



(12) 发明专利

(10) 授权公告号 CN 108021940 B

(45) 授权公告日 2023. 04. 18

(21) 申请号 201711235660.8

(22) 申请日 2017.11.30

(65) 同一申请的已公布的文献号

申请公布号 CN 108021940 A

(43) 申请公布日 2018.05.11

(73) 专利权人 中国银联股份有限公司

地址 200135 上海市浦东新区含笑路36号
银联大厦

(72) 发明人 黄自力 杨阳 陈舟 朱浩然

(74) 专利代理机构 中国专利代理(香港)有限公
司 72001

专利代理师 王星 杨美灵

(51) Int. Cl.

G06F 18/24 (2023.01)

G06F 18/27 (2023.01)

G06F 18/214 (2023.01)

G06F 18/21 (2023.01)

(56) 对比文件

CN 107292330 A, 2017.10.24

CN 104573013 A, 2015.04.29

CN 102163239 A, 2011.08.24

CN 106991296 A, 2017.07.28

CN 106548210 A, 2017.03.29

CN 102332028 A, 2012.01.25

CN 105282720 A, 2016.01.27

CN 105260746 A, 2016.01.20

US 2014369597 A1, 2014.12.18

US 2005249401 A1, 2005.11.10

US 2007127825 A1, 2007.06.07

邹权 等. 基于集成学习和分层结构的多分
类算法.《模式识别与人工智能》.2015,第28卷
(第9期),

刘宁 等. 一种多分类器协同的半监督分类
算法SSC_MCC.《河南科学》.2015,第33卷(第9
期),

Lior Rokach. Ensemble-based
classifiers.《Artificial Intelligence
Review 》.2010,

(续)

审查员 张露

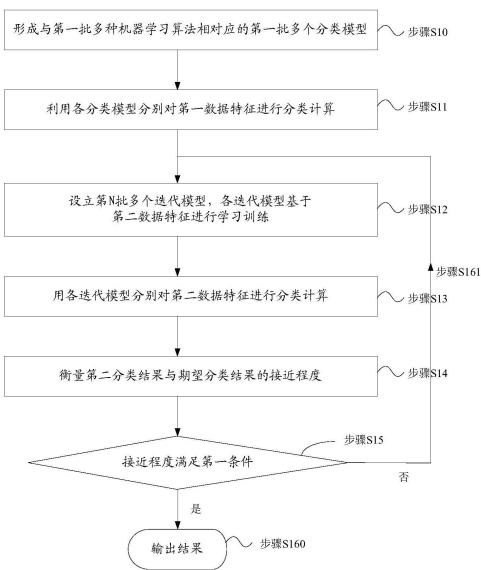
权利要求书2页 说明书7页 附图2页

(54) 发明名称

基于机器学习的数据分类方法及系统

(57) 摘要

本发明涉及一种基于机器学习的数据分类方法,包括:形成与第一批多种机器学习算法相对应的第一批多个分类模型;利用第一批多个分类模型中的各分类模型分别对第一数据特征进行分类计算;设立第N批多个迭代模型,基于第二数据特征进行学习训练;利用各迭代模型分别对第二数据特征进行分类计算;衡量第二分类结果与期望分类结果的接近程度;若接近程度满足第一条条件,方法结束;否则,迭代执行迭代模型的设立步骤。该方法能够实现“类深度学习”的效果,这极大提升了数据分类的普适性和准确性。同时,这种数据分类方法可信度高,人工干预程度低。



[转续页]

[接上页]

(56) 对比文件

JEMAL H. ABAWAJY et al. Large
Iterative Multitier Ensemble Classifiers

for Security of Big Data.《IEEE
TRANSACTIONS ON EMERGING TOPICS IN
COMPUTING》.2014,第2卷(第3期),

1. 一种用于识别恶意网页的数据分类方法,包括:

a)、采用第一批多种机器学习算法,分别对数据集进行采集、并进行学习训练,以形成与所述第一批多种机器学习算法相对应的第一批多个分类模型;

b)、利用所述第一批多个分类模型中的各分类模型分别对来自所述数据集的第一数据特征进行分类计算,以得到第一分类结果;

c)、设立第N批多个迭代模型,以所述第N批多个迭代模型中的各迭代模型分别获取所述第一分类结果和所述第一数据特征来形成第二数据特征,并基于所述第二数据特征进行学习训练;其中,N为大于等于1的正整数;

d)、利用各所述迭代模型分别对所述第二数据特征进行分类计算,以得到第二分类结果;

e)、衡量所述第二分类结果与期望分类结果的接近程度;

f)、若所述接近程度满足第一条件,则输出所述第二分类结果,所述方法结束;否则,N自增1,且以所述第二分类结果作为所述第一分类结果,所述方法回到步骤c)迭代执行,

其中,所述第一条件包括:所述第二分类结果是收敛的,

其中,所述第一批多种机器学习算法包括:逻辑回归算法;支持向量机算法;朴素贝叶斯算法;K近邻算法;BP神经网络算法;决策树算法;DNN算法;RNN算法;以及,CNN算法,

其中,所述第N批多个迭代模型中的各所述迭代模型分别采用第二批多种机器学习算法中的一个,所述第二批多种机器学习算法包括:Adaboost算法;随机森林算法;梯度提升决策树算法;以及,eXtreme GradientBoosting算法。

2. 根据权利要求1所述的方法,其特征在于,所述步骤d)进一步包括:

自主地调节各所述迭代模型所采用的参数。

3. 根据权利要求1所述的方法,其特征在于,每次迭代执行所述步骤c)时,使得第K+1批多个迭代模型中的所述迭代模型的数量小于或等于第K批多个迭代模型中的所述迭代模型的数量,其中K为大于等于1的正整数。

4. 根据权利要求3所述的方法,其特征在于,每次迭代执行所述步骤c)时,使得所述第K+1批多个迭代模型各自对应的机器学习算法在算法优先级上高于或等于所述第K批多个迭代模型各自对应的机器学习算法。

5. 根据权利要求3所述的方法,其特征在于,迭代执行所述步骤c)的次数至少为1。

6. 根据权利要求1所述的方法,其特征在于,所述步骤c)还包括:

对所述第二数据特征进行特征处理,以使得所述第N批多个迭代模型能够直接依据所述特征处理的结果来进行分类计算。

7. 根据权利要求1所述的方法,其特征在于,在所述步骤e)中,衡量所述接近程度包括:计算所述第二分类结果的分类准确率、分类召回率、以及F度量。

8. 一种用于识别恶意网页的数据分类系统,该系统包括:

数据分类模块,包括第一批多个分类模型,所述第一批多个分类模型中的各分类模型分别采用第一批多种机器学习算法中的一个对数据集进行采集、并进行学习训练,以及,对来自所述数据集的第一数据特征进行分类计算,以得到第一分类结果;

数据迭代模块,其初始化地设立或基于筛选器的通知而逐次设立第N批多个迭代模型,以所述第N批多个迭代模型中的各迭代模型分别获取所述第一分类结果和所述第一数据特

征来形成第二数据特征,并基于所述第二数据特征进行学习训练,以及,利用各所述迭代模型分别对所述第二数据特征进行分类计算,以得到第二分类结果;其中,N为大于等于1的正整数;以及

所述筛选器,其衡量所述第二分类结果与期望分类结果的接近程度,以及,在所述接近程度满足第一条件时输出所述第二分类结果,否则,N自增1,以所述第二分类结果作为所述第一分类结果,并向所述数据迭代模块发出所述通知,

其中,所述第一条件包括:所述第二分类结果是收敛的,

其中,所述第一批多种机器学习算法包括:逻辑回归算法;支持向量机算法;朴素贝叶斯算法;K近邻算法;BP神经网络算法;决策树算法;DNN算法;RNN算法;以及,CNN算法,

其中,所述第N批多个迭代模型中的各所述迭代模型分别采用第二批多种机器学习算法中的一个,所述第二批多种机器学习算法包括:Adaboost算法;随机森林算法;梯度提升决策树算法;以及,eXtreme Gradient Boosting算法。

基于机器学习的数据分类方法及系统

技术领域

[0001] 本发明涉及机器学习技术领域,更具体地说,涉及一种基于机器学习的数据分类方法及系统。

背景技术

[0002] 在大数据安全分析领域,在进行正常/危险标签识别时,如在恶意网页的识别分类,常常先要根据安全人员对场景与问题的测评分析,然后选择具体的机器学习算法来进行训练与测试,但是识别结果会过多取决于单一机器学习算法的合适程度以及模型调优的时间,而通常无法有一定的普适性,即知识迁移的能力。

[0003] 机器学习算法在进行事物识别与分类前,必须要进行业务场景的精确分析与研究,即必须由具备相关领域专业知识的技术人员以经验为向导,选择、设计或优化其中一至多种机器学习分类算法,来实现与具体领域相关的分类识别。

[0004] 以下介绍一些现有的基于机器学习的安全检测和/或识别方法,以及各自特点与缺陷。

[0005] 1. 基于特定机器学习方法

[0006] 从大数据的角度,根据具体场景,选择特定的机器学习算法(例如朴素贝叶斯在垃圾邮件识别有很强的适用性)来进行建模,训练与测试数据、并实现分类。此方法可智能检测安全威胁,但对于算法的选取与模型的优化有较高的要求,如若没有选择到合适的算法,则最终的分类效果可能并不理想。

[0007] 2. 模型参数对分类效果极其敏感

[0008] 当选定了指定的机器学习算法,对于数据的训练,需要通过技术人员根据经验,反复地调节参数,来观察模型最终的分类效果。在这一过程中,参数的调节对结果的影响是关键,这取决于专业人员的技术水平、以及所花费的时间。

[0009] 3. 某一算法对多分类器的组合效果有显著影响

[0010] 采用多机器学习切割融合方法,将一个整体架构,根据需求,拆分成多个部分,每个部分可根据不同的算法进行分析,最终将所有结果汇总组合,例如身份识别,分为人像的匹配识别算法与身份证号或姓名的匹配算法,两个部分所采用的算法不同且互不干扰,最终将识别结果汇总即可。但是,如果某一部分算法的识别效果较差,会影响汇总结果;此外,若不同部分算法产生相互矛盾的结果,分类器将左右为难,而再次需要人工介入。

[0011] 4. 主辅组合方法与场景关联过高

[0012] 采用多机器学习互补方法,以一个算法为主,其他算法为辅,分析问题。例如Alpha Go人工智能,以蒙特卡洛树搜索为主算法,再配合价值网络等算法辅助完善其功能。这种方法有主辅之分,与具体场景会关联紧密;若涉及到不同场景的检测问题,有可能需要完全改变模型,因此不具备普适性。

[0013] 5. 多种机器学习算法线性组合或简单投票

[0014] “多种机器学习算法线性组合或简单投票”方法一般会选择多个机器学习算法各

自训练与分类,将最后的结果进行“线性组合”或“简单投票”,得到最终的分类结果。即使这种方法的优点在于对多个算法进行了结合应用,但是“组合器”或“投票器”的应用又容易为分类过程引入相当程度的主观因素或经验因素,使得分类过程又偏离了机器学习-机器分类的目标。

[0015] 因此,技术人员期望获得一种结合多种机器学习算法、并能克服上述各种缺陷的数据分类方法。

发明内容

[0016] 本发明的目的之一在于提供一种基于机器学习的数据分类方法,其可信度高、人工干预程度低。

[0017] 为实现上述目的,本发明提供一种技术方案如下。

[0018] 一种基于机器学习的数据分类方法,包括:a)、采用第一批多种机器学习算法,分别对数据集进行采集、并进行学习训练,以形成与第一批多种机器学习算法相对应的第一批多个分类模型;b)、利用第一批多个分类模型中的各分类模型分别对来自数据集的第一数据特征进行分类计算,以得到第一分类结果;c)、设立第N批多个迭代模型,以第N批多个迭代模型中的各迭代模型分别获取第一分类结果和第一数据特征以形成第二数据特征,并基于第二数据特征进行学习训练;其中,N为大于等于1的正整数;d)、利用各迭代模型分别对第二数据特征进行分类计算,以得到第二分类结果;e)、衡量第二分类结果与期望分类结果的接近程度;f)、若接近程度满足第一条条件,则输出第二分类结果,方法结束;否则,N自增1,且以第二分类结果作为第一分类结果,方法回到步骤c)迭代执行。

[0019] 优选地,步骤d)进一步包括:自主地调节各迭代模型所采用的参数。

[0020] 优选地,每次迭代执行步骤c)时,使得第K+1批多个迭代模型中的迭代模型的数量小于或等于第K批多个迭代模型中的迭代模型的数量,其中K为大于等于1的正整数。

[0021] 优选地,每次迭代执行步骤c)时,使得第K+1批多个迭代模型各自对应的机器学习算法在算法优先级上高于或等于第K批多个迭代模型各自对应的机器学习算法。

[0022] 优选地,步骤c)还包括:对第二数据特征进行特征处理,以使得第N批多个迭代模型能够直接依据特征处理的结果来进行分类计算。

[0023] 优选地,第一批多种机器学习算法包括:逻辑回归算法;支持向量机算法;朴素贝叶斯算法;K近邻算法;BP神经网络算法;决策树算法;DNN算法;RNN算法;以及,CNN算法。

[0024] 本发明还公开一种基于机器学习的数据分类系统,包括:数据分类模块,包括第一批多个分类模型,第一批多个分类模型中的各分类模型分别采用第一批多种机器学习算法中的一个对数据集进行采集、并进行学习训练,以及,对来自数据集的第一数据特征进行分类计算,以得到第一分类结果;数据迭代模块,其初始化地设立或基于筛选器的通知而逐次设立第N批多个迭代模型,以第N批多个迭代模型中的各迭代模型分别获取第一分类结果和第一数据特征以形成第二数据特征,并基于第二数据特征进行学习训练,以及,利用各迭代模型分别对第二数据特征进行分类计算,以得到第二分类结果;其中,N为大于等于1的正整数;以及筛选器,其衡量第二分类结果与期望分类结果的接近程度,以及,在接近程度满足第一条条件时输出第二分类结果,否则,N自增1,以第二分类结果作为第一分类结果,并向数据迭代模块发出通知。

[0025] 本发明各实施例提供的基于机器学习的数据分类方法及系统,能够实现“类深度学习”的效果,这极大提升了数据分类的普适性和准确性。同时,这种数据分类方法可信度高,人工干预程度低。

附图说明

[0026] 图1示出本发明第一实施例提供的基于机器学习的数据分类方法的流程示意图。

[0027] 图2示出本发明第二实施例提供的基于机器学习的数据分类系统的模块结构示意图。

具体实施方式

[0028] 在以下描述中提出具体细节,以便提供对本发明的透彻理解。然而,本领域的技术人员将清楚地知道,即使没有这些具体细节也可实施本发明的实施例。在本发明中,可进行具体的数字引用,例如“第一元件”、“第二装置”等。但是,具体数字引用不应当被理解为必须服从于其字面顺序,而是应被理解为“第一元件”与“第二元件”不同。

[0029] 本发明所提出的具体细节只是示范性的,具体细节可以变化,但仍然落入本发明的精神和范围之内。术语“耦合”定义为表示直接连接到组件或者经由另一个组件而间接连接到组件。

[0030] 以下通过参照附图来描述适于实现本发明的方法、系统和装置的优选实施例。虽然各实施例是针对元件的单个组合来描述,但是应理解,本发明包括所公开元件的所有可能组合。因此,如果一个实施例包括元件A、B和C,而第二实施例包括元件B和D,则本发明也应被认为包括A、B、C或D的其他剩余组合,即使没有明确公开。

[0031] 如图1所示,本发明第一实施例提供一种基于机器学习的数据分类方法,其结合多批次多种机器学习方法对采集自数据集中的一个或多个数据特征进行分类,该方法包括如下步骤。

[0032] 步骤S10、形成与第一批多种机器学习算法相对应的第一批多个分类模型。

[0033] 具体地,在该步骤中,采用第一批多种机器学习算法,分别对数据集进行采集、并进行学习训练,进而形成上述第一批多个分类模型。除数据采集、学习训练之外,根据实际应用场景,该步骤可能还涉及数据预处理、特征工程、以及抽样与建模等子步骤,以分别得到最终分类模型。

[0034] 仅作为示例,第一批多种机器学习算法采用相对简单的机器学习算法,包括但不限于:逻辑回归算法;支持向量机算法;朴素贝叶斯算法;K近邻算法;以及,BP神经网络算法。以该5种机器学习算法为例,步骤S10中形成的第一批分类模型包括5个分类模型,它们分别采用这5种机器学习算法其中之一。

[0035] 步骤S11、利用第一批多个分类模型中的各分类模型分别对来自数据集的第一数据特征进行分类计算。

[0036] 该步骤承接上述步骤S10执行,其利用建立好的多个分类模型对从数据集中提取的第一数据特征进行分类计算,进而得到第一分类结果。

[0037] 作为示例,就单个数据特征而言,采用5种不同机器学习算法的5个分类模型将各自产生一个分类结果,所产生的5个分类结果可能彼此相同或不同。

[0038] 步骤S12、设立第N批多个迭代模型,以第N批多个迭代模型中的各迭代模型分别获取第一分类结果和第一数据特征以形成第二数据特征,并基于第二数据特征进行学习训练。

[0039] 在该步骤中,N为大于等于1的正整数。在首次执行步骤S12时,N初始化为1,随后,根据第一实施例的以下各步骤,每迭代执行一次步骤S12,N将自增1(或在步骤S161自增,也可在步骤S12递增)。

[0040] 以第一批多个迭代模型为例,它们分别接受第一批多个分类模型的输出(即第一分类结果)作为输入,同时还获取待分类的第一数据特征(原始数据特征)同样作为输入。

[0041] 类似地,在根据以下步骤S161启动迭代过程之后,第N批多个迭代模型中的每个都将分别获取第N-1批迭代模型的输出、以及原始数据特征,来形成相应的第二数据特征。进而,在以下步骤S13中可以各自进行分类计算,以获得新的分类输出结果(第二分类结果)。

[0042] 作为示例,例如,假设选定特征处理采用的函数为One-Hot编码,当分类算法为5种、分类标签(即第一分类结果)有4个,再加上原始数据特征,那么在各迭代层中,每个迭代模型可接受的特征维度至少为 $4*5+1=21$ 维。

[0043] 根据实际应用场景,可能需要对第二数据特征进行适量的特征处理,以使得第一批多个迭代模型能够直接依据这种特征处理的结果来进行分类计算。

[0044] 优选情况下,第N批多个迭代模型中的各迭代模型分别采用第二批多种机器学习算法中的一种(可以为任选的一个,且相互之间可以相同或不同),第二批多种机器学习算法可采用bagging或boosting算法(组合性质的高级机器学习算法),包括但不限于:Adaboost算法;随机森林算法;梯度提升决策树算法;以及,eXtreme Gradient Boosting算法。

[0045] 如上所述,各迭代模型结合第一分类结果和原始数据特征来形成不同于原始数据特征的第二数据特征,通常第二数据特征的特征维度将远高于第一数据特征。经由学习训练,各迭代模型都能够适合于对新形成的第二数据特征进行识别分类。在迭代模型的学习训练中,仍有可能需要一定程度的人工干预(例如获取专家经验等),但与分类模型相比,其所需的干预程度明显降级。

[0046] 步骤S13、利用各迭代模型分别对相应的第二数据特征进行分类计算,以得到第二分类结果。

[0047] 具体地,采用第二批多种机器学习算法其中之一,同一批多个迭代模型中的各迭代模型都将进行分类计算,分别得到各自的分类结果,它们可能彼此相同或不同,这些分类结果的集合形成第二分类结果。

[0048] 需要说明的是,各迭代模型不仅针对从数据集中提取的第一数据特征(原始数据特征)进行分类计算,而且还将第一分类结果作为重要因素来考虑在内。换言之,基于原始数据特征和第一分类结果来共同形成新的第二数据特征,进而针对第二数据特征来进行分类计算,本发明实际上将原始的待分类数据进行了“解析”与“拓展”。

[0049] 作为进一步的改进,通过相应的机器学习算法所进行的学习训练以及分类计算,各迭代模型还能够采用交叉验证的方式来自主地调节各种参数、权重等。而在参数或权重经一次甚至多次的调节(可迭代执行)而得到优化之后,可以更新(或优化)所计算的第二分类结果,这将作为优选实施方式。可以理解的是,参数的自主调节过程可以独立于分类计算

过程,但两者在执行时间上可能存在重叠或交叉。

[0050] 步骤S14、衡量第二分类结果与期望分类结果的接近程度。

[0051] 在步骤S14中,期望分类结果是掌握熟练分类技能的技术人员对原始数据特征的分类结果,它可以作为数据项包含在数据集的一部分中,也可以在后期提供。第二分类结果与期望分类结果的接近程度可以采用多种指标来体现,例如,第二分类结果的分类准确率、分类召回率、以及F度量,也可以包括相对简单的指标,例如,第二分类结果与期望分类结果相同部分所占的比例、两者间的方差等。

[0052] 步骤S15、确定接近程度是否满足第一条件。

[0053] 在该步骤中,第一条件可以为技术人员普遍认可的数据分类标准,或用户根据实际应用场景而设置的其他类型判定条件,达到这一标准,即意味着第二分类结果接近于掌握熟练分类技能的技术人员的期望,或者,第二分类结果是收敛的,从而使得最终能够得到一个确定的分类值。

[0054] 以下为两个选择性执行的步骤,即,根据步骤S15的判断结果,选择以下两个步骤S160、步骤S161其中之一来执行。

[0055] 步骤S160、若接近程度满足第一条件,则输出第二分类结果,方法结束。

[0056] 在输出第二分类结果之后,有可能第二分类结果已经表明了唯一的分类值,或者,也可能需要对第二分类结果再进行进一步的集成,以生成唯一确定的分类值。

[0057] 步骤S161,若接近程度未能满足第一条件,N自增1,且以第二分类结果作为第一分类结果,方法回到步骤S12迭代执行。

[0058] 在该步骤S161中,因为第二分类结果偏离技术人员的期望,而迭代执行步骤S12,即,新设一批多个迭代模型,将当前的第二分类结果作为该新设的一批多个迭代模型的输入,经由学习训练,进而该新设的一批多个迭代模型能够再进行一次分类计算。按照这种方式逐次迭代,直到分类结果与期望分类结果的接近程度满足第一条件。

[0059] 为实现步骤S12的迭代,将N自增1,将第二分类结果作为第一分类结果。

[0060] 可以理解,能够为迭代过程设置最大次数,迭代达到最大次数时将停止,并输出当前的分类结果。与第一分类结果相比,这时的分类结果将会更接近技术人员的期望。

[0061] 在需要迭代执行步骤S12的情况下,每次迭代执行时,使得第K+1批多个迭代模型中的迭代模型的数量小于或等于第K批多个迭代模型中的迭代模型的数量,其中K为大于等于1的正整数。

[0062] 作为示例,第一批多个迭代模型中的迭代模型的数量为n,第二批多个迭代模型中的迭代模型的数量也为n,第T批多个迭代模型中的迭代模型的数量为m,其中, $m < n$,T为迭代模型形成的总批数。

[0063] 此外,每次迭代执行时,使得第K+1批多个迭代模型各自对应的机器学习算法在算法优先级上高于或等于第K批多个迭代模型各自对应的机器学习算法,其中K为大于等于1的正整数。

[0064] 本文所述的“算法优先级”可以表示算法复杂度、算法的进化程度、算法为本领域技术人员所采用的频度、或用户对算法的喜好程度。

[0065] 作为示例,第二批多个迭代模型各自可以采用与第一批多个迭代模型相同算法优先级的机器学习算法,但在第T批多个迭代模型中,所采用的机器学习算法在算法优先级上

应高于前面的第1,2,⋯,T-1批迭代模型。

[0066] 采用上述关于迭代模型的数量及算法优先级的这种配置方式是为了以第T批多个迭代模型的分类结果来输出,例如,作为第二分类结果输出至筛选器或比较器中、或不经筛选器比较器而直接作为最终的分类值。然而,在实际应用中,也可以略过第T批多个迭代模型的设置,而直接由筛选器进行处理,生成最终分类值。

[0067] 优选情况下,步骤S12迭代执行的次数至少为1,即,总计形成至少2批的多个迭代模型。

[0068] 可以理解,在迭代层中不断使用多种不同机器学习算法进行迭代计算,并在迭代层最后一层以更高级机器学习算法,这能够集成多层分类计算的结果而得到最终输出。多次迭代的应用提升了算法模型的维度,达成了“类深度学习”的效果,这极大提升了数据分类的普适性和准确性。同时,这种数据分类方法可信度高,人工干预程度低。

[0069] 如图2所示,本发明第二实施例提供一种基于机器学习的数据分类系统,其包括数据分类模块21、数据迭代模块22以及筛选器23。

[0070] 数据分类模块21包括第一批多个分类模型210,211,⋯21N。其中各分类模型210,211,⋯21N分别采用第一批多种机器学习算法中的一个对数据集进行采集、并进行学习训练。随后,各分类模型210,211,⋯21N对来自数据集的第一数据特征(原始数据特征)进行分类计算,以得到第一分类结果。

[0071] 数据迭代模块22初始化地设立第一批多个迭代模型2211,2212,⋯221n,其中迭代模型的数量为n。数据迭代模块22还可以基于筛选器23的通知而设立第二批多个迭代模型,以第二批多个迭代模型中的各迭代模型分别获取第一批多个迭代模型输出的分类结果和原始数据特征,进而形成相应的第二数据特征,并基于该相应的第二数据特征进行学习训练,使得各迭代模型适合用于进行更进一步的分类计算。经这种分类计算后,可以得到第二分类结果。与分类计算的过程相独立,各迭代模型还可以自主地调节其自身所采用的参数。

[0072] 类似地,第K批多个迭代模型22K1,22K2,⋯22Km基于第K-1批迭代模型输出的分类结果和原始数据特征来形成相应的第二数据特征,随后进行分类计算。其中迭代模型的数量为m,m可以与n相同或不同。

[0073] 在需要输出至筛选器之前,数据迭代模块22设立第K+1批迭代模型。第K+1批迭代模型中迭代模型的数量优选地小于第一、第二直至第K批,而其采用的机器学习算法在算法优先级上将会高于第一、第二直至第K批。其中,作为备选,K的值可以由技术人员根据实际应用场景来具体设置。

[0074] 筛选器23衡量第二分类结果与期望分类结果的接近程度,以及,在接近程度满足第一条件时输出第二分类结果;否则,N自增1,以第二分类结果作为第一分类结果,并向数据迭代模块22发出通知,数据迭代模块22将根据这一通知而新设一批多个迭代模型,进而再执行一轮数据分类。这种方式可以多次迭代执行,直到第K+1批迭代模型输出的分类结果接近于技术人员的期望分类结果。为了防止陷入迭代的死循环,也可以为迭代过程设置最大次数,达到这一最大次数,迭代过程将停止,并以当前的分类结果来输出。

[0075] 通过设置多层迭代模型,可以实现多层学习训练、多层分类计算,并可以结合多种不同机器学习算法,这种数据分类系统建立了高维度算法模型,达成了“类深度学习”的效果,极大提升了数据分类的普适性和准确性。

[0076] 在本发明的一些实施例中,系统的至少一部分可采用通信网络所连接的一组分布式计算装置来实现,或,基于“云”来实现。在这种系统中,多个计算装置共同操作,以通过使用其共享资源来提供服务。

[0077] 基于“云”的实现可提供一个或多个优点,包括:开放性、灵活性和可扩展性、可中心管理、可靠性、可缩放性、对计算资源所优化、具有聚合和分析跨多个用户的信息的能力、跨多个地理区域进行连接、以及将多个移动或数据网络运营商用于网络连通性的能力。

[0078] 上述说明仅针对于本发明的优选实施例,并不在于限制本发明的保护范围。本领域技术人员可能作出各种变形设计,而不脱离本发明的思想及附随的权利要求。

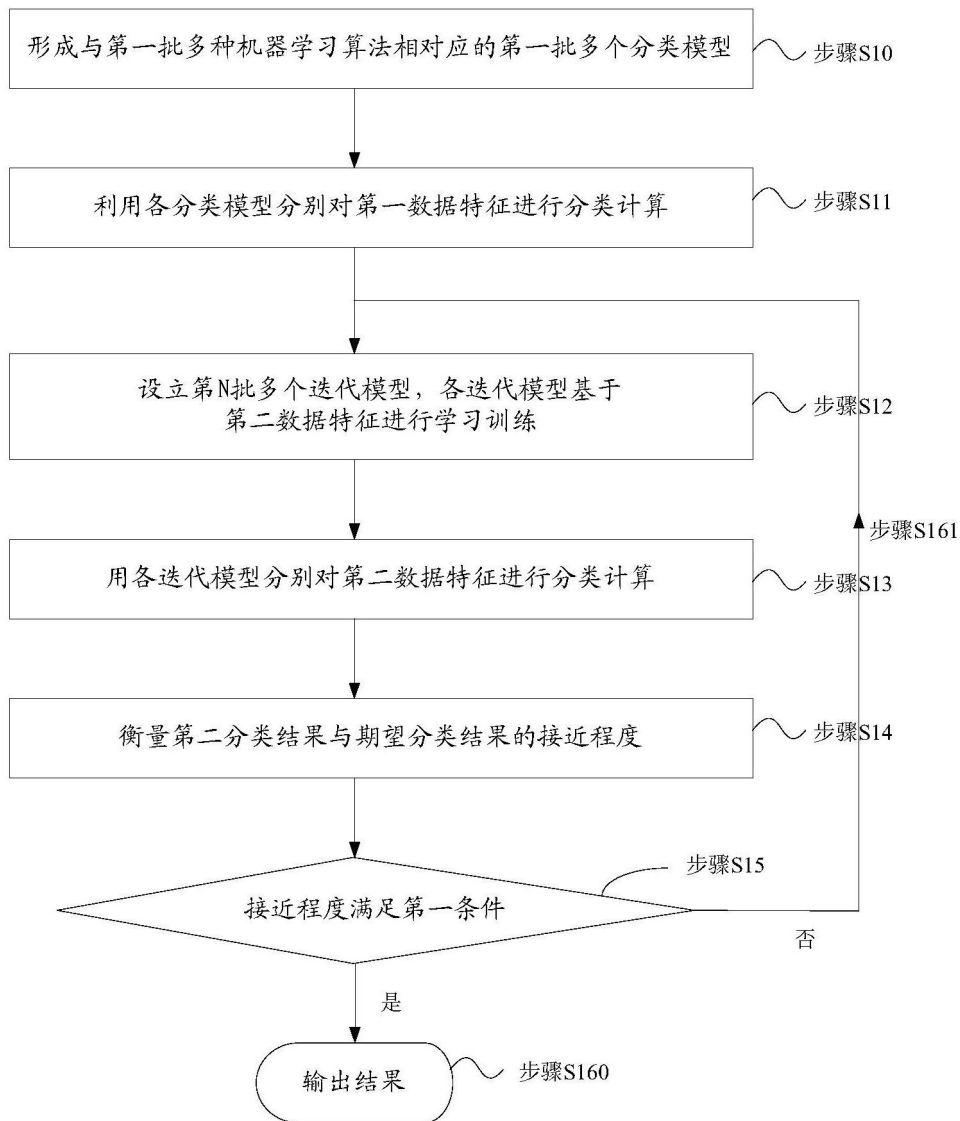


图1

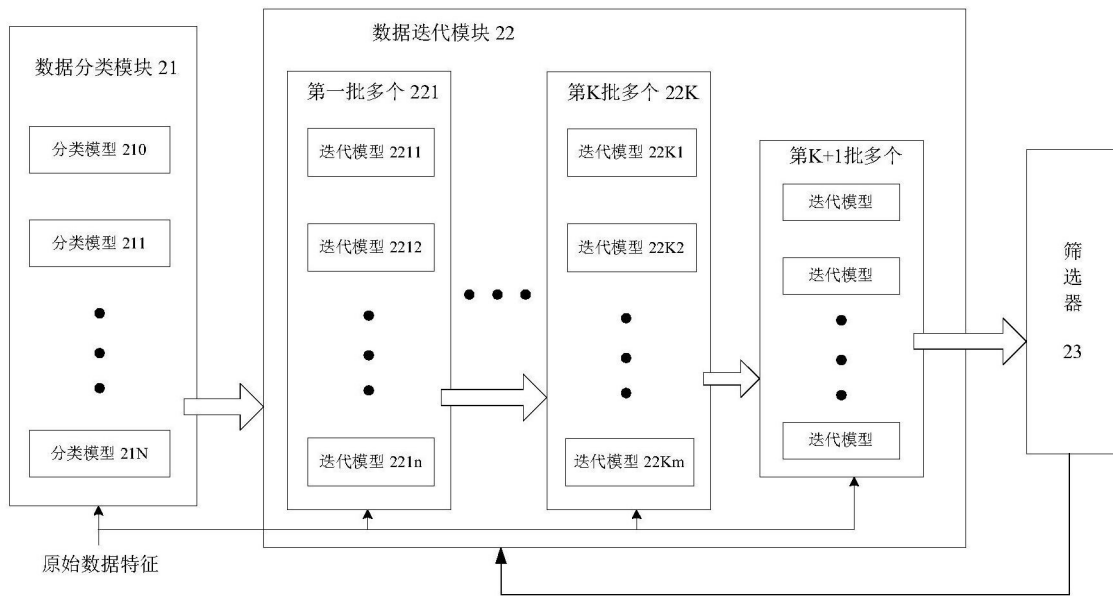


图2