森林分布：异常检测方法研究隔离森林

成飞Yao1，小青MA1，彪臣1，高晓松Zhao2，和刚BAI1（）

计算机科学，南开大学，天津，中国2120160400@mail.nankai.edu.cn，baigang@nankai.edu.cn2天津市公安职业学院，天津，中国的高校1

抽象。异常检测是指在发现不符合预期的数据模式。异常检测有多种应用领域和场景，例如网络入侵检测，欺诈检测和故障检测。本文提出了通过隔离森林（iForest）启发了新的异常检测方法分布森林（dForest）。dForest建立的所谓分发树（DTREE）特殊二叉树的合奏。我们的方法的基本思想是通过数据的每个节点的分布，引导DTREE建设。和DTREE的每个节点被视为输入空间的子空间。当dForest而建，异常具有比正常情况下更短的路径长度。dForest有其他方法不同的解释。与iForest，LOF和INNE相比，所提出的方法实现了在不同的基准数据集AUC方面的竞争结果。此外，dForest执行以及在这两个半监督和无监督异常检测模式。

关键词：异常检测马氏距离隔离森林

1引言

异常检测是数据挖掘任务的一个重要分支，受到了广泛的关注。异常检测是指不符合期望[1]数据寻找模式的问题。这些意外的模式是指异常的，嘈杂，观察，事件，物品等。异常检测起着许多应用领域和应用场景，如网络入侵检测，欺诈检测和故障检测具有重要作用。的研究异常检测日期的历史可追溯至19世纪，并且存在各种各样的在这一领域的技术。有几种类型的包含基于聚类的方法[2]，基于密度的方法[3]，基于距离的方法[4]异常检测方法和其它方法[7,8]。大多数这些方法具有较高的时间复杂度和存储成本。其中，隔离林（iForest）是在[5，6]由刘和汀提出了一种基于异常的隔离检测方法。iForest通过轴线平行的分割，而不是正常的配置文件分离株的实例异常。iForest的主要优点是，它具有低存储器要求的线性时间复杂度。iForest的局限性在于它是局部不敏感



©斯普林格自然瑞士AG2019P.-C.红豆杉等。（编辑）：APPT2019，LNCS11719，第135-147，2019年https://doi.org/10.1007/978-3-030-29611-7\_11

全球得分异常[15]。本文提出了解决iForest的框架的上述问题的新方法。类似iForest，dForest建立dTrees的合奏。在构建DTREE的过程中，几个属性随机在每一个节点选择以形成所述特征空间的子空间。然后，每个节点的划分在子空间进行。在每个节点的分割，我们引入协方差估计和使用马哈拉诺比斯距离来确定超椭圆。在超椭圆内的点分为左子，和超椭圆外的点分为右孩子。每个内部节点根据局部区域的分布划分，从而使可能的正常情况下和反常的实例被分成不同的分支尽可能。因此，异常更容易被分成比正常情况下的叶节点。与其他方法相比，dForest具有以下优点：（1）dForest充分利用多维信息。在每个节点处，几个属性是随机选择的。大量的属性组合被用于形成不同的子空间。的数据被映射到不同的那些低维子空间。（2）dForest工作良好在两个半监督和无监督模式。在无监督异常检测的情况下，正常情况下和异常实例根据实例和在不同的子空间局部分布之间的Mahalanobis距离区分。在半的情况下监督异常检测，每个子空间的局部分布的边界被认为是其中异常可能出现的位置。（3）dForest有一个直观的和明确的解释。在每个节点处，该数据被投影到由随机选择的属性来确定子空间。这些节点根据数据对所述子空间的分布划分。在不同的子空间，正常情况下和异常根据分布在不断区分。本文的其余部分安排如下。第2节总结了相关工作简要介绍。第3所介绍的过程和所提出的方法的细节。第4节主要提供与其他方法和其他相关的实验比较的结果。最后，第四节。5简要给出结论。

2相关工作

根据标签是否可用，异常检测技术可以分为以下三种类型。

监督异常检测。该种技术，列车使用的是已被贴上了数据的正常与异常类的模型。但监督异常检测技术面临的两大挑战。首先，在正常情况下和异常情况是不平衡的。其次，获得准确的标签是非常困难的，尤其是对异常情况。

半监督异常检测。相关技术假设在训练数据只有正常情况下，这些标签。一般来说，半监督异常检测技术培训的正常数据和不符合模型的异常对待数据的模型。

136C.Yao等人。

无监督异常检测。这一类的技术不要求数据被标记，并得到广泛应用。但它通常假设正常情况下的数量比异常实例的数量大得多。与仅正常情况下或与未标记的训练数据训练：在本文所提出的方法主要执行在第二种和第三种情况。根据原理，异常检测技术可分为以下几种类型。

集群的基于异常检测技术。基于聚类的异常检测技术是基于以下认可：正常情况下属于某个集群，和异常不属于任何集群。还是正常情况下属于比较大的集群，而异常属于比较小的集群。这些技术的优势在于，他们可以在无人监管的模式下工作。该限制是性能主要取决于聚类算法。

基于密度的异常检测技术。这一类的技术假定与低密度的情况下更可能是一个异常，和一个密集的密度被宣布为一个正常的实例。的公知方法基于密度是本地异常因子（LOF）。基于密度的异常检测技术的主要优点是，他们是无监督的技术，并不需要对数据的分布做任何假设。其缺点是，大多数的这些技术需要找邻居和计算距离，所以时间复杂度是比较高的。

隔离基于异常检测技术。基于隔离异常检测技术的基本思想是，异常更容易比正常情况下要进行分离。这些技术分离异常，而不是曲线正常情况。所述第一基于隔离法是iForest。iForest建立隔离树（iTree）的集合。一个iTree是从一个子样本建立了一个特殊的二进制树。在iTree每个内部节点根据对随机选择的属性随机选择的阈值划分。对于测试实例x，路径长度hðxÞ在每个iTree被收集在第一和平均路径长度EðhðxÞÞ被计算。由于在叶节点的实例的数目可以大于1是更大的，路径长度hðxÞ需要添加在叶节点使用数据构建子树的高度。对于大小的数据n，预期的平均路径长度C（N）可被表示如下。

CNdÞ¼2HN+1d2ÞN1dÞÑð1Þ

ħIDÞ¼LNIDÞþ0：5772156649ð2Þ

然后，异常分数x的s的作为式计算。（3）。越接近得分S为​​1时，更可能x是为异常。越接近得分S为​​0时，更可能x是为正常的实例。

小号X;ÑdÞ¼2E高xdÞdÞCNdÞð3Þ

iForest具有低存储器要求的线性时间复杂性，并且可以在许多问题，从而获得高精度。本文由iForest的启发，提出了一种新的异常检测方法，称为分布森林（dForest）。

森林分布：异常检测方法研究iForest137

3我们的方法

从不同的角度，提出了异常检测的新方法。树的每个节点被视为功能空间的子空间。该数据在子空间划分，在配送中心的数据分配给左子，在分布的边缘数据分为右孩子。二叉树的每个内部节点包含了数据的一部分，因此它可以被看作是局部地区的分布情况。通过节点的连续分裂，正常情况下和异常情况下在不同的子空间被连续区分。更少的数据被分为左右的孩子，这样的异常到达叶节点更快。因此，异常具有比正常情况下更短的平均路径长度。

3.1训练阶段

对于给定数据集X路，X0是大小W从数据集X.然后X0是用来建立一个DTREE一个子样本。设T为DTREE的内部节点。让铊是节点T的左孩子，让为tr在节点T节点T.所有实例的右子构成S.然后如下节点的分裂中描述了分配。首先，K属性是随机选择的，当节点被分割，以形成一个k维子空间。然后每个实例x和分布S之间的Mahalanobis距离被计算，根据式（4），其中，参数L和R被分布的均值和协方差矩阵，respectively.DMX;小号dÞ¼ffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiffiX升dÞTR1X升dÞqð4ÞThe马氏距离在[9]提出了通过马哈拉诺比斯。马哈拉诺比斯距离考虑了数据的相关性，并且可以用于计算样品和分配之间的距离。另外，马氏距离的比例不变。它被广泛用于聚类分析和分类技术[10]。马哈拉诺比斯距离可以用于异常检测[11]。一个实例可能异常，如果它已经从分布有很大马氏距离。在本文中，马哈拉诺比斯距离用于在子空间中的可能的异常和正常实例之间进行区分。如果实例XI之间的Mahalanobis距离我¼1;2;。。。;ñdTH和分布S小于P，它被分为左子。否则，它被分为左右的孩子。换言之，在特征空间中，其中，所述马哈拉诺比斯距离是等于p的地方构成超椭圆。在超椭圆内的数据被分为左子，和超椭圆之外的数据被分为右孩子。在一些情况下，如太少实例或共面的情况下，协方差矩阵可以是单数。在这种情况下，马哈拉诺比斯距离由欧几里德距离替换。

TL：喜：DM喜;小号dÞ磷;我¼1;2;。。。;·N·克ð5Þ

TR：喜：DM喜;小号dÞ[P;我¼1;2;。。。;·N·克ð6Þ

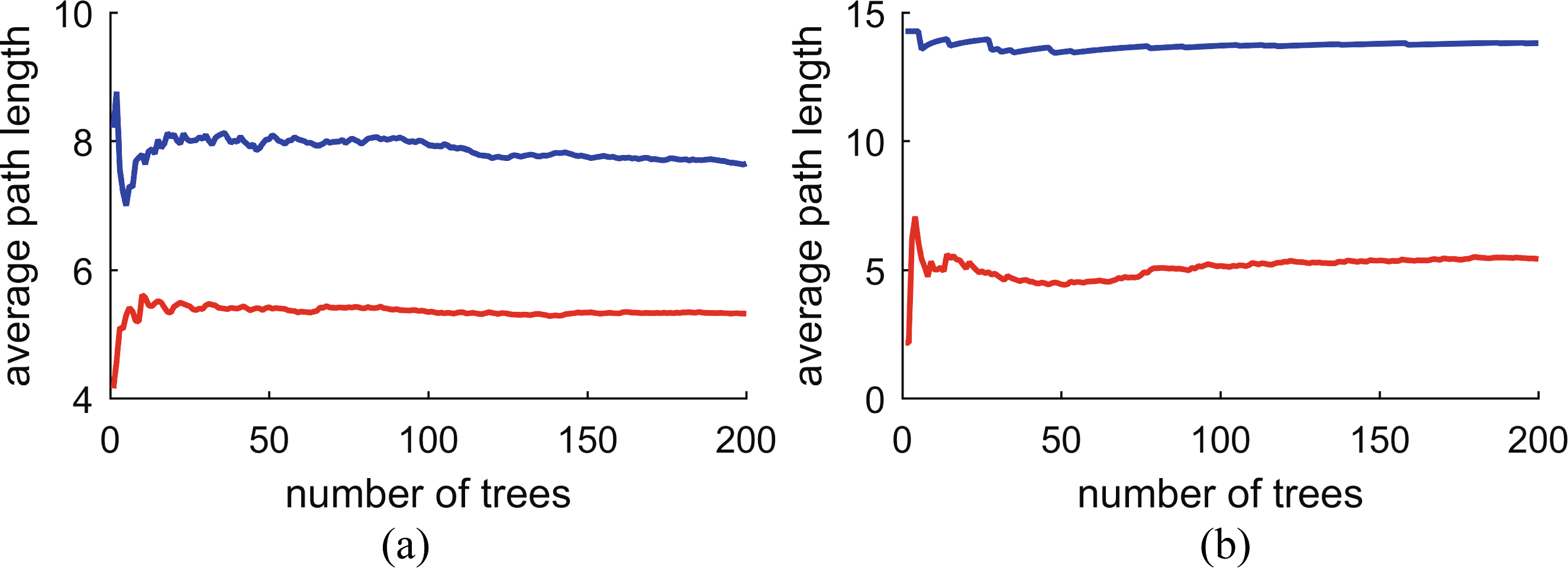
138C.Yao等人。

为了确定阈值p，一个参数v被引入。参数V的含义是分为左子中的所有数据的数据率。分为左子样品是地板为VN具有较小Mahalanobis距离dÞ样品。当k的值是1，只有一个属性是随机在每一个节点选择。此时，它是等效于确定一个属性的间隔，在时间间隔内的情况下被分成左子，并在区间外的情况下被分成右子。构建DTREE的细节在算法1和2DTREE所述的节点，直到以下停止条件之一被满足，根据上述方法递归地分割描述：

（1）只有两个实例保持在节点。在这一点上，两个实例具有相同的马氏距离，所以没有必要再划分节点。（2）在该节点的数据都在选择的属性是相同的。（3）DTREE的高度达到极限高度。像iTree，每个DTREE并不需要完全成长。主要的原因是，大多数异常可以由几个子空间被划分后到达叶节点。因此，树的高度可以被限制，以节省培训和测试时间。

dForest建立DTREE的集合。通过以上介绍，可以知道，dForest的建设需要确定三个参数，树木吨即数量，子采样尺寸w，并且比率诉株数吨控制集合大小。对于iForest，路径长度通常收敛在t¼100而由于dForest的节点划分是更有目的，它通常比收敛更快iForest。如图1中所示，平均路径长度趋于稳定作为树木的数量增加。

参数子采样大小W控制的训练样本，建立一个DTREE数量。在一般情况下，此参数选择一个较小的值可以达到相同的结果，在大多数数据集较大的值。W的合理范围是32至256通过实验评估。



森林分布：异常检测方法研究iForest139

算法1：dForest（X，T，，）输入：X-输入数据，叔-株数，-子采样大小，-分为左子输出的数据的比率：一套dTrees1.初始化森林2。为它=-1至吨3.X'←样品（X）4.森林←森林DTREE（X'，v）的5端为6.返回森林

算法2：输入：X“-输入数据，V-分为左子输出的数据的比率：一个DTREE1.如果X满足停止条件，然后返回2{exNode大小←|X”|}3.其他4.令Q是X的属性列表”5随机地选择该子空间6.随机选择ķattributes7的尺寸k。计算X的协方差矩阵Σ和平均值μ“的属性q的8上，如果秩（Σ）==K：9.10.计算每个实例和X之间的Mahalanobis距离d”11.其他12.13.计算的每个实例和X之间的欧几里德距离d“以升序16.14.端如果15.排序d让分割点是d的在17处的过滤器（X”的值，）18过滤器（X”，）19。inNode{20。21.SplitAtt←Q，22.SplitValue←，23.covMatrix←Σ，24.平均←μ，25.distType←标志}26。万一

比V是用来控制左子数据的分割后的比例。左子的数据是超椭圆内的情况下，和右子含有超椭圆外面的情况。参数v的值应大于0.5，以确保正常的点和所述异常的平均路径长度是判别。当与正常情况下和异常情况的训练，小V，则设置。因为分布估计可能不准确，更小的v是更健壮到异常。通过实验，发现一个合适的范围为

140℃Yao等人。

[0.7,1）。当与仅正常样本的训练，该分布的边缘都被视为其中异常可能出现的位置。由于分布估计是比较准确的，在这个时候，V值大效果更好。图1示出了位置，其中平均路径长度收敛具有不同参数v。

3.2测试阶段

所提出的方法采用计算路径长度和得分作为iForest的相同的方式。与iForest一致，异常会更容易达到比正常dForest实例叶节点。组合不同树木的结果后，将异常的平均路径长度通常比正常情况下要小。算法3描述了如何计算上的DTREE测试实例的路径长度。由于用于早停止的条件在训练过程中被设定，在许多叶节点的数据的大小大于1为了使结果可比，路径长度进行调整。该策略是添加C（T.size）到路径长度e。C的计算方法（T.size）在公式中，如图所示。（1）。获得平均路径长度后，将试验实例的得分可根据等式来计算。（3）。得分范围从0到1越接近得分是1时，更可能的实例是成为一个异常。越接近的比分是0，更可能的实例是正常的。

算法3：光程长度（X，T，HLIM，E）输入：X-的实例，T-一个DTREE，HLIM-高度限制，电子-电流路径长度;要被初始化为零，当第一次调用输出：X1，如果2.返回E+C（T.size）3.端的路径长度，如果4.5.如果

6.7.8.其他9.结束时，如果10分11返回的光程（X，T.left，HLIM，E+1）。12.其他13.返回路径长度（X，T.right，HLIM，E+1）14.结束时，如果

4实验

为了验证所提出的方法的性能，dForest等反映最新技术的方法在不同的数据集进行比较。首先，我们的方法是基于iForest，所以iForest被选择为比较方法。其次，使用隔离最近邻合奏（INNE）是另一种基于隔离异常检测

森林分布：异常检测方法研究iForest141

方法[15]。而INNE展示在许多数据集优异的性能。最后，本地异常因子是公知的异常检测方法。和LOF以及在检测局部异常执行，并且它被广泛地用作用于异常检测的比较方法。此外，dForest已在两种模式根据所选择的属性是否大于1在每个节点处执行的。不同的数据集选择，比较了这些方法的性能。其中，网络入侵数据的最大的数据集在文献[14]被使用。数据集航天飞机和Breastw从源[12]收集，剩余的数据集是由文献[13]所提供的异常检测基准数据集。这些数据集的详细信息可以在表1中找到。

由于正常情况下和异常不均衡，AUC被选择作为评价指标。每个数据集，随机进行10次每个方法。结果的平均值作为最终结果。此外，实验CON组在两个半监督和无监督模式涵道。

4.1半监督模式

当训练数据只包含正常情况下，如下数据集划分。正常情况下的60％的被用于训练，并且剩余的40％的正常实例和所有异常用于测试。其结果如表2所示为每个方法的关键参数的值示于表3。dForest，树木的数目t为100，并在所有的数据集的子采样尺寸w为128。比率V被搜索。对于iForest，树吨数为100上的所有数据集，并且子采样大小瓦特上不同的数据集被示出在表3中为INNE，所述集合大小为100，子采样尺寸中搜索在候选集合{2,4,8,16,32,64,128}。对于LOF，参数邻居数目k被优化。间的参数k的20个候选值中，对应于最佳结果的参数被选择作为最终值。

数据集的表1.基本信息

数的情况下的数据集属性数异常率的Http5674973枚0.39％班车490979枚7％Pageblocks5393枚109.46％Spambase420757枚39.91％胎心211421枚22％皮马7688枚34.90％Breastw6839枚34.99％电离层3513235.90％邮票34099.12％

142C.Yao等人。

正如可以从结果列于表2中可以看出，在dForest总共9点的数据集进行6个数据集最好。在对比iForest，dForest执行更好在5个数据集，并具有相同的结果iForest4点的数据集。这证明dForest的异常检测的有效性。此外，相同的参数T和W是在9点的数据集，这也证明了它更容易为dForest组参数使用。

4.2无监督模式

在无监督异常检测的情况下，训练数据包含正常情况和异常情况，并假设不提供自己的标签。类似地，每个数据集的60％被用于训练。剩余的40％用于测试。结果见表4。而且对于关键参数的设置如表5所示的结果表明，dForest也是行之有效的无监督模式。

表2不同的方法AUC性能（半监督模式）

AUCdForest随机ķdForestk=1时iForestINNELOF的Http1.001.001.001.001.00班车1.001.001.000.991.00Pageblocks0.940.930.910.900.94Spambase0.830.850.820.720.73胎心0.840.830.840.840.84皮马0.750.750.740.770.72Breastw0.990.990.990.990.990.97电离层0.950.920.950.950.94邮票0.940.930.960.94

表3.不同的方法参数（半监督模式）

dForest随机ķdForestk=1时iForestINNELOFvvwwķ的Http0.990.995122500返0.990.9925625Pageblocks0.990.99128830Spambase0.990.9912864500胎心0.70.81282150皮马0.80.91284200Breastw0.990.9912822000.99电离层1280.99420枚0.99邮票0.99128840

森林分布：异常检测方法研究iForest143

表2和表4的结果的比较可以导致这样的结论。一般来说，训练与正常情况下达到相同或更好的结果比正常和异常情况下的训练。这样做的原因的结果是分布的估计将是无异常的更准确。在半监督模式，则仅需要从局部distri-组织者那样远视为其中异常可能位于的位置的位置。因此，较大的V为选择。在无监督模式，普通实例和异常情况来区分在每个子空间。此外，异常情况下会干扰分布的估计，和一个较小的-v可以使划分更坚固。在时间复杂性方面，当k的值是1，dForest和iForest具有相同的时间复杂度。训练阶段的时间复杂度为OtWlogwdTH，以及测试阶段的时间复杂度为Ontlogwd位。在这一点上，iForest和dForest比其他方法更快。当k的值大于1，时间复杂性的增加，但时间消耗仍是一个合理的范围内。

表4.不同的方法AUC性能（监督模式）

AUCdForest随机ķdForestk=1时iForestINNELOF的Http1.001.001.001.000.42班车1.001.001.000.620.55Pageblocks0.920.900.890.900.91Spambase0.610.640.650.590.70胎心0.790.810.760.800.81皮马0.700.720.700.710.68Breastw0.980.980.980.980.960.91电离层0.870.850.890.910.94邮票0.960.920.970.94

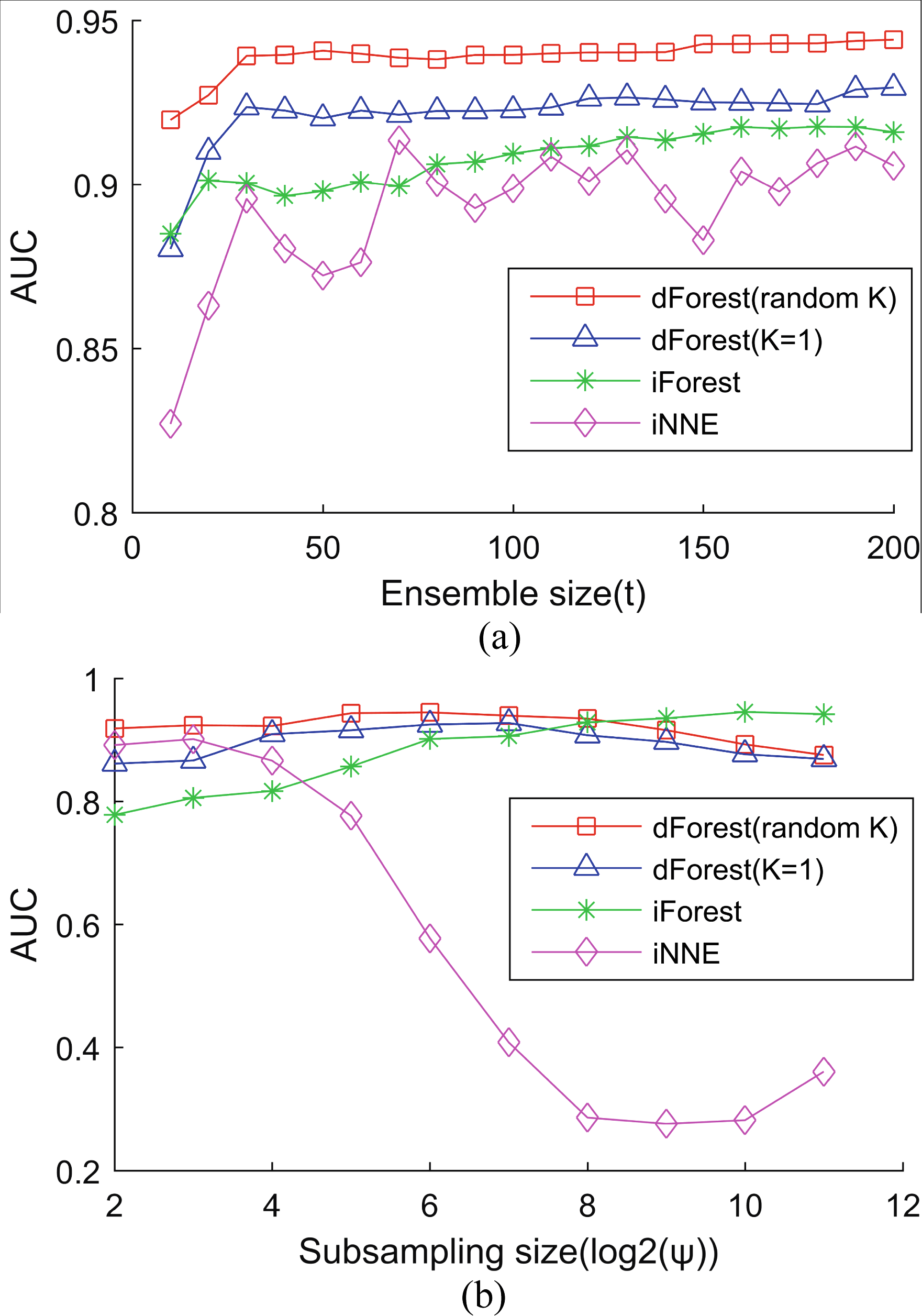
表5参数不同的方法（监督模式）

dForest随机ķdForestk=1时iForestINNELOFvvwwķ的Http0.990.9912825班车0.90.9128250Pageblocks0.90.912816100Spambase0.90.91282500胎心0.70.71282500皮马0.80.91284200Breastw0.70.912822000.8电离层1280.99820枚邮票0.80.81284150

144C.Yao等人。

4.3集合大小和子采样尺寸的影响

另外，在上述实验中，相同的值被用于在dForest参数集合大小和子采样大小设置。下面我们探讨这两个参数的影响。在iForest和INNE，这些参数还需要进行设置，因此它们被用作一个COM型坯。在Pageblocks数据集，这些方法进行了比较。合奏尺寸变化从10至200，和子采样尺寸范围从4到2048。其结果示于图2中。从结果中，当集成规模达到一定水平可以看出，dForest趋于稳定。而dForest也是稳健的采样大小的选择。小尺寸二次抽样可以实现比一个较大的值更好的结果。当二次抽样规模过大，dTrees之间的差异小，性能可能会下降。合理的范围是[32，256]。



森林分布：异常检测方法研究iForest145

5结论

在本文中，我们提出了一个新的异常检测方法，即dForest。该方法的主要内容如下。（1）dForest充分利用多维信息。dForest对待DTREE的每个节点作为子空间。随机选择的属性的不同组合构成不同的子空间。（2）dForest使用本地分配的异常和正常实例之间进行区分。在dForest，节点根据局部区域的分布划分。由马氏距离确定的超椭圆内的情况被分为左子，其余的情况下被分为左右的孩子。（3）dForest有一个直观的和明确的解释。数据被连续地投影到子空间不同，以及正常和异常实例在子空间连续区分。实验评价很好地说明了在这两个半监督和无监督模式dForest执行。相对于LOF，iForest和INNE，dForest实现对不同的数据集有竞争力的AUC。当选定的属性的数目是1，dForest具有线性时间复杂度和低存储器成本，同时实现较高的检测精度。因此，dForest可以应用于大规模的数据。

确认。这项工作部分是由天津市自然科学基金（No.18ZXZNGX00200）的支持，中国的国家重点研究发展计划（2016YFC0400709），天津滨海新区科学技术委员会（BHXQKJXM-PT-ZJSHJ-2017005），天津市自然科学基金（18YFYZCG00060），南开大学（91922299）。

1.Chandola，V.，班纳吉，A.，库马尔，V.:异常检测：调查。ACMCOMPUT。监测网。41（3），1-58（2009年）他，Z.，徐，X.，邓，S.:发现基于簇的局部异常值。模式Recogn。快报。24（9-10），1641至1650年（2003）3。BREUNIG，M.M.，克里格尔，H.P.，伍，R.T.:LOF：识别基于密度的局部异常值。在：对数据的管理ACMSIGMOD国际会议。ACM（2000年）4家乐，E.M.，吴，室温，Tucakov，V：基于距离的离群值：算法和应用。VLDBJ.8（3-4），237-253（2000）5，刘，F.T.，楷，M.T.，周，Z.H：隔离森林。在：数据挖掘第八届IEEE国际会议（2009年）6刘，F.T.，亭，K.M.，周，Z.H：隔离基于异常检测。ACM跨。Knowl。Discov。数据6（1），1-39（2012）7.Scholkopf，B.:估计高维分布的支持。神经COMPUT。13（7），1443年至1471年（2014）8，威廉姆斯，G.，巴克斯特，R.，他，H.，霍金斯，S.，谷，L。：RNN的用于数据挖掘离群点检测的比较研究。在：数据挖掘，2003年ICDMIEEE（2002年）9马氏，P.C：在统计的广义距离2002年IEEE国际会议论文集。PROC。国家科。研究所。科学。印度12，49-55（1936）

146C.Yao等人。

10.Maesschalck，R.D.，儒安-兰波，D.，Massart的，D.L。：马哈拉诺比斯距离。Chemometr。INTELL。实验室。SYST。50（1），1-18（2000）11.帕蒂尔，N.，DAS，D.，Pecht中，M.:异常检测用于使用Mahalanobis距离的IGBT。Microelectron。Reliab。55（7），一〇五四年至1059年（2015）12度瓦，D.，KarraTaniskidou，E。：UCI机器学习知识库。信息和加州大学计算机科学学院，尔湾，加利福尼亚州（2017年）的学校。HTTP：//archive.ics.uci。EDU/ml13。Yamanishi，K.，竹内，J.I.，威廉姆斯，G.，Milne的，P。：在线使用有限的混合物与贴现学习算法的无监督异常检测。数据最小。Knowl。Discov。8（3），275-300（2004）14.Swersky，L.，马克斯，H.O.，砂光机，J.，Campello的，R.J.G.B.，Zimek，A.:在孤立点检测和一类分类方法的评估。在：对数据科学与高级分析IEEE国际会议。IEEE（2016）15Bandaragoda，T.R.，亭，K.M.，阿尔布雷希特，D.，刘，F.T.，朱，Y.，韦尔斯，J.R.:隔离基于使用近邻合奏异常检测。COMPUT。INTELL。34，968-998（2018）

森林分布：异常检测方法研究iForest147