对分类算法的研究现状及关键问题进行综述,详细讨论了决策树、贝叶斯、基于关联规则、支持向量机以及神经网络等分类算法的研究发展[23-25]。在数据挖掘应用中,用户要根据数据的特点,选择合适的分类算法或混合交互分类算法。在今后的工作中,为进一步提高分类的准确率、降低计算复杂度,更应该综合多领域技术,将分类算法与多学科相互交叉相互渗透,使之向着更多样化方向发展。

参考文献:

[1] 郭炜星. 数据挖掘分类算法研究[D]. 杭州:浙江大学,2008.

[2] Quinlan J R. Induction of decision tree[J]. Machine Learning,1986,1(1):81-106.

[3] Han J W,Micheline K. Data mining——concepts and techniques[M]. 北京:高等教育出版社,2001.

[4] 张丽娟,李丹军. 分类方法的新发展:研究综述[J]. 计算机科学,2006,33(10):11-12.

[5] 丁春荣,李龙澍,杨宝华. 基于粗糙集的决策树构造算法[J]. 计算机工程,2010,36(11):75-77.

[6] 季桂树,陈沛玲,宋航. 决策树分类算法研究综述[J]. 科技广场,2007(1):9-12.

[7] Quinlan J R. Learning efficient classification procedures and their application to chess and games [C]//Michalski R S, Carbonell J G, Mitchell T M. Machine learning:an artificial intelligence approach, CA:Morgan Kaufmann, 1983:463-482.

[8] 史忠植. 知识发现[M]. 北京:清华大学出版社,2002:1-265.
[9] 余瑞康,施润身. 聚类思想在贝叶斯算法中的应用[J]. 计算机工程与应用,2006,28:159-160.
[10] 樊建聪,张问银,梁永全. 基于贝叶斯方法的决策树分类算法[J]. 计算机应用,2005,25(12):2882-2884.
[11] 郭鸿,黄桂敏,周娅. 基于 Kademlia 的下关联规则挖掘算法研究[J]. 计算机工程与设计,2011,32(1):221-223.
[12] 武建华,沈钧毅,方加沛. 提取有效规则的关联分类算法[J]. 西安交通大学学报,2009,43(4):22-25.
[13] 武建华,沈均毅,王元元. 一种改进的关联分类算法[J]. 计算机工程,2009,35(9):63-65.
[14] 安德智. 改进的 Apriori 算法在 IDS 中的应用[J]. 河北理工大学学报:自然科学版,2011,33(11):95-99.
[15] 张珍珍,董才林,陈增照,等. 改进的结合密度聚类的 SVM 快速分类方法[J]. 计算机工程与应用,2011,47(2):136-138.
[16] 郭亚琴,王正群. 一种改进的支持向量机 BS-SVM[J]. 微电子学与计算机,2010,27(6):54-56.
[17] 唐发明,陈绵云,王仲东. 基于支持向量机的仓储害虫声音识别[J]. 华中科技大学学报:自然科学版,2005,33(2):34-36.
[18] 郑继明,俞佳. 基于小波变换和支持向量机的音频分类[J]. 计算机工程与应用,2009,45(16):161.
[19] Samaneh S J,Amtrhassan M J,Zahra Z J J. A model for adoption of mobile banking services using classification and regression trees[J]. US-China Public Administration,2010(11):66-73.
[20] 范明,孟小峰. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001.
[21] 禹蒲阳. CBA 分类算法的一种改进[J]. 计算机应用与软件,2010,27(8):241-243.
[22] 温泉彻,彭宏,黎琼. 基于粗糙集和遗传算法的神经网络模型研究[J]. 计算机工程与设计,2007,28(11):2652-2654.
[23] 陶春梅,王洪炼. 基于组织进化和信息熵的数据驱动分类算法[J]. 重庆邮电大学学报:自然科学版,2009,21(4):512-517.
[24] 王柯柯,崔贯勋,倪伟,等. 基于单元的快速的大数据集离群数据挖掘算法[J]. 重庆邮电大学学报:自然科学版,2010,22(5):673-677.
[25] 万红新,彭云,聂承启基于模糊集和粗糙集的关联规则挖掘策略[J]. 江西师范大学学报:自然科学版,2005,29(1):23-25,30.

A Review on Classification Algorithms in Data Mining

LI Ling-Li

(Dept. of Information Management, Guangdong Justice Police Vocational College, Guangzhou 510520, China)

Abstract: In this paper, we analyzed some key problems that must be solved in classification. Then, the idea and characteristic of main kinds of classification algorithms are reviewed. Decision tree algorithm can handle noise data well but is only effective to small datasets. Bayesian has the merits of high accuracy, fast speed, low mistake rate and demerits of low accuracy. Classification based on association rule has advantages of high accuracy but is limited to random access memory. Support vector machine has the merits of high accuracy, low complexity but shows bad time complexity. According to the advantages and disadvantages of the well-known algorithms, some recent proposed classification algorithms which achieve better performance are addressed, such as multi-decision fusion technology, the hybrid classification algorithm based on Bayesian and information gain, and neural network classification algorithm based on rough set and genetic algorithm etc. Finally, research emphasis in the future is discussed.

Key words: data mining; classification; review

(责任编辑 游中胜)