上海大学

SHANGHAI UNIVERSITY

课程论文

COURSE THESIS

题目：基于机器学习的分类算法研究综述

学院	计算机工程与科学学院
专业	计算机科学与技术
课程名称	计算机科学进展
学号	**20121802**
学生姓名	严昕宇

基于机器学习的分类算法研究综述

严昕宇
（上海大学 计算机工程与科学学院）

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

A Survey of Ensemble Learning Approaches

Yan Xinyu
(Shanghai University, School of Computer Engineering and Science)

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; ensemble classification algorithm

目录

1 引言 **3**

2 机器学习与分类问题 **3**

3 单一分类算法 **4**
3.1 人工神经网络分类 4
3.2 朴素贝叶斯分类 4
3.3 K最邻近分类 4
3.4 决策树分类 5
3.5 支持向量机分类 5
3.6 单一分类算法的比较及应用 6

4 集成分类算法 **7**
4.1 Bagging分类 7
4.2 随机森林分类 8
4.3 Boosting分类算法 8
4.4 Bagging与Boosting的区别 9
4.5 集成学习的应用 10
4.5.1 在时间序列上的应用 10
4.5.2 在医疗健康上的应用 11
4.5.3 在入侵检测上的应用 11

5 挑战与展望 **12**

6 总结 **12**

1 引言

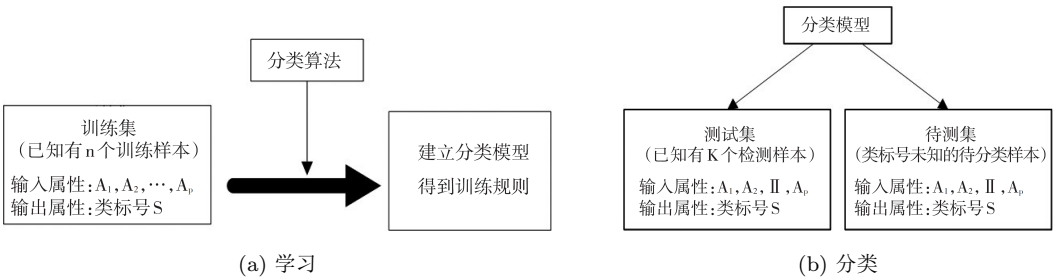
在人类的生产和生活中存在着各种分类问题，对分类方法的需求并不比回归方法少。分类方法已经得到广泛研究，如判别分析和 Logistic 回归等。但是，传统分类方法的分类准确度有限，且应用范围较窄。随着互联网和大数据的发展，数据的丰富度和覆盖面远超出了人工可以观察和总结的范畴。结合了统计学、数据库科学和计算机科学的机器学习已成为人工智能和数据科学发展的主流方向之一。分类问题作为机器学习的一部分，成为了研究的重点。近年来，机器学习分类算法相关研究发展迅猛，并广泛应用于实践。因此，对机器学习分类算法相关研究进行整理和评述，对研究以及实际应用都具有较大的意义。

2 机器学习与分类问题

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。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。机器学习作为人工智能的重要研究内容，经过近半个世纪的发展，现今已和模式识别、数据挖掘、统计学习、计算机视觉、自然语言处理等多个领域相互影响、交织发展，无论是理论上还是实践上都取得了巨大的进展，并广泛地应用于文本分类、语音识别、图像解译、医学诊断等多个领域。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。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，其最终的目标是使分类准确度达到最好。分类的实现过程主要分为两个步骤（如图1所示）：一是“学习”，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”，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。



(a) 学习 (b) 分类

图1：分类的实现过程

3 单一分类算法

从模型的个数和性质角度来看，可以将机器学习模型划分为单模型（Single Model）和集成模型（Ensemble Model）。所谓单模型，是指机器学习模型仅包括一个模型，基于某一种模型独立进行训练和验证。

3.1 人工神经网络分类

人工神经网络（Artificial Neural Network，ANN）是模仿生物神经系统中神经元的一种数学处理方法。ANN由大量处理单元（神经元）互联组成非线性大规模自适应动态系统，依靠系统的复杂程度，通过调整大量单元间相互连接关系，从而达到处理信息的目的。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。ANN模型的性能主要与神经元的特性、神经元之间相互连接形式以及为适应环境而改善性能的学习规则有关。ANN构成原理与功能更接近人脑，擅于适应环境、总结规律以处理模糊复杂、推理规则不明确的问题，不用考虑各变量之间是否独立及满足正态分布等条件，并且ANN可给出结构参数，这也与传统的统计分析不一样。而在此基础上进行算法拓展，进一步兴起了深度学习，它运用分层抽象的思想，被应用在图像识别、语音识别等领域，具有更高的识别精度。

然而，因ANN具有黑箱运行的特点，ANN在解释推理过程和推理依据及其存储知识的意义时存在一定困难。目前，常见的神经网络主要有向前神经网络、BP神经网络、卷积神经网络、Hopfield神经网络等。

3.2 朴素贝叶斯分类

Maron和Kuhns以贝叶斯理论为基础，提出了依据概率原则进行分类的朴素贝叶斯算法（Naive Bayesian Algorithm）。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。朴素贝叶斯方法是在贝叶斯算法的基础上进行了相应的简化，即假定给定目标值时属性之间相互条件独立。也就是说没有哪个属性变量对于决策结果来说占有着较大的比重，也没有哪个属性变量对于决策结果占有着较小的比重。虽然这个简化方式在一定程度上降低了贝叶斯分类算法的分类效果，但是在实际的应用场景中，极大地简化了贝叶斯方法的复杂性。朴素贝叶斯算法的改进算法主要有TAN算法、BAN算法、半朴素贝叶斯算法、贝叶斯信念网络等。

3.3 K最邻近分类

Cover和Hart提出了基于距离度量的K最近邻（K-Nearest Neighbor，KNN）分类算法。K近邻分类算法指假设一个样本在特征空间中的K个最相似(即特征空间中最近邻)的样本中大多数属于某一类别，则该样本也属于这个类别。该算法有3个基本要素：K值、距离度量和分类决策规则。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。其中K值的选择与结果密切相关，K值较小意味着只有与输入实例较接近的训练实例才会对预测结果有作用，但容易发生过拟合;而K值较大时，学习的估计误差减少，但近似误差增大，此时与输人实例较远的训练实例也会对预测起作用，可导致预测发生错误。一般来说，K值常选用一个较小数值，通常采用交叉验证方法选择最佳K值。该算法的特点为:首先，无需参数估计与训练，比较简单有效，精度高，且对噪声不敏感;其次，因KNN依靠周围有限的邻近样本，因而对类域交叉或重叠较多的待测样本来说，KNN相对更理想，也更适合多分类问题。

但KNN分类算法也存在着解释性较差、计算量大、当样本不均衡时可能导致结果偏差等问题。近年来，针对KNN算法的不足之处，衍生出了一系列新算法，如K-D树KNN算法、快速KNN算法、ML-KNN算法等。其中，ML-KNN模型为近年来典型的多标签学习算法，具有简单、易行、错误率较低等优点，是在传统的单标签KNN模型基础上结合贝叶斯算法发展起来。

3.4 决策树分类

决策树（Decision Tree，DT）是一种相对简单的机器学习分类算法。其构建思想来源于人们的决策过程，在已知各种情况发生的概率基础上，通过构成决策树来评价项目风险，判断可行性的决策方法。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。因这种决策分支构成的图形很像一棵树干，故称为决策树。树中每个节点表示一个样本属性，每个分支则代表对该属性的判断，而每个叶子结点则对应最终的类别，通常将其看作一个预测模型，代表对象属性与对象值之间的一种映射关系]。该模型易于理解和实现，用户即使未学习相关背景知识，也能发掘其简单直观的分类规则，只要通过适当的解释，用户就能理解决策树所表达的意义。其次，算法运行速度快，易转化为分类规则，分类准确率较高，只要沿着根结点向下一直走到叶子结点，沿途分裂条件是唯一且确定的。但决策树在处理大样本集时，易出现过拟合现象，从而降低分类的准确性。

决策树分类算法一般有两个步骤：一是利用训练集从决策树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）；二是利用建好的决策树对待分类样本集进行分类。

Breiman等提出了早期的决策树（Decision Tree）分类算法——CART算法，其使用树结构算法将数据分成离散类。Quinlan引入信息增益提出了ID3算法和C4.5算法。目前已发展到C5.0算法，其运行效率等得到进一步完善。决策树的改进算法还有EC4.5、SLIQ算法、SPRINT算法、PUB-LIC算法等。

3.5 支持向量机分类

支持向量机（Support Vector Machine，SVM）是一种基于统计学的VC维（Vapnik-Chervonenkis Dimension）理论与结构风险最小原理的有监督二分类器，由Cortes和Vapnik在1995年正式提出。SVM是机器学习领域若干技术集大成者，它具有严格的理论和数学基础，可较好地实现结构风险最小化思想，能较为合理地解决小样本、非线性、高维数和局部最小等实际问题。SVM的关键在于针对样本数据不可分情况，利用核函数把一个复杂的分类任务映射，使之能转化成一个线性可分问题。

支持向量机分类算法如下：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$f(x) = sgn(g(x))$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。

SVM主要的优势在于：（1）专门针对有限样本设计，目标是获得现有信息下的最优解，根据有限的样本信息在模型的复杂性及学习能力之间寻求最佳折衷。（2）算法将实际问题通过非线性变换映射到高维特征空间，并构建线性最佳逼近来解决原来空间的非线性逼近问题，这样既保证了机器学习取得最好推广能力（泛化能力），并且SVM的算法复杂性与数据维数无关，因而能较好解决维数灾难问题。但是SVM主要解决二分类问题，对多分类问题存在困难，此时可考虑通过多个二类支持向量机的组合或构造多个分类器的组合来解决。近年来，发展出多种改进支持向量机算法，如GSVM、FSVM、TWSVMs、RS-VM等。

3.6 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。朴素贝叶斯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等。决策树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、机械故障诊断、人体行为的分类识别等。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。

4 集成分类算法

尽管单一分类算法取得了飞速发展，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）。

实际上，通过集成学习思想进行决策在文明社会开始时就已经存在了，例如：在民主社会中，公民们通过投票来选举官员或制定法律，对于个人而言，在重大医疗手术前通常咨询多名医生。这些例子表明，人们需要权衡并组合各种意见来做出最终的决定。其实，研究人员使用集成学习的最初目的和人们在日常生活中使用这些机制的原因相似。Dietterich从数学角度解释了集成方法成功的3个基本原因：统计、计算和代表性。此外，亦可通过偏差方差分解对集成学习的有效性进行分析。

基于以上所述，1979年，Dasarathy和Sheela首次提出集成学习思想。1990年，Hansen和Salamon展示了一种基于神经网络的集成模型，该集成模型具有更低的方差和更好的泛化能力。同年，Schapire证明了通过Boosting方法可以将弱分类器组合成一个强分类器，该方法的提出使集成学习成为机器学习的一个重要研究领域。此后，集成学习研究得到迅猛发展，出现了许多新颖的思想和模型。1991年Jacobs提出了混合专家模型。1994年，Wolpert提出了堆叠泛化模型。1995年，Freund和Schapire提出了Adaboost算法，该算法运行高效且实际应用广泛，该算法提出后，研究人员针对该算法进行了深入的研究。1996年，Breiman提出了Bagging算法，该算法从另一个角度对基学习器进行组合。1997年，Woods提出了一种动态分类器选择方法。2001年，Breiman提出了随机森林算法，该算法被誉为最好的算法之一。

随着时代的发展，更多的集成学习算法被提出，并且在诸多领域都取得了重大突破。由于数据结构复杂、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。

4.1 Bagging分类

Breiman最早提出Bagging方法。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。

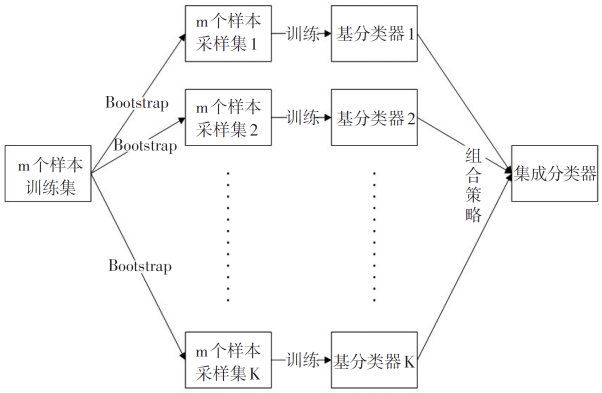


图2：Bagging算法流程图

该算法中，由于使用自助采样法来产生新的训练子集，一些实例会被多次采样，而其他实例会被忽略，因此，对于特定的子空间，个体学习器会具有很高的分类精度，而对于那些被忽略的部分，个体学习器难以正确分类。但是，最终的预测结果是由多个个体学习器投票产生的，所以当个体学习器效果越好且它们之间的差异越大时，该集成算法的效果就会越好。由于不稳定的学习算法对于训练集比较敏感，训练集只要产生一些微小的变化，就会导致其预测结果发生很大的改变，所以Bagging算法对于不稳定学习算法非常有效。 Bagging算法适合于解决训练集较小的问题，但对于具有大量训练集的问题，其效果就会下降，因此Breiman基于Bagging设计了Pasting Small Votes算法，该算法能够有效地应对数据量较大的机器学习问题。与此同时，由于Bagging训练得到的个体学习器是强相关的，因此Bagging方法在这种情况下通常表现较差。

4.2 随机森林分类

随机森林（Random Forest）算法是关注决策树的集成学习，由Breiman于2001年提出。随机森林算法将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，还出现了一些随机森林的推广算法，如表2所示。

表2：随机森林的推广算法

算法名称	与RF的不同
Random Survival Forest(RSF)	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	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 (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。

4.3 Boosting分类算法

Boosting算法是一种将弱学习器转换为强学习器的迭代方法，它通过增加迭代次数，产生一个表现接近完美的强学习器。其中，弱学习器是指分类效果只比随机猜测效果稍好的学习器，即分类准确率略高于50%。在实际训练中，获得一个弱学习器比获得一个强学习器更加容易。因此，对于Boosting系列算法的研究意义非凡。Boosting算法除了具有良好的实际性能外，还具有强大的理论基础和算法特点。

Schapire和Freundfenbi最早提出了两种Boosting算法。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。其原理是，每一个分类器都是针对前一次未被正确分类的样本进行学习，因此该算法可以有效地降低模型的偏差，但随着训练的进行，整体模型在训练集上的准确率不断提高，导致方差变大，不过通过对特征的随机采样可以降低分类模型间的相关性，从而降低模型整体的方差。当主分类器不能被信任，无法对给定对象进行分类时，例如，由于其结果中的置信度低，则将数据输送到辅助分类器，按顺序添加分类器。

下面将对Boosting算法的流程进行简单介绍。Boosting算法反复运行一个弱学习器来处理不同分布的训练数据，然后按照顺序将每次产生的弱学习器组合成一个复合强学习器，如下图所示：

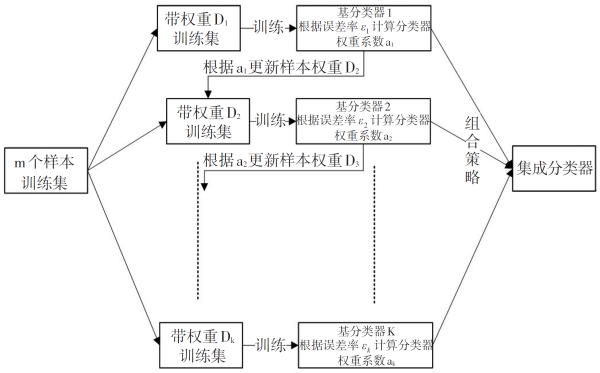


图3：Boosting算法流程图

改进的Adaboost算法有实Adaboost算法、LogitBoost算法、BrownBoost算法等。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。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）。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。

4.4 Bagging与Boosting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（见下表）。Bagging 通过采样得到多个训练子集，由于各个训练子集相互独立，降低了基分类器的方差，改善了泛化误差，并且重采样方法可以有效降低原始训练集中随机波动导致的误差，使得不稳定的学习器具有更好的学习效果因为该算法中的基学习器权重相同，所以基学习器的选择会直接影响集成结果，不稳定的基学习器不仅能提供良好的学习效果，而且能根据训练集的不同产生多样性，因此Bagging与不稳定的学习算法相结合通常能产生一个强大的学习模型，并且具有良好的抗噪能力，而且各个基学习器可以并行生成，提高运行效率Boosting在每轮训练中使用的训练集不变，但训练集中每个样例会根据上一轮的学习结果进行调整，使新学习器针对已有学习器判断错误的样本进行学习这种方法能够显著提高弱学习器的学习效果，但很容易受到噪声的影响产生过拟合现象，并且每个基学习器只能顺序生成，训练效率相对较差。

Bagging、随机森林、Adaboost三种主要集成分类算法的优缺点也各不相同（见下表）。其中，随机森林和Bagging作为Bagging系列算法的不同在于：一是随机森林的基分类器都是CART决策树；二是随机森林在Bagging随机采样的基础上，又加上了特征随机采样。

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

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。随机森林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。

4.5 集成学习的应用

时间序列分析、医疗健康和入侵检测三大领域与人们生活息息相关，因此被研究人员广泛关注。但由于这3个领域都具有数据维度高、数据结构复杂和特征模糊等特点，难以进行人工分析与处理，因此机器学习方法的引入使得3个领域的研究取得了突破。在各种机器学习方法中，集成学习作为一种可以最大化提升学习效果的技术，推动了诸多领域的快速发展，因此集成学习也被广泛应用于这3个领域并取得了良好的效果。本文将集成学习在这3个领域中的应用进行了概述。

4.5.1 在时间序列上的应用

集成学习被广泛用于预测天气、电气设备和化工装置的损耗，卫星通信和雷达数据等时间序列。时间序列预测与普通预测的区别在于，随着时间的推移，可能会在概念或分布上发生一些偏差。处理这些问题最有效的方法是使用多个学习器共同决策，因为多个学习器相比于单个学习器具有更好的泛化能力。在时间序列预测方面，Md M Islam 等提出了CNNE（Cooperative Neural Network Ensemble）算法，该算法通过使用一个特定的构造方法来决定神经网络中隐藏节点个数，以此保证每个神经网络的准确性，并且使用基于负相关的增量训练法来获取不同周期的神经网络，从而确保个体神经网络的多样性。D Kim等提出了一种GFPE（Genetic Fuzzy Predictor Ensemble）算法，该算法分为两阶段，第一阶段生成一个尽可能涵盖多的训练样例的模糊规则库，第二阶段建立微调隶属函数来降低预测误差，最后加权组合每个模糊预测器的结果进一步的减少预测误差。Assaad等将boosting和RNN相结合，进一步提升了R、RNN在时间序列预测方面的效果。熊志斌提出了一种基于ARIMA和神经网络集成的GDP时间序列预测算法，该集成算法优于单一模型并且能够有效用于GDP预测时间序列分类方面，J Neugebauer等设计出了一个通用的级联分类模型，它可以将高维分类任务分解为一系列低维任务，因此它不仅适用于时间序列，也可以用于处理其他具有复杂数据结构的数据集。A Bagnall等提出了一种将监督式集成学习和转换数据空间相结合的方法，该方法首先将原始时间序列在时间、频率、变化和shapelet四个领域上进行转换，然后再使用监督式集成学习方法生成模型。

时间序列聚类方面，Y Yang等提出了一种结合RPCL 网络和集成学习的方法，该方法用多个RPCL网络分别聚类不同表征学习后的时间序列，从而产生多个竞争学习的结果并将其组合，然后使用几个共识函数（Consensus Function）在组合结果上学习，最后采用一种最优的选择函数来输出结果。针对基于隐马尔科夫的时间数据聚类问题，一种双加权集成方法和混合元聚类集成方法被提出，这两种方法都能有效地解决模型初始化和模型选择的问题。

4.5.2 在医疗健康上的应用

医疗领域数据通常具有数据量大、数据质量不高和数据不精确等特点．然而，用于医疗诊断的专家系统需要具有高度精确性，因为其结果将直接关系到患者的生命，现有的专家诊疗系统有:计算机辅助诊断系统（Computer Aided Diagnosis，CAD）、临床决策支持系统（Clinical Decision Support System，CDSS）等．因此，在该领域利用集成学习方法集成多个“专家”进行决策的系统相比只根据单个“专家”进行决策的系统具有更高的准确性．Y Li等针对帕金森病设计出一种基于语音数据的分类算法。B Lssac等提出了一种基于特征选择和数据分类的集成学习方法，此方法用于癌症诊断和评分。D Subramanian等在预测心力衰竭方面提出了一个结合时间序列测量方法和多元逻辑回归集成模型的方法。Guo Haoyan等设计了一种针对肺结节检测的多层次特征选择和集成学习的计算机辅助诊断系统。李霞等提出了一种基于递归分类树的特征基因选择的集成方法，该方法不仅能够寻找疾病相关基因，而且能显著提高疾病的分类准确率。A C Tan等将Bagging、Boosting和C4.5决策树算法相结合，在癌基因表达谱中利用微阵列分析对癌症基因组数据进行分类。并且，集成学习在生物信息方面也具有广泛的应用。

4.5.3 在入侵检测上的应用

互联网，手机，电子商务，基于PC的通信和信息系统已成为日常生活的一部分。这些系统的广泛使用使通信变得更加方便快捷，增加了数据传输和信息共享量，并提高了生活质量。由于通信系统在许多领域都有使用，因此当它们受到诸如病毒，蠕虫，特洛伊木马等各种攻击的困扰时，便会对日常生活造成很大的影响。关于发现攻击并消除其影响的研究被称为入侵检测系统（IDS）。可将IDS研究看作是将网络的正常行为与攻击行为分开的分类任务。机器学习和数据挖掘算法在IDS研究中被广泛使用，其主要目的是解决IDS泛化能力差的问题。由于攻击者会不断改变并提高他们的能力，所以对IDS的研究将是永无止境的。P Mehetrey等研究了一种基于集成学习方法的协同入侵检测系统，它允许使用未开发的云资源来检测云系统上的各种类型的攻击。谷雨等针对于新的入侵行为，提出了一种集成PCA、ICA和增量式支持向量机的分类系统。P Sornsuwit等针对于U2R和R2L这两种难以检测的网络入侵攻击，采用AdaBoost算法创建一个可以保护安全性并提高分类器性能的模型。SChebrolu等提出了一种基于集成学习方法的混合体系结构，它将不同的特征选择算法组合，通过确定重要的输入特征，从而建立一个高效的入侵检测系统。G Giacinto等针对网络流量中不同的协议与服务，提出了一种基于模块化多分类器系统（MCS）的无标签网络异常入侵检测系统，其中每个模块都会模拟一组特定的网络协议或者服务类型。

5 挑战与展望

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。

6 总结

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。

人类做出重大决定前会寻求多种意见来辅助决策，集成学习算法就是模仿这种行为而产生的。七十年代后期，模式识别、统计学和机器学习等学科的研究人员开始对集成学习方法进行研究。随着研究热情不断增长，并且对于集成学习的研究不断深入，多种集成学习方法被提出并被广泛应用于各个领域。集成学习通过结合多个学习器来为各种机器学习问题提供解决方案，其模型能够解决很多单一模型无法解决的问题。由于大部分集成学习算法对基础学习器的类型没有限制，并且它对于诸多成熟的机器学习框架都具有良好的适用性，因此集成学习也被称为“无算法的算法”。因此，在传统分类算法改进和发展的同时，集成学习也将得到更广泛的应用和发展。

参考文献

[1] 徐继伟,杨云.集成学习方法:研究综述[J].云南大学学报(自然科学版),2018,40(06):1082-1092.

[2] 梁云,门昌骞,王文剑.基于模型决策树的AdaBoost算法[J/OL].山东大学学报(理学版):1-9[2022-05-12].http://kns.cnki.net/kcms/detail/37.1389.N.20220419.1258.004.html

[3] 曹莹,苗启广,刘家辰,高琳.AdaBoost算法研究进展与展望[J].自动化学报,2013,39(06):745-758.

[4] 吕红燕,冯倩.随机森林算法研究综述[J].河北省科学院学报,2019,36(03):37-41.

[5] 尤璞,刘星甫.Stacking集成学习在销售预测中的应用[J].软件导刊,2022,21(04):103-108.

[6] 王琦琪,戴家佳,崔熊卫.基于集成学习模型的糖尿病患病风险预测研究[J].软件导刊,2022,21(04):62-66.

[7] 夏淑洁,杨朝阳,周常恩,辛基梁,张佳,杜国栋,李灿东.常见机器学习方法在中医诊断领域的应用述评[J].广州中医药大学学报,2021,38(04):826-831.DOI:10.13359/j.cnki.gzxbtcm.2021.04.032.

[8] 肖梁,韩璐,魏鹏飞,郑鑫浩,张上,吴飞.基于Bagging集成学习的多集类不平衡学习[J].计算机技术与发展,2021,31(10):1-6.

[9] 张晓龙,任芳.支持向量机与AdaBoost的结合算法研究[J].计算机应用研究,2009,26(01):77-78+110.

[10] 涂承胜,刁力力,鲁明羽,陆玉昌.Boosting家族AdaBoost系列代表算法[J].计算机科学,2003(03):30-34+145.