

外文翻译稿

**题目： 大数据流分类技术文献综述与分析**

**学 院** 启迪数字学院

**专业班级** 软件工程Q177

**学生学号** 201709601295

**学生姓名** 周静茹

**指导教师** 李 明

**日 期** 2020.12.11～2021.4.30

**评阅成绩**

**指导老师签名：**

**批改日期：**

# 摘要

技术和信息的快速发展，使人类见证了数据速度、数据量、数据种类的提高增长。商业机构的数据展示了大数据应用的发展。由于应用需求的提高，对复杂的流式大数据的分析往往成为数据挖掘的一个重要领域。其中一个重要的研究方面就是采用深度学习的方法来有效提取复杂的数据表示。因此，本调查对大数据分类方法进行了详细的评述，如基于深度学习的技术、基于卷积神经网络(CNN)的技术、基于K-最近邻(KNN)的技术、基于神经网络(NN)的技术、基于模糊的技术和基于支持向量的技术等。此外，还对参数进行了详细的研究，如评价指标、实现工具、采用的框架、利用的数据集、采用的分类方法以及各种技术获得的精度范围等。最后，提出了各种大数据分类方案的研究空白和问题。

**关键词：**分类技术，分类数据集，分类方案，分类框架

## 1.引言

万维网(WEB)的快速完善，互联网数据不断上升，成为大数据。根据大数据的各种特性，大数据分类利用在网络带宽、安全过滤网络数据管理、分类管理、网络信誉管理、绿色互联网等方面。大数据是用速度、数量和种类来描述的，[57]。如果加强数据的种类、速度和数量，则利用最新的技术来管理数据的处理以及存储。大数据分析这个名词并没有什么，只是理解和分析海量数据集的特征，以提取统计和几何模式的过程[56]。大部分产生的数据原本是流式数据。这是因为代表行动、测量和交互的数据，来自于互联网。数据是由一个时间间隔产生的。在流式框架下，高速数据，和算法必须处理非常严格的时间和空间的限制。流式算法利用数据结构来给出速度，和最佳答案[44]。

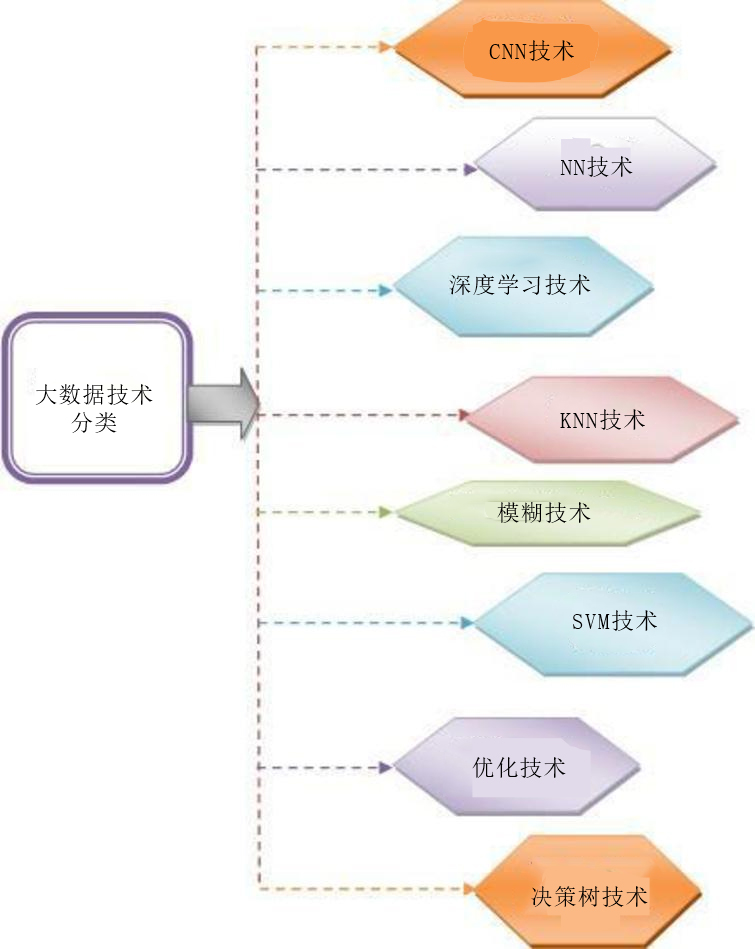


图1-1 大数据技术的分类

在市场营销、生物医学和社会化媒体的应用中，大数据分类是有益的任务之一。为解决大数据的挑战，通常采用的模型是单一的传统分类模型[58]。用于大数据分类的分类器有奈夫贝叶斯（NB）、KNN、SVM等。NB分类器广泛应用于信息。大数据分类中包含了数据融合机器分析[11]，NB用于云计算中的分类和目标跟踪[12]以及机器人控制[13]。大数据分析需要图像和文本，对于网络分析则包括弹性学习和可扩展方法[14，10]。从本质上看，KNN分类器是一种慢速分类器，且没有固定大小的小训练模型进行测试[9]。SVM作为二元分类器被采用，但现在，多类SVM(MSVM)被用来将向量分类为不同的集合，与使用训练好的口令，也被广泛研究[5]。在SVM中，训练活动是通过引入一个最优超平面来执行的[7]。

本研究的主要意图是对各种大数据分类方法进行详细调查。本次审查审议了不同研究工作中采用的现有方法，通过分析研究作品中提供的采用的框架、实施工具、评价指标等进行调查。 此外，采取所采用的框架和精度范围对大数据分类方法进行性能评价。此外，对所建议的大数据分类方法进行性能评价时，也采取了所采用的框架和精度范围。对现有的方法进行了明显的分类，然后，对研究的空白和问题进行了进一步的调查。 因此，它作为未来扩展有效的大数据管理的动力。

本文的组织结构为。第1节简要介绍了本文，第2节对现有分类方法进行了文献回顾，第3节讨论了研究的不足和问题。第4节对调查情况进行分析和讨论，第5节是本文的结尾。

## 2.不同大数据分类方法的文献调查

本节广泛讨论不同大数据分类技术的调查。不同的大数据分类系统的分类图解如图。1. 大数据分类技术分为八种，即基于CNN的技术、基于NN的技术、基于深度学习的技术、基于KNN的技术、基于模糊的技术、基于SVM的技术、基于优化的技术和基于决策树的技术。通过对大数据分类技术的回顾，可以清晰直观地了解所建议的方法及其意义和缺点。

## 2.1大数据分类技术的分类情况

本节阐述了采用各种大数据分类技术的特色研究工作。

## 2.1.1基于CNN的分类技术

下面对在大数据分类系统中采用基于CNN技术的研究论文进行论述。

Plathottam等人[16]提出了针对庞大数据集的CNN。这里，CNN被用于控制复杂的电网。CNN由电力数据进行训练，用于执行来自中洲ISO（MISO）的多类多标签分类。Tensor Flow被用来构建CNN和训练网络。

Qin等人[29]开发了CNN，从原始数据中控制特征，减轻外部资源以及工具箱的应用。利用实体标签特征(ETF)来指示目标标称的位置，虽然简单，但比位置特征(PF)更有效。Dropout策略用于缓解过拟合问题。

Cui等[32]在单一框架下开发了多尺度卷积神经网络(MCN)，用于特征的提取和分类。基于可学习的卷积层和多分支层，MCNN可以自动提取不同频率和尺度的特征进行特征表示。MCNN的计算效率很高，因为它天然地利用了图形处理单元（GPU）计算。然而，该方法没有使用来自各种来源的其他侧信息的时间序列分类，如图像、文本和语音。Deepak Puthal等人[43]从同步质数开发了基于动态密钥长度的安全框架（DLSeF）。在这里，密钥以很短的间隔动态更新，以确保端到端安全。使用CNN模型进行训练，从白天的图像中预测夜间的灯光，同时学习特征用于预测贫困。

Dong[46]从电子病历(EMRs)出发，提出了基于CNN的多类分类方法。该框架包括两个阶段。在第一阶段，用词嵌入来表示样本，进行预处理；在第二阶段，将训练数据分割成不同的子集，进行训练。第二阶段，将训练数据分割成不同的子集，并对每个子集进行基于CNN的训练。

## 2.1.2基于NN的分类技术

下面介绍了采用基于NN技术的大数据分类系统的独特研究工作，Shan等[21]解决了随机优化方法，并将Shuffled Frog-Leaping Algorithm(SFLA)应用于基于NN的分类器的参数优化中。初步开发了NN分类器，并与SVM进行比较。NN对于庞大的数据库是必不可少的，对于高层次数据的提取有一定难度。然后，建立数据集，涵盖语音和癌症数据。这两个数据库都有大量的样本数据，低级方差复杂。之后，基于修改版的SFLA对NN参数进行优化。

Sharma和Mangat[31]开发了基于数据挖掘方法的相关性向量机(Relevance Vector Machine，RVM)，利用埃博拉病毒的流行性爆发进行大数据分类，并与其他流行性疾病进行比较，基于RVM分类器进行泛化误差和类内分离。Dai等[50]开发了并行随机化NN进行分类。在这个框架中，引入了对RNN并行化的优化，以减少Shuffle阶段的通信开销，如基于缓存的优化、分区，从而提高了性能和可扩展性。

## 2.1.3 基于深度学习的分类方法

Alsheikh等[2]提出了基于火花的大数据深度学习的可扩展方法。深度学习使用一种工具从原始移动大数据中增加价值。特别是在多个spark worker上实现了基于Map Reduce的分布式深度学习。每个worker都能够在整体移动的一个分区上学习一个局部的深度模型，并为所有局部模型建立一个主深度框架。

Li等[6]利用张量表示法开发了深度卷积计算模型，用于从向量空间到张量空间的大数据学习特征。利用基于张量的方法来表示每个异质对象，以揭示隐藏的相关性。张量CNN使用当前的CNN来共享参数。此外，还设计了一个张量全连接层和一个张量池层，用于学习高阶特征。建立了反向传播的方法来训练深度卷积计算模型的参数。

双投深度计算模型（DPDCM）由Zhang等人[8]开发，用于特征学习。DPDCM在双投影层的基础上，投射出非线性子空间用于学习交互特征，以替换各层。然后，利用权重传感器进行映射，对交互进行建模。之后，引入反向传播的方法来训练DPDCM的参数。最后，建立了隐私保护双深计算模型（PPDPDCM），以防止隐私信息的泄露。

Deng等[13]开发了模糊深度神经网络(Fuzzy Deep Neural Network，FDNN)，用于从神经和模糊表示中提取数据。从这两种观点获得的知识被合并形成最终的表示。在这里，模糊表示减轻了不确定性，神经表示消除了原始数据中的噪声。然后，所开发的方法使用两种最佳表示法进行最终分类。因此，FDN对于应用数据噪声，和模糊性是必不可少的。

Ravi等人[14]开发了数据分类的深度学习方法。最初，流水线从惯性传感器收集原始数据。然后，将输入数据提取成片段，其中，来自浅层学习和深度学习方法的特征被并行估计。最后，将这两个特征集组合在一起，并基于软最大和全连接层进行分类。

Long等人[18]设计了深度适应网络(Deep Adaptation Network，DAN)，以增强特征转移性。嵌入隐藏表示以产生内核Hilbert空间，在该空间中对各种域分布的嵌入进行分组。然后基于最优多核选择方法进行平均嵌入匹配的域差异最小化。

Xiong等人[24]研究了地球科学数据的深度学习方法。这种方法被用来从大量的输入数据中学习有意义的模式，以绘制中国福建西南金属有机区的矿物图。该框架对于从原始数据中发现检测和分类所需的困难特征具有相关性。

Koyamada等人[25]开发了基于DNN的主题转移解码器。应用大规模功能磁共振成像(fMRI)数据库，以达到比其他基线方法最大的解码精度。为了解决基于解码器获得的知识，应用主灵敏度分析(Principal Sensitivity Analysis，PSA)来可视化数据集中所有主体的判别特征。

Zou等人[26]模拟了腾讯深度学习方法，用于训练、简化和支持腾讯的大型模型。Mariana包括三个框架。它们分别是多GPU模型并行和数据并行框架、大规模DNN的CPU集群框架。

Kuang和He[30]提出了用于特征提取和分类的DBN。该框架在区分注意力缺陷多动障碍(ADHD)的亚型和对照方面是有效的。在ADHD-200比赛中发表的结果，提高了准确性。

Shafiq和Torunski[49]提出了结构和组织日志的形式模型。然后，贝叶斯技术被用于预测和检测任何可能的故障。同时，引入了基于Map Reduce的分布式、单次递增的方法来构建、训练和实现所开发的框架。

## 2.1.4 基于KNN的分类技术

利用基于KNN的方法进行大数据分类系统的研究论文阐述如下。

Maillo等人[3]开发了用于大数据分类的KNN分类器。在这里，map阶段在数据的不同分割中确定k-NN。然后，减少阶段用于计算地图阶段所获得的列表中的确定邻域。这种方法采用KNN进行扩展，只需增加更多的计算节点。这种执行方式涉及到在原有KNN模型的基础上获得更好的分类率。

Ramírez-Gallego等人[5]采用apache spark的KNN分类器处理海量高速数据流。这种方法是针对高速、大规模和流媒体问题设计的。该方法基于分布式度量树组织实例，其中包含用于查询路由的顶层树和用于执行并行搜索的一组分布式子树。然后，介绍了实例选择方法，用于更新和淘汰案例库中过时的实例。

Hassanat[9]开发了KNN分类器，用于将训练实例插入到二进制搜索树中，以加快文本样本的搜索时间。这里，利用两种方法对训练实例进行排序。最主要的是基于规范的二元树(NBT)，用于计算最大或最小缩放规范。具有1-norm的实例被安排在右边的子节点，而具有0-norm的实例被安排在左边的子节点。这个过程重复进行，直到得到叶子节点中的相似规范。次要的是改进的NBT。在这里，每个例子被插入到搜索树取决于最大，和最小的欧氏规范。最后，顺序KNN分类器用于对特征向量进行分类。

Popescu和Keller[54]提出了大数据分类的随机投影模糊KNN（RPFKNN）融合策略。该方法是利用类成员的值引入的，是在每个投影中利用分类精度和FKNN实现的。

## 2.1.5 基于模糊的分类方案

下面就采用基于模糊技术的大数据分类技术的研究工作进行讨论。

Fernandez等[17]设计了基于模糊规则的进化分类系统（FR-BCSs）进行分类。它由Chi-FRBCS-Bi Data方法的人构建的知识库(KB)，以及基于2-tuples表示的数据库的遗传调整组成。然后，将每个地图的知识库合并在一起，产生一个集合分类器。

Iniguez等[19]提出了用于分类的Chi大数据支持过滤。最初，考虑分类精度和几何平均值。然后，使用规则库大小分析系统的可解释性。最后，根据经过的训练时间进行可扩展性分析。

El Bakry等人[20]开发了基于Map Reduce的模糊KNN分类器进行分类。该方法由映射器和还原器组成。映射器用于将数据集划分为块，并生成中间记录。这些记录是利用 "key，dat "集形式的map函数产生的。Mapper执行计算过程并将结果传递给reducer。然后，从reducer获得的输出，实现最终的输出。

Akil Kumar等人[22]提出了用于解决大数据中问题的FRBCS，称为Chi-FRBCS-BigData。这种方法使用可解释的模型来管理海量数据，并在Hadoop框架下以最佳的性能时间提供准确性。

Segatori等[28]开发了基于Map Reduce框架的分布式模糊决策树（FDT），从大数据中生成多路径和二元方式的FDT。该方法采用分布式模糊慎密化，利用模糊信息为每一个重复的属性产生模糊分区。模糊分区利用FDT学习在决策节点选择属性。

Azar和Hassanien[36]开发了Linguistic Hedges Neuro-Fuzzy Classifier with Selected Features(LHNFCSF)，用于降维、选择特征和分类。在这个框架中，连接权重、传播和激活函数与NN不同。

## 2.1.6 基于SVM的分类方法

下面讨论了实践基于SVM技术的大数据分类的研究工作。

Rebentrost等人[1]采用优化的二元分类器进行分类。SVM的最小平方公式允许使用相位计算和量子矩阵反演方法。量子方法的速度很高，当使用较少的主成分数支配内核矩阵。量子技术不需要表示每个训练实例的所有特征，但它可以生成适当的数据结构。生成内核矩阵后，训练数据的各个特征对用户完全隐藏。

Bishwas等人[7]采用全对量子多类-SVM对大数据进行分类。在这里，建立了k(k-1)/2分类器，用于对给定的隐藏量子查询状态进行分类。经典的多类-SVM可以用多项式运行时间来执行，本框架采用量子版本来加快系统的速度。本框架使用了多种分类方法，相对于经典的分类方法来说，达到了提速的效果。

Bishwas等人[12]开发了多类SVM和量子一元分类方法，用于处理大数据的分类。首先，引入了k个量子二元分类器。然后，利用这些分类器对未知量子状态进行分类，并根据最大置信度得分进行预测。

Zou[38]开发了Max-Relevance-Max-Distance(MRMD)特征排序技术，该技术可以管理特征预测以及排序任务的稳定性和准确性。该框架为保证特征选择的稳定性，避免了计算的困难。

## 2.1.7 基于优化的分类技术

下面讨论了实践基于优化技术的大数据分类的研究工作。

Meera和RosilineJeetha[23]开发了加速人工蜂群-人工神经网络（AABC-ANN）来解决高维度和特征选择问题。该框架由预处理、特征选择和分类组成。在预处理步骤中，利用KNN方法去除不相关的，和冗余的属性，以管理未知值。之后，基于AABC进行特征选择，以生成最优值。选取的特征被输入到训练和测试中。基于ANN分类器，对特征进行更精确的分类。

Hegde和Mundada[47]开发了认知粒子群优化（PSO）与深度学习相结合的大数据分类方法。这种方法被用来从给定的数据集中提取特征，并基于深度学习进行分类。所开发的方法可以调整其学习速率，从获得的数据中有效地提取特征。

Manoj等人[52]开发了基于蚁群优化(ACO)和人工神经网络(ANN)的混合算法，用于文本分类的特征选择。这种混合算法被发现是最佳的和有效的。在这里，ACO具有迅速聚集的能力，因为它在状态空间问题中具有搜索能力，可以有效地找到最小的特征子集。

Lin等[55]开发了改进的猫群优化（CSO）用于大数据分类。这里应用改进的CSO（ICSO）进行特征选择，最后使用SVM进行分类。该方法没有考虑其他方法来改进传统的CSO模式。

## 2.1.8 基于决策树的分类方法

下面就采用决策树分类器进行大数据分类的研究工作进行讨论。

Triguero等人[4]开发了基于apache spark的大数据方法来解决密度不足的问题。最初，将整个训练数据集划分为chunks，并使用正样本进行提取。之后，将正样本集进行广播，使所有节点都有不同的正样本的内存副本。对于每一个负分块数据，根据正样本集得到一个子集。然后，应用进化采样法（EUS）来减小两个类的大小，提高分类性能。最后，合并各种模型以预测测试集的类别。

Bhagat和Patil[11]开发了用于分类的Synthetic Minority Oversampling Technique（SMOTE）。最初，引入Binarization技术，将原始数据集分解为二进制类子集。然后，对每个不平衡的二进制类子集给出SMOTE方法进行数据平衡。最后，利用随机森林（RF）分类器进行分类。

Bifet等人[39]介绍了流式大数据的评估技术。这个框架分析了不平衡的数据流，其中修改在不同的时间尺度上达到，数据来自不同模型的测试和训练。Leszek等人[40]提出了数据流的决策树。利用不纯度量引入了一种新的划分树节点的措施，称为错误误分类。利用这种方法来决定从整个数据流中确定的较好属性。

## 2.1.9 其他分类方法

下面对大数据分类采用的其他分类方法进行阐述。

Liu等人[10]基于NB分类器开发了简易的整机系统，用于对庞大的数据集进行分类。这里，需要额外的模块来实现实验的自动化。Yan等[15]开发了可扩展的分类器合集技术，用于结合各种分类器的结果对大数据进行分类。最初，利用训练和验证输出的特征和分类器生成各种 "判断器"。然后，将这些判断器进行排序并加入到一个提升的分类器中。最后，介绍了基于火花的大数据分类框架。

Peixoto等[27]针对大数据分类发展了层次化多标签分类（HMC），称为语义层次化多标签分类。层次化、矢量化、索引化、解析和实现，是语义HMC所涉及的步骤。在这里，主要的三个阶段从数据源进行标签分层。其余的两个阶段则是用来根据层次化标签进行新项目的分类。

Fong等人[33]提出了轻量级特征选择，称为Swarm Search with Accelerated Particle Swarm Optimization，用于分类。在这里，从在最短的预处理时间内达到最高精度的角度出发，对两组分类算法，即批处理学习和增量或数据流学习进行了性能比较分析。

Triguero等人[34]提出了用于分类的原型减少（MRPR）的Map Reduce解决方案。这种方法用于解决由于内存和运行时间的限制，没有包含几千个例子的问题。Map Reduce具有透明、简单、有效的原型还原计算环境。同时分析了Join、Fusion和Filtering等三种不同的Reduce类型，以提供更精确的预处理集。

Wu等[35]针对大数据革命的特点开发了HACE，从数据挖掘的角度引入大数据处理模型。这种方法采用了挖掘、需求驱动的信息源聚合、安全以及隐私的考虑、用户兴趣的建模。

Chen[37]提出了利用Flink是图形处理单元(GPU)、内存集群计算平台之一的层次极限学习机(H-ELM)。为了提高性能，采用了各种优化技术，如合理的分区策略、基于缓存的方法和内存映射，用于映射特定的Java虚拟机对象进行缓冲。这个框架用于加速ELM和机器学习方法的几种变体。

Puthal等人[41]开发了基于动态质数的安全验证（DPBSV），用于大数据流。该方法使用共同的共享密钥进行动态更新，生成同步的质数对。Li等[42]开发了云计算中的隐私保护外包分类（POCC）方法，用于处理和存储。利用代理全态加密机制Using Gentr's来保存隐私数据。基于POCC，利用Crypto Service Provider(CSP)和S评价器从各种公钥的加密数据中训练一个分类模型。然后，这个模型以加密格式存储在评估器中，用于为客户提供服务。然而，在CSP和S之间，存在一些联系。该方法未能减轻计算成本和通信。

Vu等人[45]开发了第一个用于学习决策规则的分布式流算法。该框架利用水平和垂直并行的方式在集群上分发自适应模型规则（AMRules）。AM引入的决策规则是可理解的模型，其中前项规则是属性值上的连词，而后项则什么都不是，而是属性的组合。

Twardowski和Ryzko[48]提出了大数据处理的多代理架构。该模型使用lambda系统，它显示了利用自主代理来增强架构，实时提供数据的稳健处理。

Zhai等人[51]提出了基于模糊的ELM分类器进行分类。最初，对于每一个正向的情况，利用Map Reduce识别其最近的邻居敌方，并在其敌方最近的邻居中根据均匀分布以随机的方式产生p个正向。之后，建立平衡的Datasets，利用构建的数据子集进行非迭代学习，训练l个分类器。最后，将训练好的分类器利用模糊积分进行组合，对未见的情况进行分类。

Bifet等人[53]开发了用于大数据流挖掘的可扩展高级大规模在线分析（SAMOA）工具。该框架用于收集分布式流方法，用于机器学习和数据挖掘任务，如分类、聚类、回归和编程抽象。同时，这个方法还措施了一个可插拔的设计，以分布式的方式运行在各种处理引擎上，即Samza以及Storm S4。

## 3. 确定的研究差距

本节论述了文学作品中存在的各种研究差距和挑战，具体如下：

基于深度学习的方法的局限性如下。在[2]中，利用深度学习分类器来提高速度，深度学习模型由若干隐藏层和计算集群上的数百万个参数组成。在[13]中，该方法没有考虑线性回归模型来进行价格值预测。在[30]中，DBN没有考虑个人特征数据和基于功能磁共振成像（fMRI）的信息进行分析。

基于 CNN 的分类技术面临的挑战如下。[29]中的方法没有考虑分类的目标名词，也没有在系统中加入一些语言学的先验知识。基于NN的数据分类方法面临的主要挑战是。NN[21]未能考虑更多的数据集，将Shuffled Frog-Leaping Approach(SFLA)与随机搜索算法进行比较。基于KNN的数据分类面临的挑战是，KNN分类器[3]没有使用最新的技术，如spark，以加快计算速度。

基于模糊的分类方法的主要挑战如下。该方法没有考虑其他编程模型，如Spark，以提高效率和扩展性[17]。在[20]中，模糊KNN分类器在减速器中没有考虑模糊机制。[22]中的技术需要使用FCBF和Bagging算法等技术进一步提高性能。在[36]中的方法没有估计到神经模糊在各种医学诊断问题，如微阵列基因选择等数据挖掘问题。

基于优化的数据分类技术面临的挑战问题是，ACO[52]没有引入其他分类器，未能使用其他种类的数据，如音频、视频、图像等。

基于决策树的数据分类技术面临的挑战如下。[4]中的方法没有考虑有效的大数据策略，可以在大数据背景下部署预处理程序，如欠采样以及过采样方法。合成少数派超采样技术（SMOTE）[11]没有解决由超采样率引起的问题，因为超采样率随数据集而变化。该方法未能考虑回归、多标签和多目标学习[39]。

其他方法面临的挑战是。[10]中的方法没有考虑到文本和图像的信息融合、使用云计算的网络分析以及分布式机器人应用。语义HMC未能在Reduce和Map阶段之间增加集成，以最大限度地减少计算时间和数据到Redducers的大小[27]。在[45]中，该方法没有考虑处理高维和大规模数据集的新方法。在[34]中，该方法没有使用结构化和多目标学习。

## 4. 分析和讨论

本节根据所采用的框架、利用的数据集、准确度范围、评价指标、实现工具和采用的分类方法对大数据分类的不同研究工作进行了分析和讨论。

## 4.1 根据采用的方法

本小节阐述了在所采用的大数据分类方案基础上进行的分析。图2描述了为对大数据进行有效分类而采用的不同分类方法。从图2中可以看出，在55篇研究论文中，约有11篇研究论文，即27%的论文采用了深度学习分类器，11%的研究作品采用了Fuzzy分类器。12%的研究采用了CNN分类器，而7%的研究，即3篇研究论文采用了NN分类器，其余作品采用了其他技术，如KNN、SVM和优化、决策树分类器等。

## 4.2基于实施工具

本小节讨论考虑到研究论文中采用的模拟工具进行的分析。表1代表了为进行有效的大数据分类而采用的不同的实现工具。常用的实现工具有Java、Cloudera、MATLAB、Hadoop和Weka。从表1可以清楚地看出，MATLAB工具是有效大数据分类经常采用的实现工具。

## 4.3基于评价指标

本小节展示了基于重搜索论文中使用的不同评价指标的分析。

表4-1 基于大数据分类的实现工具分析

| 实现工具 | 研究数据 |
| --- | --- |
| Java | [[2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref002)] |
| Cloudera | [[3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref003), [4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref004), [5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005)] |
| Matlab | [[23](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref023), [36](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref036), [41](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref041), [44](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref044), [46](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref046), [51](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref051)] |
| Weka | [[38](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref038)] |
| Others | [[6](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref006), [8](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref008), [14](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref014), [16](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref016), [26](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref026), [29](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref029), [21](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref021), [29](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref029), [53](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref053)] |

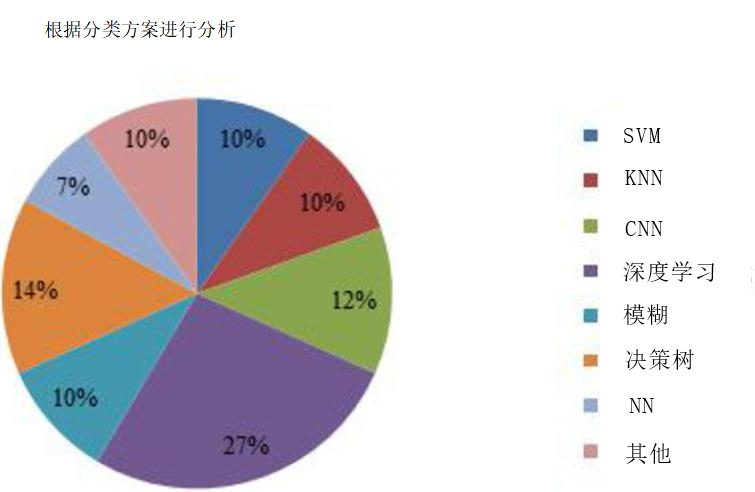


图4-1 根据分类方案进行分析

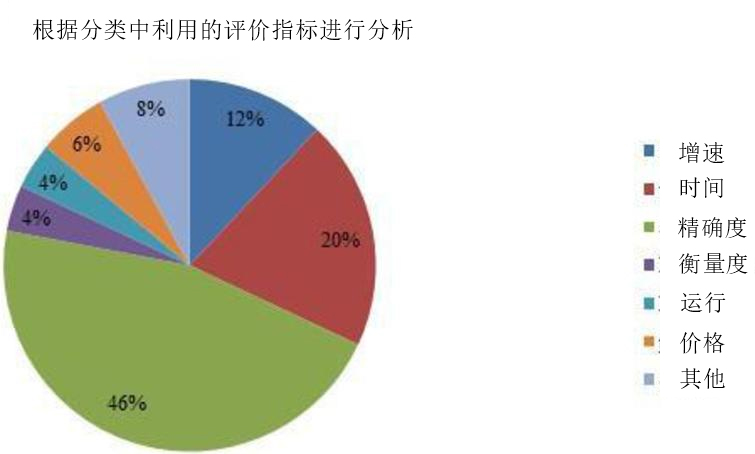


图4-2 根据分类中利用的评价指标进行分析

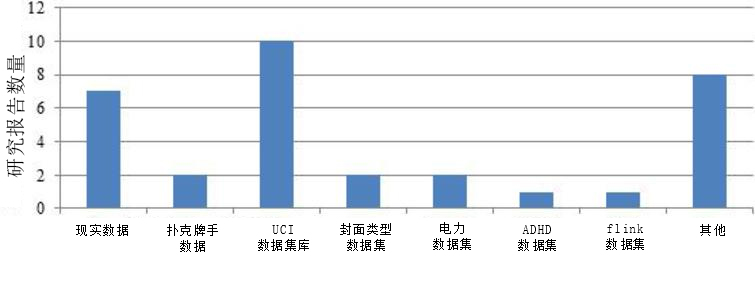


图4-3 根据分类方法中利用的数据集进行分析

图4-2是基于时间、速度、精度、F-measure、运行时间、精度等评价指标的大数据分类系统研究论文的分析图。从图3中可以看出，在55篇研究论文中，约有23篇，即46%的研究论文采用了准确度指标，20%的研究作品采用了时间指标。12%的研究作品采用了速度指标，而8%的研究作品，即4篇研究论文采用了其他指标。近，6%的研究作品采用精度度量，4%的论文采用F度量，2篇研究作品采用运行时间。由此可见，大部分研究论文都采用了精度度量来进行大数据分类。

## 4.4以精度范围为基础

本小节阐述了在准确度范围方面进行的分析。表2代表了基于分类精度的分析。从表中可以看出，在60%-70%的范围内，有5篇研究论文的准确率达到了70%-80%，有8篇研究作品的准确率达到了80%。然后，准确率在80%-90%之间的有2篇研究作品，5篇研究论文的准确率在90%-99.9%之间。

## 4.5基于所使用的数据集

表4-2基于分类精度的分析

| 精度范围 | 研究数据 |
| --- | --- |
| 60%–70% | [[19](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref019), [25](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref025), [30](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref030), [40](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref040), [50](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref050)] |
| 70%–80% | [[8](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref008), [9](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref009), [13](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref013), [17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017), [20](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref020), [22](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref022), [33](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref033), [39](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref039)] |
| 80%–90% | [[10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010), [18](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref018)] |
| 90%–99.9% | [[5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005), [6](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref006), [16](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref016), [23](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref023), [47](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref047)] |

本小节将对不同研究工作所采用的数据集进行分析。图4描述了大数据分类所使用的不同数据集，不同的研究工作中常用的数据集有真实世界、扑克牌手、UCI数据集库、封面类型、电力、ADHD和flink数据集。从图4可以看出，最常使用的数据集是UCI数据集库和真实世界数据集。

## 4.6基于所采用的框架

本小节利用表3对大数据分类所采用的框架进行分析。通过分析，可以清楚地看出，在55篇研究论文中，约有12篇研究论文采用了MR框架，6篇研究作品采用了Hadoop框架，然后，4篇研究论文采用了Apache Hadoop框架，7篇研究作品采用了Apache Spark框架。然后，4篇研究论文采用了Apache Hadoop框架，7篇研究作品利用了Apache Spark框架。由此可见，处理大数据的常用框架是MR框架。

表4-3基于大数据分类使用的框架进行分析

| 框架 | 研究数据 |
| --- | --- |
| Map Reduce 框架 | [[3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref003), [10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010), [11](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref011), [17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017), [19](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref019), [20](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref020), [27](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref027), [37](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref037), [48](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref048), [49](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref049), [50](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref050), [51](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref051)] |
| Apache Hadoop 框架 | [[17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017), [21](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref021), [34](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref034), [35](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref035)] |
| Apache spark 框架 | [[2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref002), [4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref004), [5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005), [11](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref011), [15](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref015), [28](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref028), [39](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref039)] |
| Hadoop 框架 | [[10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010), [22](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref022), [44](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref044), [45](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref045), [48](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref048), [54](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref054)] |

## 5. 结束语

本研究阐明了为有效分类大数据而采用的不同数据流挖掘方案的调查。本文的原始目标是通过对IEEE、谷歌Scholar、Elsevier和Science Direct等多篇研究论文的分析，回顾、分类和学习用于大数据分类的独特分类方法。对评价指标、采用的分类方法、使用的数据集、实现工具以及新的数据流挖掘方案进行了分析和讨论。就评估指标、采用的分类方法、使用的数据集、实施工具、使用的框架和精度范围进行了分析和讨论。

此外，本研究通过考虑研究的空白和问题，为有效的大数据分类提供了重要的未来前景。通过回顾和分析，可以得出结论，在大多数探索大数据分类有效性的研究工作中，准确性是广泛关注的评价指标。被广泛采用的框架是MR，常用的分类方法有深度学习分类器和模糊分类器。