(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112000808 B
(45)授权公告日 2024.04.16

(21)申请号 202011051021.8

(22)申请日 2020.09.29

(65)同一申请的已公布的文献号
申请公布号 CN 112000808 A

(43)申请公布日 2020.11.27

(73)专利权人 迪爱斯信息技术股份有限公司
地址 200233 上海市徐汇区钦江路333号41幢三层

(72)发明人 杜浸 王聚全 邱祥平 雷霆 彭明喜 苏永煜 邱雷 索涛 刘冉东 杨博 陈健 孙骞 张利

(74)专利代理机构 上海硕力知识产权代理事务所(普通合伙) 31251
专利代理师 童素珠

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/215(2019.01)
G06F 18/214(2023.01)
G06N 20/20(2019.01)

(56)对比文件
CN 106682192 A,2017.05.17
CN 110457675 A,2019.11.15
CN 111291185 A,2020.06.16
WO 2019233297 A1,2019.12.12
WO 2014149972 A1,2014.09.25
CN 102402713 A,2012.04.04
CN 105426826 A,2016.03.23
CN 110610193 A,2019.12.24
CN 110826494 A,2020.02.21
CN 111144475 A,2020.05.12
CN 111444945 A,2020.07.24
US 2012089552 A1,2012.04.12

审查员 汪安

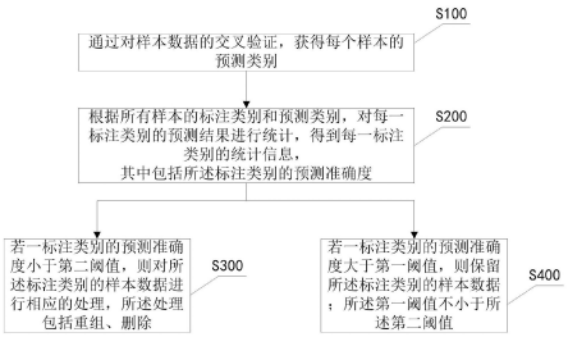
权利要求书2页 说明书10页 附图3页

(54)发明名称

一种数据处理方法及装置、可读存储介质

(57)摘要

本发明提供了一种数据处理方法及装置、可读存储介质，包括：通过对样本数据的交叉验证，获得每个样本的预测类别；根据所有样本的标注类别和预测类别，对每一标注类别的预测结果进行统计，得到每一标注类别的统计信息，其中包括所述标注类别的预测准确度；若一标注类别的预测准确度小于第二阈值，则对所述标注类别的样本数据进行相应的处理，所述处理包括重组、删除。本发明可以解决因标注错误导致的训练数据质量不高的问题，提升用于分类模型的训练数据的质量，从而提高分类模型的分类准确度。



CN 112000808 B

1.一种用于文本分类的数据处理方法,其特征在于,包括:

通过对样本数据的交叉验证,获得每个样本在每一类别下的概率信息;将每个样本的最大概率信息所对应的类别,作为所述样本的预测文本类别;所述样本数据为文本数据;

根据所有样本在同一类别下的概率信息,得到所述类别的概率阈值;

当一样本的预测文本类别的概率信息大于与所述预测文本类别相同的类别的概率阈值,则所述样本为有效预测样本;

根据所有有效预测样本的标注文本类别和预测文本类别,对每一标注文本类别的预测结果进行统计,得到每一标注文本类别的统计信息,所述标注文本类别的统计信息包括所述标注文本类别的样本数据被预测为各个类别的统计信息,其中所述标注文本类别的样本数据被预测为自身类别的统计信息记为所述标注文本类别的预测准确度;

若一标注文本类别的预测准确度小于第二阈值,则根据所述标注文本类别的统计信息对所述标注文本类别的样本数据进行相应的处理,所述处理包括重组、删除。

2.根据权利要求1所述的数据处理方法,其特征在于:

所述的根据所述标注文本类别的统计信息对所述标注文本类别的样本数据进行相应的处理,包括:

根据所述标注文本类别的统计信息,获取其中预测占比最大的两类类别的统计信息;

若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注文本类别的样本数据的标注文本类别更新为预测占比最大值对应的预测文本类别。

3.根据权利要求1所述的数据处理方法,其特征在于:

所述的根据所述标注文本类别的统计信息对所述标注文本类别的样本数据进行相应的处理,包括:

若所述标注文本类别的统计信息中,存在一个不小于2的N值,使得预测占比最大的N个值组成样本的样本方差小于第四阈值,则删除所述标注文本类别的样本数据。

4.根据权利要求1所述的数据处理方法,其特征在于,包括:

若一标注文本类别的预测准确度大于第一阈值,则保留所述标注文本类别的样本数据;所述第一阈值不小于所述第二阈值。

5.根据权利要求1所述的数据处理方法,其特征在于,根据所有有效预测样本的标注文本类别和预测文本类别,对每一标注文本类别的预测结果进行统计,得到每一标注文本类别的统计信息,包括:

根据所有有效预测样本构建混淆矩阵,其中,标注文本类别和预测文本类别分别为所述混淆矩阵的行列特征,每个元素为满足所述混淆矩阵的行列特征要求的样本数量或满足所述混淆矩阵的行列特征要求的样本数量占对应标注文本类别的样本总数的比例;

根据所述混淆矩阵得到每一标注文本类别的样本数据被预测为各个类别的统计信息。

6.根据权利要求5所述的数据处理方法,其特征在于,若一标注文本类别的预测准确度小于第二阈值,则根据所述标注文本类别的统计信息对所述标注文本类别的样本数据进行相应的处理,包括:

若所述混淆矩阵的行特征为标注文本类别,且一行对角线元素的值小于第二阈值,且该行最大两个元素的差值的绝对值大于第三阈值,则将该行所代表的标注文本类别的样本数据的标注文本类别更新为该行最大元素值对应的预测文本类别;或,

若所述混淆矩阵的列特征为标注文本类别,且一列对角线元素的值小于第二阈值,且该列最大两个元素的差值的绝对值大于第三阈值,则将该列所代表的标注文本类别的样本数据的标注文本类别更新为该列最大元素值对应的预测文本类别。

7.根据权利要求6所述的数据处理方法,其特征在于,还包括:

若所述混淆矩阵的行特征为标注文本类别,且一行对角线元素的值小于第二阈值,且该行最大的N个值组成样本的样本方差小于第四阈值,N不小于2,则删除该行所代表的标注文本类别的样本数据;或,

若所述混淆矩阵的列特征为标注文本类别,且一列对角线元素的值小于第二阈值,且该列最大的N个值组成样本的样本方差小于第四阈值,N不小于2,则删除该列所代表的标注文本类别的样本数据。

8.一种用于文本分类的数据处理装置,其特征在于,包括:

交叉验证模块,用于通过对样本数据的交叉验证,获得每个样本在每一类别下的概率信息;将每个样本的最大概率信息所对应的类别,作为所述样本的预测文本类别;所述样本数据为文本数据;根据所有样本在同一类别下的概率信息,得到所述类别的概率阈值;当一样本的预测文本类别的概率信息大于与所述预测文本类别相同的类别的概率阈值,则所述样本为有效预测样本;

信息统计模块,用于根据所有有效预测样本的标注文本类别和预测文本类别,对每一标注文本类别的预测结果进行统计,得到每一标注文本类别的统计信息,所述标注文本类别的统计信息包括所述标注文本类别的样本数据被预测为各个类别的统计信息,其中所述标注文本类别的样本数据被预测为自身类别的统计信息记为所述标注文本类别的预测准确度;

数据处理模块,用于若一标注文本类别的预测准确度小于第二阈值,则根据所述标注文本类别的统计信息对所述标注文本类别的样本数据进行相应的处理,所述处理包括重组、删除。

9.根据权利要求8所述的数据处理装置,其特征在于:

所述信息统计模块,进一步用于根据所述标注文本类别的统计信息,获取其中预测占比最大的两类类别的统计信息,并判断所述预测占比最大的两类类别的统计信息的差值是否大于第三阈值;

所述数据处理模块,进一步用于若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注文本类别的样本数据的标注文本类别更新为预测占比最大值对应的预测文本类别。

10.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1-7任一项所述的用于文本分类的数据处理方法的步骤。

一种数据处理方法及装置、可读存储介质

技术领域

[0001] 本发明涉及自然语言处理技术领域,尤指一种数据处理方法及装置、可读存储介质。

背景技术

[0002] 文本分类问题是自然语言处理领域关键的任务之一,程序通过理解文本中的自然语言,可以为文本选择最匹配的类别。文本分类任务被广泛应用于推荐算法、数据分析、垃圾信息过滤等诸多领域。然而,在实际应用中文本分类器的效果由于训练数据质量不稳定的原因,效果难以把控。

[0003] 在警情分类任务中,面临着以下问题:(1)警情类别众多,类别划分详细,且某些类之间的边界不明确,有时人都无法区分,因此会出现部分标注错误的噪音数据;(2)每个类别的数据不均衡,部分类别的样本很少,使得现有分类模型的分类效果差。

[0004] 一般的数据处理方法对数据质量高和数量较高的类别分类效果较好,反之则显得乏善可陈。

发明内容

[0005] 本发明的目的是提供一种数据处理方法及装置、可读存储介质,用于解决现有训练数据中因标注错误导致的数据噪音问题。

[0006] 本发明提供的技术方案如下:

[0007] 一种数据处理方法,包括:通过对样本数据的交叉验证,获得每个样本的预测类别;根据所有样本的标注类别和预测类别,对每一标注类别的预测结果进行统计,得到每一标注类别的统计信息,其中包括所述标注类别的预测准确度;若一标注类别的预测准确度小于第二阈值,则对所述标注类别的样本数据进行相应的处理,所述处理包括重组、删除。

[0008] 进一步地,所述标注类别的统计信息包括所述标注类别的样本数据被预测为各个类别的统计信息;

[0009] 所述的对所述标注类别的样本数据进行相应的处理,包括:根据所述标注类别的统计信息,获取其中预测占比最大的两类类别的统计信息;若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注类别的样本数据的标注类别更新为预测占比最大值对应的预测类别。

[0010] 进一步地,所述标注类别的统计信息包括所述标注类别的样本数据被预测为各个类别的统计信息;所述的对所述标注类别的样本数据进行相应的处理,包括:若所述标注类别的统计信息中,存在一个不小于2的N值,使得预测占比最大的N个值组成样本的样本方差小于第四阈值,则删除所述标注类别的样本数据。

[0011] 进一步地,若一标注类别的预测准确度大于第一阈值,则保留所述标注类别的样本数据;所述第一阈值不小于所述第二阈值。

[0012] 进一步地,所述的通过对样本数据的交叉验证,获得每个样本的预测类别,包括:

通过对样本数据的K折交叉验证,获得每个样本在分类模型中每一类别下的概率信息;将每个样本的最大概率信息所对应的类别,作为所述样本的预测类别。

[0013] 进一步地,在得到每个样本的预测类别之后,还包括:根据所有样本在同一类别下的概率信息,得到所述类别的概率阈值;当一样本的预测类别的概率信息小于与所述预测类别相同的类别的概率阈值,则所述样本为无效预测样本,并删除所述无效预测样本。

[0014] 进一步地,在获得每一类别的概率阈值之后,还包括:当一样本的预测类别的概率信息大于与所述预测类别相同的类别的概率阈值,则所述样本为有效预测样本;所述的根据所有样本的标注类别和预测类别,对每一标注类别的预测结果进行统计,包括:根据所有有效预测样本的标注类别和预测类别,对每一标注类别的预测结果进行统计。

[0015] 本发明还提供一种数据处理装置,包括:交叉验证模块,用于通过对样本数据的交叉验证,获得每个样本的预测类别;信息统计模块,用于根据所有样本的标注类别和预测类别,对每一标注类别的预测结果进行统计,得到每一标注类别的统计信息,其中包括所述标注类别的预测准确度;数据处理模块,用于若一标注类别的预测准确度小于第二阈值,则对所述标注类别的样本数据进行相应的处理,所述处理包括重组、删除。

[0016] 进一步地,所述标注类别的统计信息包括所述标注类别的样本数据被预测为各个类别的统计信息;所述信息统计模块,进一步用于根据所述标注类别的统计信息,获取其中预测占比最大的两类类别的统计信息,并判断所述预测占比最大的两类类别的统计信息的差值是否大于第三阈值;所述数据处理模块,进一步用于若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注类别的样本数据的标注类别更新为预测占比最大值对应的预测类别。

[0017] 本发明还提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现如前所述的数据处理方法的步骤。

[0018] 通过本发明提供的一种数据处理方法及装置、可读存储介质,至少能够带来以下有益效果:

[0019] 1、本发明通过识别不同类别间的相关程度,并根据该相关程度对训练数据进行保留、重组、删除等处理,提升了训练数据质量,从而让模型可以更好地学习到类别的特征,提高了分类模型的分类准确度,解决了现有训练数据中因标注错误导致的数据噪音问题。

[0020] 2、本发明通过剔除无效预测样本,进一步提升了训练数据的质量,提高了分类模型的分类准确度。

附图说明

[0021] 下面将以明确易懂的方式,结合附图说明优选实施方式,对一种数据处理方法及装置、可读存储介质的上述特性、技术特征、优点及其实现方式予以进一步说明。

[0022] 图1是本发明的一种数据处理方法的一个实施例的流程图;

[0023] 图2是本发明的一种数据处理方法的另一个实施例的流程图;

[0024] 图3是本发明的一种数据处理装置的一个实施例的结构示意图;

[0025] 图4是10折交叉验证的一种示意图。

[0026] 附图标号说明:

[0027] 100.交叉验证模块,200.信息统计模块,300.数据处理模块。

具体实施方式

[0028] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对照附图说明本发明的具体实施方式。显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图,并获得其他的实施方式。

[0029] 为使图面简洁,各图中只示意性地表示出了与本发明相关的部分,它们并不代表其作为产品的实际结构。另外,以使图面简洁便于理解,在有些图中具有相同结构或功能的部件,仅示意性地绘制了其中的一个,或仅标出了其中的一个。在本文中,“一个”不仅表示“仅此一个”,也可以表示“多于一个”的情形。

[0030] 本发明的一个实施例,如图1所示,一种数据处理方法,包括:

[0031] 步骤S100通过对样本数据的交叉验证,获得每个样本的预测类别。

[0032] 通常训练后的模型对于训练集的拟合程度还是挺好的,但是对于训练集之外的数据的拟合程度就不那么令人满意。因此通常会对所有的数据集进行分组,一部分作为训练集,另一部分作为验证集,首先用训练集对分类模型进行训练,再利用验证集测试训练得到的模型,以此来评价分类模型的性能。这种思想称为交叉验证。

[0033] 具体地,通过对样本数据的交叉验证,获得每个样本在每一类别下的概率信息;将每个样本的最大概率信息所对应的类别,作为所述样本的预测类别。

[0034] 可选地,采用K折交叉验证。定义好分类器模型,将整个样本数据等比例划分为K份,选取其中一份作为测试数据,另外K-1份作为训练数据,这样一共可以得到K个分类器。每个分类器在训练完毕后,可以对测试数据中的每个样本在每个类别上的概率做出预测,根据K组测试数据的预测结果,就可以得到整个样本数据中每个样本在每一类别下的概率信息。

[0035] 如图4所示,以常用的10折交叉验证为例,将全样本数据集分成十份,轮流将其中9份做训练1份做验证。测试结果1包括D10中每个样本在每一类别下的概率;测试结果2包括D9中每个样本在每一类别下的概率,依次类推,根据10次的测试结果,就可以得到全样本数据集中每个样本在每一类别下的概率。

[0036] 从一个样本在所有类别下的概率信息中选取最大概率信息所对应的类别,作为所述样本的预测类别。

[0037] 步骤S200根据所有样本的标注类别和预测类别,对每一标注类别的预测结果进行统计,得到每一标注类别的统计信息,其中包括所述标注类别的预测准确度。

[0038] 具体地,在对分类模型训练前需要对样本数据中的每个样本进行类别标注,通常由人工标注,该类别称为标注类别。若无标注错误,该类别相当于真实类别。

[0039] 根据所有样本的标注类别和预测类别,对每一标注类别的样本数据的预测结果进行统计,得到每一标注类别的统计信息。

[0040] 标注类别的统计信息包括对应标注类别的预测准确度。标注类别的预测准确度是指标注类别的样本数据被预测为自身类别的统计信息,该统计信息可以是被预测为自身类别的样本数,也可以是被预测为自身类别的样本数相对该标注类别样本总数的比例。

[0041] 比如,整个样本数据存在3种类别,分别为类别1-3;对每个样本进行标注,标注类别也对应有3种,且标注类别1对应类别1,其他依次类推;预测类别也对应3种,且预测类别1

对应类别1,其他依次类推。按标注类别划分,整个样本数据可分为标注类别1-3的样本数据。

[0042] 以标注类别1的样本数据为例,其中有的样本的预测类别等于标注类别1,即被预测为自身类别的样本;有的样本的预测类别不等于标注类别1,即被预测为其他类别的样本。其中被预测为自身类别的样本,说明分类模型对该样本预测准确。所以预测类别等于标注类别1的样本数,或预测类别等于标注类别1的样本数相对标注类别1的样本总数的比例,可以作为标注类别1的预测准确度。

[0043] 可选地,标注类别的统计信息包括所述标注类别的样本数据被预测为各个类别的统计信息。

[0044] 承接上述例子,标注类别1的统计信息除了包括标注类别1的样本数据中被预测为类别1的统计信息,还包括被预测为类别2、3的统计信息。

[0045] 步骤S300若一标注类别的预测准确度小于第二阈值,则对所述标注类别的样本数据进行相应的处理,所述处理包括重组、删除。

[0046] 若一标注类别的预测准确度小于第二阈值,说明该标注类别的样本数据存在较大标注错误,影响了分类模型的预测准确度,所以需要对该标注类别的样本数据进行重组或删除等处理,以提高样本数据的质量。

[0047] 对一标注类别的样本数据进行重组,是指将该标注类别的样本数据中每个样本的标注类别改为另一标注类别,比如,标注类别1的样本数据大部分实际上为类别2的数据,则需要将其中每个样本的标注类别改为标注类别2。

[0048] 对一标注类别的样本数据进行删除,是指从样本数据中删除该标注类别的样本数据。比如,标注类别1的样本数据很杂,有的为类别1的数据,有的为类别2的数据,有的为类别3的数据,标注错误分布较均匀,在分类中带来白色噪音的干扰效果,所以将标注类别1的样本数据全部删除。

[0049] 可选地,根据标注类别的统计信息,获取其中预测占比最大的两类类别的统计信息;若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注类别的样本数据的标注类别更新为预测占比最大值对应的预测类别。

[0050] 可选地,若所述标注类别的统计信息中,存在一个不小于2的N值,使得预测占比最大的N个值组成样本的样本方差小于第四阈值,则删除所述标注类别的样本数据。

[0051] 步骤S400若一标注类别的预测准确度大于第一阈值,则保留所述标注类别的样本数据;所述第一阈值不小于所述第二阈值。

[0052] 第一阈值与第二阈值可相等,也可不等,取决于分类模型的精度要求。

[0053] 若一标注类别的预测准确度大于第一阈值,比如第一阈值为90%,说明该标注类别的样本数据大部分都预测准确,标注类别全部正确或大部分正确,对这类样本数据需要全部保留。

[0054] 对样本数据进行上述处理后,训练数据质量得到提升,再将处理后的样本数据用于分类模型的训练,可提高分类模型的预测准确度。

[0055] 本实施例,通过交叉验证获取每个样本的预测类别,根据所有样本的标注类别和预测类别,得到每一标注类别的预测准确度;根据每一标注类别的预测准确度对该标注类别的样本数据进行保留、重组、删除等处理,实现了根据类别间的相关程度对数据进行合并

或删除,提升了训练数据质量,从而让模型可以更好地学习到类别的特征,提高了分类模型的分类准确度。

[0056] 本发明的另一个实施例,如图2所示,一种数据处理方法,包括:

[0057] 步骤S110通过对样本数据的K折交叉验证,获得每个样本在每一类别下的概率信息。

[0058] 步骤S120将每个样本的最大概率信息所对应的类别,作为所述样本的预测类别。

[0059] 步骤S130根据所有样本在同一类别下的概率信息,得到所述类别的概率阈值。

[0060] 比如,对所有样本在同一类别下的概率信息取平均值,得到所述类别的概率阈值。

[0061] 可选地,根据每个样本在每一类别下的概率信息,得到一个样本概率矩阵,该矩阵的每一行代表一个样本,每一列表示该样本出现在该类别下的概率。通过对每一列取平均值,得到每一类别下的概率阈值。

[0062] 步骤S140当一样本的预测类别的概率信息小于与所述预测类别相同的类别的概率阈值,则所述样本为无效预测样本,并删除所述无效预测样本。

[0063] 步骤S150当一样本的预测类别的概率信息大于与所述预测类别相同的类别的概率阈值,则所述样本为有效预测样本。

[0064] 步骤S210根据所有有效预测样本的标注类别和预测类别,对每一标注类别的预测结果进行统计,得到每一标注类别的统计信息,其中包括所述标注类别的样本数据被预测为各个类别的统计信息。

[0065] 可选地,将所有标注类别的统计信息通过混淆矩阵的形式呈现。比如,通常以标注类别为行特征、预测类别为列特征,构建混淆矩阵。当然理论上也可以标注类别为列特征、预测类别为行特征,构建混淆矩阵。

[0066] 以标注类别为行特征、预测类别为列特征为例,混淆矩阵的每一行代表了数据的真实类别(即标注类别),矩阵的每一列代表模型的预测类别,每一个单元则代表某一标注类别与某一预测类别出现重叠的样本数量。每一行之和表示标注类别为该类别的样本数量。每一列之和表示预测类别为该类别的样本数量。

[0067] 示例,假设有150个样本数据,预测为1,2,3类各为50个,得到如下混淆矩阵,每个元素采用满足要求的样本数量:

[0068]

		预测类别		
		类 1	类 2	类 3
真实类别	类 1	43	2	0
	类 2	5	45	1
	类 3	2	3	49

[0069] 混淆矩阵中每个元素可以是满足要求的样本数量,也可以是满足要求的样本数量占对应标注类别的样本数据的总数的比例。对角线元素是标注类别与预测类别一致的样本数量或占对应标注类别的样本数据的总数的比例,其值反映了该标注类别的预测准确度。

[0070] 为了后续描述方便,在本实施例中将每个元素的取值统一定义为满足要求的样本数量占对应标注类别的样本数据的总数的比例;将混淆矩阵的行特征定义为标注类别。不过本申请并不限定将元素的取值定义为满足要求的样本数量,对应的各种阈值做相应调整

即可;也不限定将混淆矩阵的列特征定义为标注类别,只要对应的判断措施做相应的调整即可。

[0071] 将元素取值定义为比例,则相当于对混淆矩阵的每一行进行归一化处理,即使每一行所有单元内的数据之和为1。

[0072] 步骤S310若一标注类别的预测准确度小于第二阈值,则根据所述标注类别的统计信息,获取其中预测占比最大的两类类别的统计信息;

[0073] 步骤S320若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注类别的样本数据的标注类别更新为预测占比最大值对应的预测类别。

[0074] 具体地,若某行对角线元素的值小于规定概率阈值P0,且该行内元素值最大的两个元素的差值的绝对值大于规定概率差值P1(0<P1<1),则表明该行所代表类别的判定受另一个类别数据的影响较大,该类类别记为情况C。

[0075] 对于情况C的类别,因为该类类别受另一个类别ci数据的影响较大,会将该类类别数据并入类别ci。

[0076] 步骤S330若一标注类别的预测准确度小于第二阈值,且所述标注类别的统计信息中,存在一个不小于2的N值,使得预测占比最大的N个值组成样本的样本方差小于第四阈值,则删除所述标注类别的样本数据。

[0077] 若某行对角线元素的值小于规定概率阈值P0,且该行内元素值最大的N个值组成样本的样本方差小于规定方差阈值P2,则判断该行所代表类别受其他多个类别数据的影响较大,该类类别记为情况D。

[0078] 对于情况D的类别,因为该类类别受其他多个类别数据ci,ci+1,…,cj的影响较大,会将该类类别数据视为噪音数据进行删除。

[0079] 步骤S400若一标注类别的预测准确度大于第一阈值,则保留所述标注类别的样本数据;所述第一阈值不小于所述第二阈值。

[0080] 具体地,若混淆矩阵某行对角线元素的值大于规定概率阈值P0(0<P0<1),则判断该行所代表类别受其他类别数据的影响较小,该类类别记为情况S。

[0081] 对于情况S的类别,因为该类类别受其他类别数据影响的程度较小,对此类类别数据会进行完整的保留。

[0082] 本实施例,通过对样本数据进行K折交叉验证,获取每一个样本在不同类别下的概率信息以及每个类别的阈值,根据每个类别的阈值识别有效预测样本,剔除无效预测样本;根据有效预测样本组建混淆矩阵;如果混淆矩阵对角线上的元素的值大于规定的阈值,判断该类判定受其他类别数据的影响小,则保留对该类别数据;如果混淆矩阵对角线上元素的值小于规定阈值但本行最大两个元素的差值大于规定阈值,判断该类别判定受另外一个类别数据的影响比较大,则将该类别数据并入另外一个类别中;如果混淆矩阵对角元素的值小于规定阈值且该行值最大的N个值组成样本的样本方差小于规定阈值,判断该类别受其他多个类别数据影响很大,则将该类别数据作为整体的噪音数据进行删除。因此,即使训练数据中存在人为的标注错误和/或相关的类别交叠情况,本实施例都能有效地减少模型训练数据中的噪声,能够明显地提高模型的拟合程度以及预测数据的准确率。因为该方法实际上将类别进行了重组,根据类别间的相关程度进行合并或删除,从而让模型可以更好地学习到类别的特征。

[0083] 本发明的一个实施例,如图3所示,一种数据处理装置,包括:

[0084] 交叉验证模块100,用于通过对样本数据的交叉验证,获得每个样本的预测类别。

[0085] 通常训练后的模型对于训练集的拟合程度还是挺好的,但是对于训练集之外的数据的拟合程度就不那么令人满意。因此通常会对所有的数据集进行分组,一部分作为训练集,另一部分作为验证集,首先用训练集对分类模型进行训练,再利用验证集测试训练得到的模型,以此来评价分类模型的性能。这种思想称为交叉验证。

[0086] 具体地,通过对样本数据的交叉验证,获得每个样本在每一类别下的概率信息;将每个样本的最大概率信息所对应的类别,作为所述样本的预测类别。

[0087] 可选地,采用K折交叉验证。定义好分类器模型,将整个样本数据等比例划分为K份,选取其中一份作为测试数据,另外K-1份作为训练数据,这样一共可以得到K个分类器。每个分类器在训练完毕后,可以对测试数据中的每个样本在每个类别上的概率做出预测,根据K组测试数据的预测结果,就可以得到整个样本数据中每个样本在每一类别下的概率信息。

[0088] 如图4所示,以常用的10折交叉验证为例,将全样本数据集分成十份,轮流将其中9份做训练1份做验证。测试结果1包括D10中每个样本在每一类别下的概率;测试结果2包括D9中每个样本在每一类别下的概率,依次类推,根据10次的测试结果,就可以得到全样本数据集中每个样本在每一类别下的概率。

[0089] 从一个样本在所有类别下的概率信息中选取最大概率信息所对应的类别,作为所述样本的预测类别。

[0090] 信息统计模块200,用于根据所有样本的标注类别和预测类别,对每一标注类别的预测结果进行统计,得到每一标注类别的统计信息,其中包括所述标注类别的预测准确度。

[0091] 具体地,在对分类模型训练前需要对样本数据中的每个样本进行类别标注,通常由人工标注,该类别称为标注类别。若无标注错误,该类别相当于真实类别。

[0092] 根据所有样本的标注类别和预测类别,对每一标注类别的样本数据的预测结果进行统计,得到每一标注类别的统计信息。

[0093] 标注类别的统计信息包括对应标注类别的预测准确度。标注类别的预测准确度是指标注类别的样本数据被预测为自身类别的统计信息,该统计信息可以是被预测为自身类别的样本数,也可以是被预测为自身类别的样本数相对该标注类别样本总数的比例。

[0094] 比如,整个样本数据存在3种类别,分别为类别1-3;对每个样本进行标注,标注类别也对应有3种,且标注类别1对应类别1,其他依次类推;预测类别也对应3种,且预测类别1对应类别1,其他依次类推。按标注类别划分,整个样本数据可分为标注类别1-3的样本数据。

[0095] 以标注类别1的样本数据为例,其中有的样本的预测类别等于标注类别1,即被预测为自身类别的样本;有的样本的预测类别不等于标注类别1,即被预测为其他类别的样本。其中被预测为自身类别的样本,说明分类模型对该样本预测准确。所以预测类别等于标注类别1的样本数,或预测类别等于标注类别1的样本数相对标注类别1的样本总数的比例,可以作为标注类别1的预测准确度。

[0096] 可选地,标注类别的统计信息包括所述标注类别的样本数据被预测为各个类别的统计信息。

[0097] 承接上述例子,标注类别1的统计信息除了包括标注类别1的样本数据中被预测为类别1的统计信息,还包括被预测为类别2、3的统计信息。

[0098] 数据处理模块300,用于若一标注类别的预测准确度小于第二阈值,则对所述标注类别的样本数据进行相应的处理,所述处理包括重组、删除。

[0099] 若一标注类别的预测准确度小于第二阈值,说明该标注类别的样本数据存在较大标注错误,影响了分类模型的预测准确度,所以需要对该标注类别的样本数据进行重组或删除等处理,以提高样本数据的质量。

[0100] 对一标注类别的样本数据进行重组,是指将该标注类别的样本数据中每个样本的标注类别改为另一标注类别,比如,标注类别1的样本数据大部分实际上为类别2的数据,则需要将其中每个样本的标注类别改为标注类别2。

[0101] 对一标注类别的样本数据进行删除,是指从样本数据中删除该标注类别的样本数据。比如,标注类别1的样本数据很杂,有的为类别1的数据,有的为类别2的数据,有的为类别3的数据,标注错误分布较均匀,在分类中带来白色噪音的干扰效果,所以将标注类别1的样本数据全部删除。

[0102] 可选地,根据标注类别的统计信息,获取其中预测占比最大的两类类别的统计信息;若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注类别的样本数据的标注类别更新为预测占比最大值对应的预测类别。

[0103] 可选地,若所述标注类别的统计信息中,存在一个不小于2的N值,使得预测占比最大的N个值组成样本的样本方差小于第四阈值,则删除所述标注类别的样本数据。

[0104] 数据处理模块300,进一步用于若一标注类别的预测准确度大于第一阈值,则保留所述标注类别的样本数据;所述第一阈值不小于所述第二阈值。

[0105] 第一阈值与第二阈值可相等,也可不等,取决于分类模型的精度要求。

[0106] 若一标注类别的预测准确度大于第一阈值,比如第一阈值为90%,说明该标注类别的样本数据大部分都预测准确,标注类别全部正确或大部分正确,对这类样本数据需要全部保留。

[0107] 对样本数据进行上述处理后,训练数据质量得到提升,再将处理后的样本数据用于分类模型的训练,可提高分类模型的预测准确度。

[0108] 本实施例,通过交叉验证获取每个样本的预测类别,根据所有样本的标注类别和预测类别,得到每一标注类别的预测准确度;根据每一标注类别的预测准确度对该标注类别的样本数据进行保留、重组、删除等处理,实现了根据类别间的相关程度对数据进行合并或删除,提升了训练数据质量,从而让模型可以更好地学习到类别的特征,提高了分类模型的分类准确度。

[0109] 本发明的另一个实施例,如图3所示,一种数据处理装置,包括:

[0110] 交叉验证模块100,用于通过对样本数据的K折交叉验证,获得每个样本在每一类别下的概率信息;将每个样本的最大概率信息所对应的类别,作为所述样本的预测类别;根据所有样本在同一类别下的概率信息,得到所述类别的概率阈值;当一样本的预测类别的概率信息小于与所述预测类别相同的类别的概率阈值,则所述样本为无效预测样本;当一样本的预测类别的概率信息大于与所述预测类别相同的类别的概率阈值,则所述样本为有效预测样本。

[0111] 比如,对所有样本在同一类别下的概率信息取平均值,得到所述类别的概率阈值。

[0112] 可选地,根据每个样本在每一类别下的概率信息,得到一个样本概率矩阵,该矩阵的每一行代表一个样本,每一列表示该样本出现在该类别下的概率。通过对每一列取平均值,得到每一类别下的概率阈值。

[0113] 信息统计模块200,用于根据所有有效预测样本的标注类别和预测类别,对每一标注类别的预测结果进行统计,得到每一标注类别的统计信息,其中包括所述标注类别的样本数据被预测为各个类别的统计信息。

[0114] 可选地,将所有标注类别的统计信息通过混淆矩阵的形式呈现。比如,通常以标注类别为行特征、预测类别为列特征,构建混淆矩阵。当然理论上也可以标注类别为列特征、预测类别为行特征,构建混淆矩阵。

[0115] 以标注类别为行特征、预测类别为列特征为例,混淆矩阵的每一行代表了数据的真实类别(即标注类别),矩阵的每一列代表模型的预测类别,每一个单元则代表某一标注类别与某一预测类别出现重叠的样本数量。每一行之和表示标注类别为该类别的样本数量。每一列之和表示预测类别为该类别的样本数量。

[0116] 混淆矩阵中每个元素可以是满足要求的样本数量,也可以是满足要求的样本数量占对应标注类别的样本数据的总数的比例。对角线元素是标注类别与预测类别一致的样本数量或占对应标注类别的样本数据的总数的比例,其值反映了该标注类别的预测准确度。

[0117] 为了后续描述方便,在本实施例中将每个元素的取值统一定义为满足要求的样本数量占对应标注类别的样本数据的总数的比例;将混淆矩阵的行特征定义为标注类别。不过本申请并不限定将元素的取值定义为满足要求的样本数量,对应的各种阈值做相应调整即可;也不限定将混淆矩阵的列特征定义为标注类别,只要对应的判断措施做相应的调整即可。

[0118] 将元素取值定义为比例,则相当于对混淆矩阵的每一行进行归一化处理,即使每一行所有单元内的数据之和为1。

[0119] 数据处理模块300,用于删除所述无效预测样本;若一标注类别的预测准确度小于第二阈值,则根据所述标注类别的统计信息,获取其中预测占比最大的两类类别的统计信息;若所述预测占比最大的两类类别的统计信息的差值大于第三阈值,则将所述标注类别的样本数据的标注类别更新为预测占比最大值对应的预测类别。

[0120] 具体地,若某行对角线元素的值小于规定概率阈值P0,且该行内元素值最大的两个元素的差值的绝对值大于规定概率差值P1(0<P1<1),则表明该行所代表类别的判定受另一个类别数据的影响较大,该类类别记为情况C。

[0121] 对于情况C的类别,因为该类类别受另一个类别ci数据的影响较大,会将该类类别数据并入类别ci。

[0122] 数据处理模块300,进一步用于若一标注类别的预测准确度小于第二阈值,且所述标注类别的统计信息中,存在一个不小于2的N值,使得预测占比最大的N个值组成样本的样本方差小于第四阈值,则删除所述标注类别的样本数据。

[0123] 若某行对角线元素的值小于规定概率阈值P0,且该行内元素值最大的N个值组成样本的样本方差小于规定方差阈值P2,则判断该行所代表类别受其他多个类别数据的影响较大,该类类别记为情况D。

[0124] 对于情况D的类别,因为该类类别受其他多个类别数据ci,ci+1,…,cj的影响较大,会将该类类别数据视为噪音数据进行删除。

[0125] 数据处理模块300,进一步用于若一标注类别的预测准确度大于第一阈值,则保留所述标注类别的样本数据;所述第一阈值不小于所述第二阈值。

[0126] 具体地,若混淆矩阵某行对角线元素的值大于规定概率阈值P0(0<P0<1),则判断该行所代表类别受其他类别数据的影响较小,该类类别记为情况S。

[0127] 对于情况S的类别,因为该类类别受其他类别数据影响的程度较小,对此类类别数据会进行完整的保留。

[0128] 本实施例,通过对样本数据进行K折交叉验证,获取每一个样本在不同类别下的概率信息以及每个类别的阈值,根据每个类别的阈值识别有效预测样本,剔除无效预测样本;根据有效预测样本组建混淆矩阵;如果混淆矩阵对角线上的元素的值大于规定的阈值,判断该类判定受其他类别数据的影响小,则保留对该类别数据;如果混淆矩阵对角线上元素的值小于规定阈值但本行最大两个元素的差值大于规定阈值,判断该类别判定受另外一个类别数据的影响比较大,则将该类别数据并入另外一个类别中;如果混淆矩阵对角元素的值小于规定阈值且该行值最大的N个值组成样本的样本方差小于规定阈值,判断该类别受其他多个类别数据影响很大,则将该类别数据作为整体的噪音数据进行删除。因此,即使训练数据中存在人为的标注错误和/或相关的类别交叠情况,本实施例都能有效地减少模型训练数据中的噪声,能够明显地提高模型的拟合程度以及预测数据的准确率。因为该方法实际上将类别进行了重组,根据类别间的相关程度进行合并或删除,从而让模型可以更好地学习到类别的特征。

[0129] 需要说明的是,本发明提供的数据处理装置的实施例与前述提供的数据处理方法的实施例均基于同一发明构思,能够取得相同的技术效果。因而,所述数据处理装置的实施例的其它具体内容可以参照前述数据处理方法的实施例内容的记载。

[0130] 在本发明的一个实施例中,一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时可实现如前述实施例记载的数据处理方法。也即是,当前述本发明实施例对现有技术做出贡献的技术方案的部分或全部通过计算机软件产品的方式得以体现时,前述计算机软件产品存储在一个计算机可读存储介质中。所述计算机可读存储介质可以为任意可携带计算机程序代码实体装置或设备。譬如,所述计算机可读存储介质可以是U盘、移动磁盘、磁碟、光盘、计算机存储器、只读存储器、随机存取存储器等。

[0131] 应当说明的是,上述实施例均可根据需要自由组合。以上所述仅是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

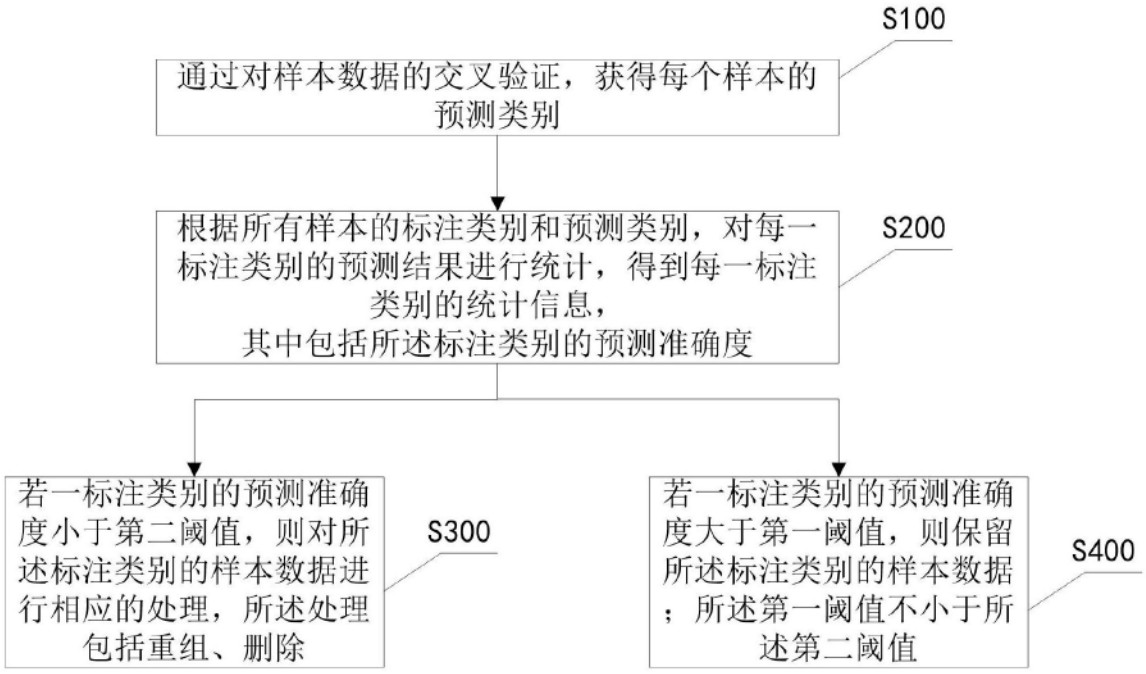

S100
通过对样本数据的交叉验证，获得每个样本的预测类别
S200
根据所有样本的标注类别和预测类别，对每一标注类别的预测结果进行统计，得到每一标注类别的统计信息，
其中包括所述标注类别的预测准确度
若一标注类别的预测准确度小于第二阈值，则对所述标注类别的样本数据进行相应的处理，所述处理包括重组、删除
S300
若一标注类别的预测准确度大于第一阈值，则保留所述标注类别的样本数据；所述第一阈值不小于所述第二阈值
S400

图1

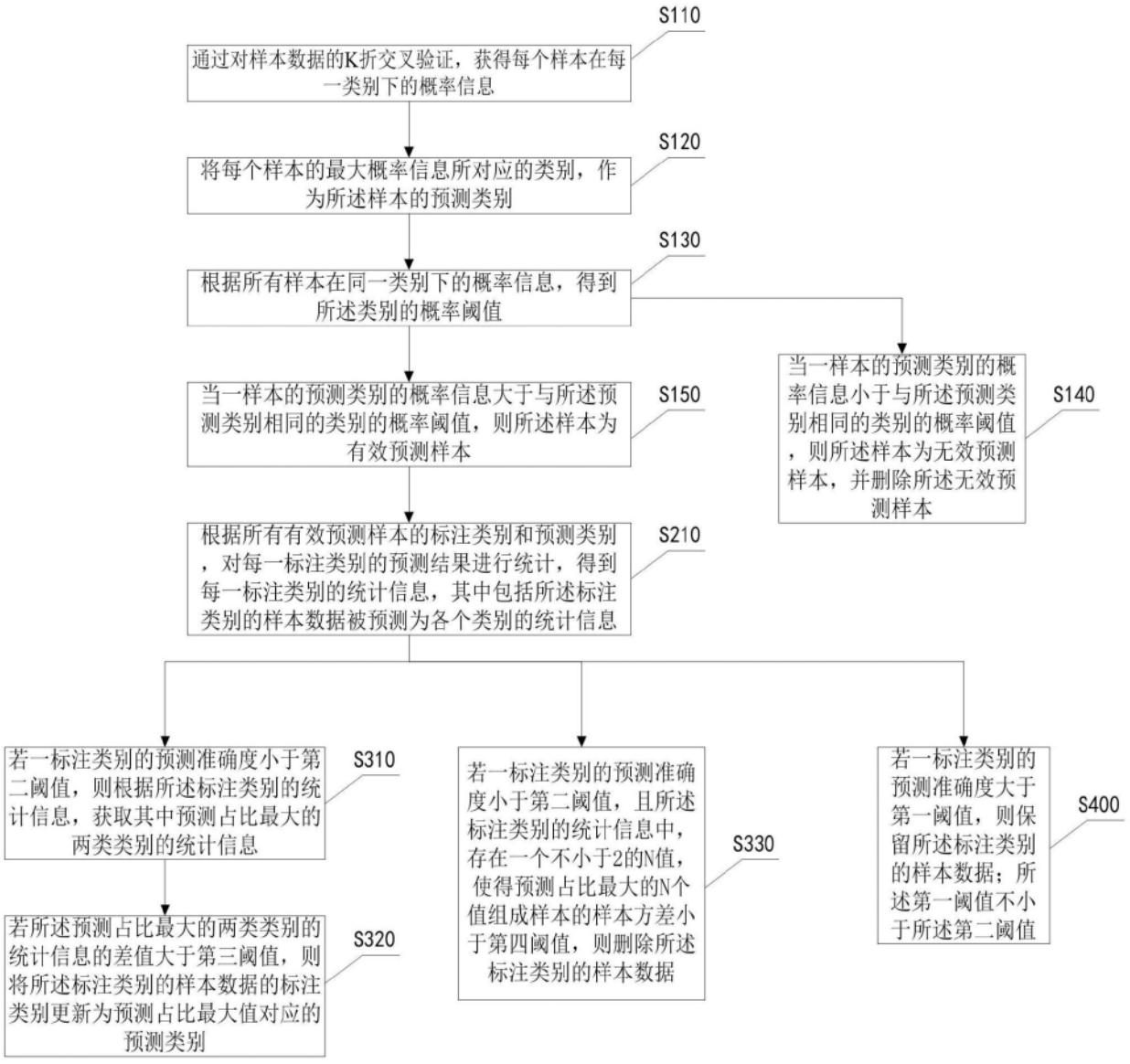

S110
通过对样本数据的K折交叉验证，获得每个样本在每一类别下的概率信息
S120
将每个样本的最大概率信息所对应的类别，作为所述样本的预测类别
S130
根据所有样本在同一类别下的概率信息，得到所述类别的概率阈值
S150
当一样本的预测类别的概率信息大于与所述预测类别相同的类别的概率阈值，则所述样本为有效预测样本
S140
当一样本的预测类别的概率信息小于与所述预测类别相同的类别的概率阈值，则所述样本为无效预测样本，并删除所述无效预测样本
S210
根据所有有效预测样本的标注类别和预测类别，对每一标注类别的预测结果进行统计，得到每一标注类别的统计信息，其中包括所述标注类别的样本数据被预测为各个类别的统计信息
S310
若一标注类别的预测准确度小于第二阈值，则根据所述标注类别的统计信息，获取其中预测占比最大的两类类别的统计信息
S320
若所述预测占比最大的两类类别的统计信息的差值大于第三阈值，则将所述标注类别的样本数据的标注类别更新为预测占比最大值对应的预测类别
S330
若一标注类别的预测准确度小于第二阈值，且所述标注类别的统计信息中，存在一个不小于2的N值，使得预测占比最大的N个值组成样本的样本方差小于第四阈值，则删除所述标注类别的样本数据
S400
若一标注类别的预测准确度大于第一阈值，则保留所述标注类别的样本数据；所述第一阈值不小于所述第二阈值

图2


100. 交叉验证模块
200. 信息统计模块
300. 数据处理模块

图3

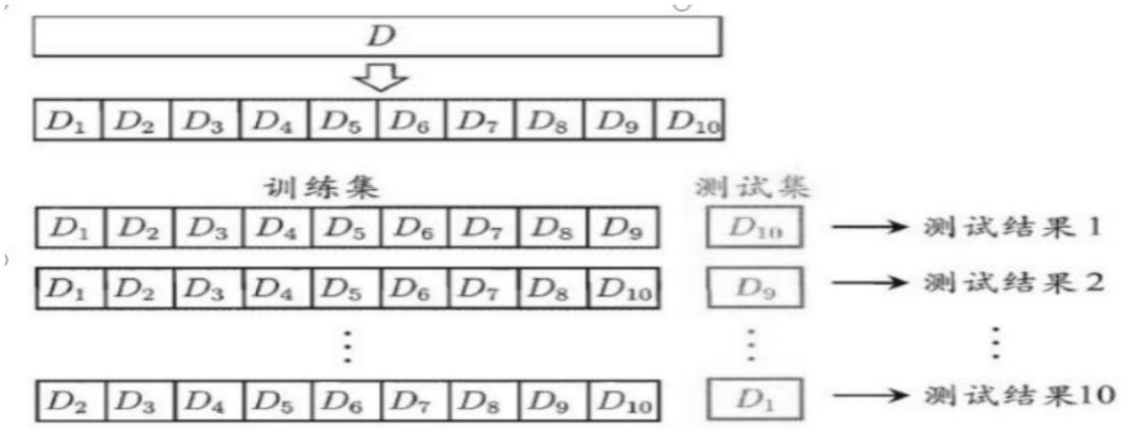

D
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10
训练集
测试集
D1 D2 D3 D4 D5 D6 D7 D8 D9
D10
测试结果1
D1 D2 D3 D4 D5 D6 D7 D8 D10
D9
测试结果2
D2 D3 D4 D5 D6 D7 D8 D9 D10
D1
测试结果10

图4