DOI:10.13546/j.cnki.tjyjc.2019.06.008

机器学习分类问题及算法研究综述

杨剑锋，乔佩蕊，李永梅，王　宁
（郑州大学 商学院，郑州 450001）

摘　要：分类问题及其算法是机器学习的一个重要分支，其应用越来越广泛，相关算法及应用研究取得了长足进展。文章对近年来机器学习分类算法的研究成果进行了回顾，从单一分类算法到集成分类算法分别进行总结，比较了不同分类算法的核心思想、优缺点以及实际应用，并分析了机器学习分类算法研究所面临的挑战和发展趋势。

关键词：机器学习；分类算法；单一分类算法；集成分类算法
中图分类号：TP181；F222.3　　**文献标识码**：A　　**文章编号**：1002-6487(2019)06-0036-05

0　引言

在人类的生产和生活中存在着各种分类问题，对分类方法的需求并不比回归方法少。分类方法已经得到广泛研究，如判别分析和Logistic回归等[1]。但是，传统分类方法的分类准确度有限，且应用范围较窄。随着互联网和大数据的发展，数据的丰富度和覆盖面远超出了人工可以观察和总结的范畴。结合了统计学、数据库科学和计算机科学的机器学习已成为人工智能和数据科学发展的主流方向之一。分类问题作为机器学习的一部分，成为了研究的重点。近年来，我国的机器学习分类算法相关研究发展迅猛，并广泛应用于实践。因此，对国内机器学习分类算法相关研究进行整理和评述，对学术研究以及实际应用都具有较大的指导意义。

1　机器学习及分类问题概述

机器学习(Machine Learning)是研究计算机如何模仿人类的学习行为，获取新的知识或经验，并重新组织已有的知识结构，提高自身的表现[2]。机器学习可以通过计算机在海量数据中学习数据的规律和模式，从中挖掘出潜在信息，广泛用于解决分类、回归、聚类等问题。机器学习一般包括监督、半监督、无监督学习问题。在监督学习问题中，数据输入对象会预先分配标签，通过数据训练出模型，然后利用模型进行预测。当输出变量为连续时，被称为回归问题，当输出变量为离散时，则称为分类问题。无监督学习问题中，数据没有标签。其重点在于分析数据的隐藏结构，发现是否存在可区分的组或集群。半监督学习[3]也是机器学习的一个重要分支。与标记数据相比，未标记数据较容易获得。半监督学习通过监督学习与无监督学习的结合，利用少量的标记数据和大量的未标记数据进行训练和分类。

机器学习算法最初多用于解决回归问题。近年来，分类问题的研究也越来越多。在机器学习中，分类通常被理解为监督学习，但无监督学习和半监督学习也可以获得更好的分类器。无监督分类是一种用来获取训练分类器标签或推导分类模型参数的方法[4]。半监督分类中的分类器构建既使用了标记样本又使用了未标记样本，逐渐成为了研究热点。本文主要讨论了监督分类问题中的算法。

从监督学习的观点来看，分类是利用有标记的信息发现分类规则、构造分类模型，从而输出未含标记信息的数据属性特征的一种监督学习方法[3]，其最终的目标是使分类准确度达到最好[5]。分类的实现过程[6]主要分为两个步骤(见下页图1)：一是“学习步”，即归纳、分析训练集，找到合适的分类器，建立分类模型得到分类规则；二是“分类步”，即用已知的测试集来检测分类规则的准确率，若准确度可以接受，则使用训练好的模型对未知类标号的待测集进行预测。

2　单一的分类算法

2.1　单一分类算法概述

2.1.1　ANN分类

ANN是一种模拟生物神经网络进行信息处理的数学模型，简称为神经网络。ANN是经典的机器学习算法。McCulloch和Pitts最早提出MP模型证明了单个神经元能执行逻辑功能。ANN分类根据给定的训练样本，调整人工神经网络参数，使网络输出接近于已知样本类标记。用

基金项目：国家自然科学基金联合项目(U1504703)；河南省软科学研究计划项目(172400410334)
作者简介：杨剑锋(1970—)，男，山东博兴人，博士，副教授，研究方向：智能制造、质量管理。
乔佩蕊(1993—)，女，甘肃白银人，硕士研究生，研究方向：智能制造、质量管理。
李永梅(1995—)，女，河南信阳人，硕士研究生，研究方向：智能制造、质量管理。
(通讯作者)王　宁(1983—)，男，辽宁葫芦岛人，博士，副教授，研究方向：质量管理。

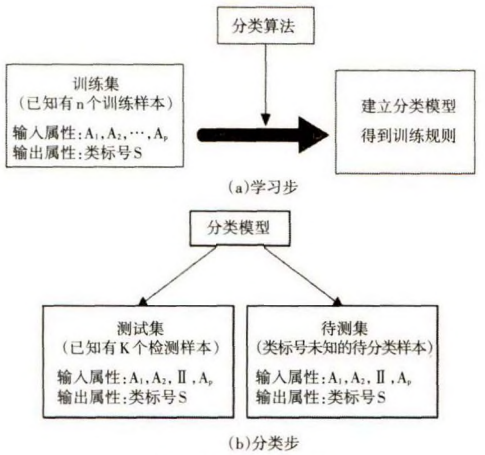

图1 分类的实现过程

于分类的ANN算法有BP神经网络、RBF径向基神经网络、FNN模糊神经网络、ANFIS自适应神经网络等，其中BP神经网络由于其良好的非线性逼近能力和分类能力得到了最广泛的应用[7]。

2.1.2 朴素贝叶斯分类

Maron和Kuhns（1960）以贝叶斯理论为基础，提出了依据概率原则进行分类的NB算法[8]。对于待分类样本，根据已知的先验概率，利用贝叶斯公式求出样本属于某一类的后验概率，然后选择后验概率最大的类作为该样本所属的类。NB改进算法主要有TAN算法、BAN算法、半朴素贝叶斯算法、贝叶斯信念网络等。

2.1.3 K近邻分类

Cover和Hart提出了基于距离度量的KNN分类算法[9]。KNN算法将整个数据集作为训练集，确定待分类样本与每个训练样本之间的距离，然后找出与待分类样本距离最近的K个样本作为待分类样本的K个近邻。待分类样本类别是占比最大的类别。KNN算法采用曼哈顿、闵可夫斯基以及欧式距离，其中欧式距离最常用。针对KNN算法的缺点，近邻规则浓缩法、产生或者修改原型法、多重分类器结合法等改进KNN算法被提出。

2.1.4 决策树分类

Breiman等提出了早期的决策树（DT）分类算法—CART算法，其使用树结构算法将数据分成离散类[10]。Quinlan引入信息增益提出了ID3算法和C4.5算法[11]。目前已发展到C5.0算法，其运行效率等得到进一步完善[12]。DT的改进算法还有EC4.5、SLIQ算法、SPRINT算法、PUBLIC算法等。决策树是一种倒置的树形结构，由决策节点、分支和叶子节点组成。DT分类算法一般有两个步骤：一是利用训练集从DT最顶层的根节点开始，自顶向下依次判断，形成一棵决策树（即建立分类模型）；二是利用建好的DT对待分类样本集进行分类[13]。

2.1.5 支持向量机分类

Cortes和Vapnik在1995年正式提出了支持向量机（SVM）。SVM是基于统计学的VC维理论与结构风险最小原理的有监督二分类器。根据给定训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots(x_l,y_l),\}\in(X,Y)^l$，寻找 R^n 上的一个实值函数，$g(x)$ 以便使用分类函数 $f(x)=\mathrm{sgn}(g(x))$ 推断任意一个模式x相对应的y的值。当数据线性可分时，SVM通过在原始特征空间中构建一个最优分割超平面，并将其作为决策面，最大化正负样本之间的边缘距离，采用训练集构建分类器对样本数据进行分类。当数据线性不可分时，SVM使用核函数将样本数据映射到一个高维空间，然后寻找一个最优分类超平面隔离不同类别样本数据，从而进行分类。近年来，发展出多种改进SVM算法，如GSVM、FSVM、TWSVMs、RS-VM等[14]。

2.2 *码单一分类算法的比较及应用*

五种单一分类方法的比较，如表1所示：

表1 五种单一分类方法优缺点对比

算法	优点	缺点
ANN	分类准确度高，学习能力强；对噪声数据鲁棒性和容错性较强；有联想能力，能逼近任意非线性关系；对未经训练的数据也具有较好的预测分类能力	参数较多（权值和阈值）；黑箱过程，不能观察中间结果，可解释性差；训练时间较长，有可能陷入局部极小值；不能直接利用神经网络生成规则，输入属性值必须是数值型
NB	训练和分类仅仅是特征概率的数学运算，分类速度快；支持增量式运算，可以对新增的样本进行训练；在对大样本进行处理时有很大的优势；对结果解释容易理解	使用样本属性独立性的假设，样本属性有关联时，会导致分类性能降低
KNN	对数据的分布无要求；直接使用训练集对数据样本进行分类，训练阶段较快	不建立分类模型，不易发现特征之间的关系；分类阶段需要逐个计算与训练样本的相似程度，计算量大且速度慢；数据不均衡时，预测偏差比较大；K值不易选择
DT	结构简单，可以可视化分析；容易提取出分类规则；适合处理量比较大的数据；可以同时处理标签型和数值型数据；测试数据集时，运算速度比较快；分类精确度较高	不易处理缺失数据；易出现过拟合；忽略了数据集中属性的相互关联；根据具有大量水平的特征进行划分时往往是有偏的
SVM	解决小样本、非线性问题；无局部极小值问题；可以很好的处理高维数据集；泛化能力比较强	对核函数的高维映射解释力不强，尤其是径向基函数；对缺失数据敏感；适用于二分类问题，对于多分类问题容易产生过拟合

ANN分类作为机器学习的重要方法被广泛应用于模式识别、故障诊断、图像处理、人脸识别和入侵检测等领域。近年来，深度神经网络由于其优异的算法性能逐渐成为了学术界的研究热点，已经广泛应用于图像分析、语音识别、目标检测、语义分割、人脸识别、自动驾驶等领域[15]。NB分类算法经常被用于文本分类，另外也被用于故障诊断、入侵检测、垃圾邮件分类等。KNN及其改进分类算法被大量应用于文本分类和故障诊断等领域，如判别粮食作物隐蔽性虫害[16]等。DT分类主要应用于遥感影像分类、遥感图像处理以及客户关系管理中的客户分类等领域，如地表沙漠化信息提取[17]、机械故障诊断[18]、人体行为的分类识别[19]等。SVM则主要用于二分类领域，在故障诊断、文本分类、模式识别、入侵检测、人脸识别等领域有广泛的应用。也扩展到了财务预警、医学以及机器人等领域。

3 集成分类算法

尽管单一分类方法取得了飞速发展，但实际中仍会遇到这些方法不能有效解决的问题。Hansen和Salamon提

出了新的机器学习方法——集成学习(Ensemble Learning)[20]。随着数据结构复杂、数据量大、数据质量参差不齐等问题愈加突出,集成学习成为了大数据分析的强有力工具。集成学习算法是通过某种方式或规则将若干个基分类器的预测结果进行综合,进而有效克服过学习、提升分类效果。集成算法按照基分类器是否存在依赖关系分为两类:基分类器之间没有依赖关系的Bagging系列算法和有依赖关系的Boosting系列算法。Bagging系列算法中用于分类的主要有Bagging算法和随机森林(Random Forest, RF)算法。对于复杂数据,集成分类算法通常优于单一分类方法,但预测速度明显下降,随着基分类器数目增加,所需存储空间也急剧增加。因此,选择性集成被提出,利用少量基本学习机进行集成提高性能[21]。鉴于篇幅限制,本文不讨论选择性集成分类算法。

3.1 Bagging系列算法

3.1.1 Bagging分类

Breiman最早提出Bagging方法[22]。其原理是,首先对原始训练集使用自助法抽样(Bootstrap Sampling)的方式得到多个采样集,然后用这些采样集分别对多个基分类器进行训练,最后通过基分类器的组合策略得到最终的集成分类器(见图2)。在分类问题中,Bagging通常使用投票法,按照少数服从多数或票要过半的原则来投票确定最终类别。

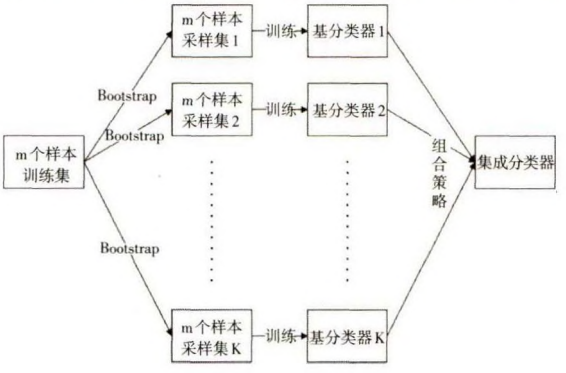

图2 Bagging算法流程

3.1.2 RF分类

随机森林(RF)算法是关注决策树的集成学习,由Breiman于2001年提出[23]。RF算法将CART算法构建的没有剪枝的分类决策树作为基分类器,将Bagging和随机特征选择结合起来,增加决策树模型的多样性。其原理是,首先从原始样本集中使用Bootstrap方法抽取训练集,然后在每个训练集上训练一个决策树模型,最后所有基分类器投出最多票数的类别或类别之一为最终类别。除此之外,还出现了一些RF的推广算法,如表2所示。

表2 RF的推广算法

算法名称	与RF的不同
Random Survival Forest(RSF)[24]	建树规则与RF类似,RSF中的每棵决策树都是二分类的生存树,用以处理生存数据,对于高维生存数据,其优于其他生存分析方法。
Extra trees[25]	RF的一个变种,与RF的区别:一般不采用自助法抽样,每个决策树都采用原始训练集,并且只随机的选择一个样本特征来划分决策树。
Isolation Forest[26](IForest)	用类似于RF的方法来检验异常值,与RF的区别:采用自助法抽样对训练集进行采样,但采样个数与RF(等于训练集个数)不一样,而是远远小于训练集个数;对于每个决策树的建立,采用随机选择一个划分特征,对划分特征随机选择一个划分阈值。

3.2 Boosting系列算法

Schapire和Freundfenbi最早提出了两种Boosting算法[27]。利用重赋权法迭代训练基分类器,然后采用序列式线性加权方式对基分类器进行组合。由于Boosting算法都要求事先知道弱分类算法分类正确率的下限,但实际中难以确定。Freund等基于Boosting思想进一步提出了AdaBoost算法[28]。其原理是,先给训练数据中每个样本赋予权重,并把样本权重初始化为相等值,训练得到第一个基分类器;通过计算错误率确定第一基分类器权重后,重新调整每个样本权重,增大被错分样本的权重,从而使被错分样本在下一次学习中能够尽可能正确分类。重复上述步骤,直至获得足够好的分类器。

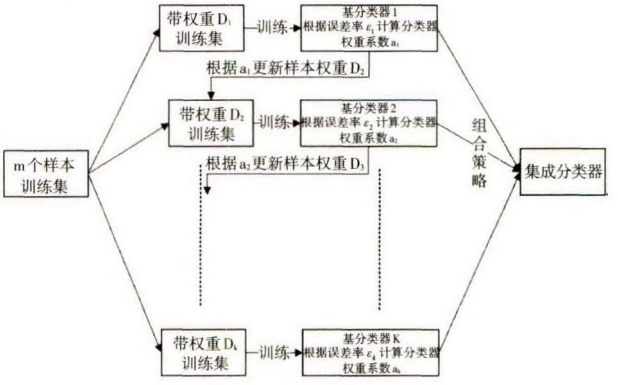

图3 Adaboost算法流程

改进的Adaboost算法有实Adaboost算法、LogitBoost算法、BrownBoost算法等[29]。近年来,由于AdaBoost.M1、AdaBoost.M2和AdaBoost.MH算法可用于解决多分类问题而受到了极大关注。此外,Friedman提出了Gradient Boosting算法,提出在前次建模的损失函数梯度下降方向进行建模,从而不断进行改进模型[30]。Adaboost算法和Gradient Boosting算法分别与决策树结合形成了提升树和梯度提升决策树(Gradient Boosting Decision Tree, GBDT)[31]。由于GBDT具有较强的泛化能力,适于多种分类问题,被越来越多地关注。

3.3 Bagging系列算法和Boosting系列算法的区别

Boosting与Bagging都是提高弱分类算法准确度的方法,但存在着一定区别(见下页表3)。Bagging、RF、Adaboost三种主要集成分类算法的优缺点也各不相同(见下页表4)。其中,RF和Bagging作为Bagging系列算法的不同在于:一是RF的基分类器都是CART决策树;二是RF在Bagging随机采样的基础上,又加上了特征随机采样。

Bagging算法主要被用于人脸识别和个人信用评估等领域,也被广泛应用于不平衡数据分类问题,如针对不平衡数据分类问题的基于Bagging组合学习方法[32]。RF作为一种优秀的非线性机器学习建模工具,广泛用于模式识别、图像分类、故障诊断等领域。Adaboost算法主要用于人脸检测、人脸识别、车俩检测、行人检测、目标检测、人眼检测、肤色分割等二分类或多分类问题。目前,决策树和神经网络是使用最广泛的Adaboost基分类器。

表3　　Bagging与Boosting的区别

算法名称	样本选择	样本权重	基分类器权重	并行计算
Bagging	从原始集中使用Bootstrap方法抽取训练集，选出的各轮训练集之间是独立的	每个样本的权重相等	所有基分类器的权重相等	可以并行生成
Boosting	每一轮的训练集不变，只是训练集中每个样本在分类器中的权重发生了变化	根据错误率不断调整样本的权重，并且错误率越大则权重越大	每个基分类器都有相应的权重	只能按顺序生成

表4　　三种集成分类算法优缺点对比

算法名称	优点	缺点
Bagging	每次都进行采样来训练模型，泛化能力强，模型方差较低；对噪声不敏感	训练集的拟合程度较差，即模型的偏倚比较大
RF	不需对数据预处理，能够很好地容忍噪声和异常值；对多元共线性不敏感，结果对缺失数据和非平衡数据比较稳健；只选择最重要的特征，适用于高维或大样本的情况以及高维小样本的问题；训练快速、分类准确高；模型方差较小，且泛化能力强	模型不容易解释；需要花费工夫使模型符合数据；如果某特征的取值划分较多，则其对RF的决策影响更大
Adaboost	可以将不同的分类算法作为基分类器；很好的利用了基分类器进行级联；分类精度较高；相对于bagging算法、RF算法，其充分考虑了每个分类器的权重；不容易发生过拟合	迭代次数(基分类器数目)不好设定；数据不平衡导致分类精度下降；训练时间较长

4　机器学习分类算法面临的挑战及展望

尽管机器学习分类算法可以处理很多复杂的分类问题，但随着数据变得更加复杂多样，机器学习分类算法在学习目标和分类效率方面遇到了新的挑战：

(1)高维小样本。不同应用领域的数据都呈现出高维度的特点。数据中的冗余、无关信息的增多，使得机器学习分类算法的性能降低，计算复杂度增加。机器学习分类算法一般需要利用大样本才能进行有效学习，大数据并不意味着训练样本数量充足。当样本量较小且特征中含有大量无关特征或噪声特征时，可能导致分类精度不高，出现过拟合。

(2)高维不平衡。机器学习分类算法一般假定用于训练的数据集是平衡的，即各类所含的样本数大致相等，但现实中数据往往是不平衡的。现有研究通常将不平衡问题和高维问题分开处理，但是实践中经常存在具有不平衡和高维双重特性的数据。

(3)高维多分类。除了常见的二分类问题，实际应用中存在着大量的多分类问题，尤其是高维数据的多分类问题，这给现有的机器学习分类算法带来了挑战。

(4)特征工程。目前的机器学习分类算法应用中的数据实例是由大量的特征来表示的。良好的分类模型依赖于相关度大的特征集合，剔除不相关和多余特征，不仅能提高模型精确度，而且能减少运行时间。因此，特征选择的研究对机器学习分类算法的发展越来越重要。

(5)属性值缺失。属性值缺失容易降低分类模型的预测准确率，是分类过程中一类常见的数据质量问题。正确解决分类过程中出现的属性值缺失是一个具有挑战性的问题。

机器学习是人工智能的重要组成部分，分类是其最重要的任务之一。通过讨论了不同机器学习分类算法的特点及应用，可以发现没有一种算法可以解决所有问题。此外，数据降维、特征选择将分类算法的发展产生更大的影响。因此，在实际应用中，必须结合实际情况比较和选择适当的分类算法和数据预处理方法以便更加有效地实现分类目标。在传统分类算法改进和发展的同时，集成学习将得到更广泛的应用和发展。

参考文献：

[1]吴喜之. 应用回归及分类:基于R [M]. 北京:中国人民大学出版社,2016.

[2]Etherm A. 机器学习导论[M].范明,昝红英,牛常勇译. 北京:机械工业出版社,2009.

[3]Zaslavsky G M. A Survey of Classification Methods in Data Streams [M]. Berlin:Springer, 2009.

[4]赵春霞,钱乐祥. 遥感影像监督分类与非监督分类的比较[J]. 河南大学学报(自然科学版), 2004, (3).

[5]Armitage W. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification[J]. Computer Communication Review, 2006, 30(1).

[6]李玲俐. 数据挖掘中分类算法综述[J]. 重庆师范大学学报(自然科学版), 2011, 28(4).

[7]张润,王永滨. 机器学习及其算法和发展研究[J]. 中国传媒大学学报(自然科学版), 2016, 23(2).

[8]Maron M E, Kuhns J L. On Relevance Probabilistic Indexing and Information Retrieval[J]. Journal of the ACM (JACM), 1960, 7(3).

[9]Cover T M, Hart P E.Nearest Neighbor Pattern Classification[J]. IEEE Transactions on Information Theory, 1967, 13(1).

[10]Breiman L, Friedman J, Olshen R A,et al.Classification and regression trees[M].Belmont:Wadsworth,1984.

[11]Quinlan J R.C4.5: Programs for Machine Learning [M].Morgan Kauffman, 1993.

[12]Thombre, A. Comparing Logistic Regression, Neural Networks, C5.0 and m5′ Classification Techniques [J]. Lecture Notes in Computer Science, 2012, 7376.

[13]Xie N, Liu Y. Review of Decision Trees[J].Computer Science and Information Technology (ICCSIT),2010, (5).

[14]丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J].电子科技大学学报,2011,40(1).

[15]朱虎明,李佩焦,李成等. 深度神经网络并行化研究综述[J].计算机学报,2018, 41(8).

[16]王锋,王艳娜,梁义涛等. 基于KNN算法的小麦隐蔽性虫害分类器设计[J].农机化研究,2014,36(7).

[17]吕利利,颉耀文,黄晓君等. 基于CART决策树分类的沙漠化信息提取方法研究[J].遥感技术与应用,2017, 32(3).

[18]徐翌,张斌.基于约简矩阵和C4.5决策树的故障诊断方法[J].计算机技术与发展,2018,(2).

[19]王忠民,张琮,衡霞. CNN与决策树结合的新型人体行为识别方法研究[J]. 计算机应用研究,2017, 34(12).

[20]Hansen L K, Salamon P. Neural Network Ensembles[J].IEEE Trans on Pattern Analysis and Machine Intelligence,1990, 12(10).

[21]张春霞,张讲社.选择性集成学习算法综述[J].计算机学报,2011,34

(8).

[22]Breiman L. Bagging Predicators[J]. Machine Learning,1996, 24(2).

[23]Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1).

[24]Ishwaran H, Kogalur U B, Blackstone E H, et al. Random Survival Forests [J]. The Annals of Applied Statistics,2008, 2(3).

[25]Desir C, Petitjean C, Heutte, L, et al. Classification of Endomicroscopic Images of the Lung Based on Random Subwindows and Extra-Trees [J]. IEEE Transactions on Biomedical Engineering,2012, 59(9).

[26]Liu F, Ting K, Zhou Z. Isolation Forest [C]. Proceeding of the 8th IEEE International Conference on Data Mining,2008.

[27]沈学华,周志华,吴建鑫等. Boosting和Bagging综述[J]. 计算机工程与应用, 2000, (12).

[28]Freund Y, Schapire R E. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting [J]. Journal of Computer and System Sciences,1997,55(1).

[29]于玲,吴铁军. 集成学习:Boosting算法综述[J].模式识别与人工智能,2004,17(1).

[30]Friedman J, Hastie T.Additive Logistic Regression:A Statistical View of Boosting(With Discussions) [J].Annals of Statistics,2000, (28).

[31]Ma X, Ding C, Luan S, et al. Prioritizing Influential Factors for Freeway Incident Clearance Time Prediction Using the Gradient Boosting Decision Trees Method[J].IEEE Trans on Intelligent Transportation Systems, 2017,99.

[32]秦姣龙,王蔚. Bagging组合的不平衡数据分类方法[J]. 计算机工程,2011,37(14).

(责任编辑/刘柳青)

A Review of Machine-learning Classification and Algorithms

Yang Jianfeng, Qiao Peirui, Li Yongmei, Wang Ning

(Business School, Zhengzhou University, Zhengzhou 450001, China)

Abstract: Classification and its algorithm are an important branch of machine learning, whose application is more and more extensive, and related algorithms and application research have made great progress. This paper firstly reviews the research results of machine-learning classification algorithm in recent years, and then makes a respective summary on single classification algorithm and integrated classification algorithm. The paper also makes a comparison of the core ideas, advantages and disadvantages and practical applications of different classification algorithms. Finally the paper analyzes the challenges that the research on machine-learning classification algorithm is facing and its future developing trend.

Key words: machine learning; classification algorithm; single classification algorithm; integrated classification algorithm