描述了一种从不平衡数据集构建分类器的方法。如果分类类别的代表性不近似相等，则数据集是不平衡的。通常，真实世界的数据集主要由“正常”示例组成，只有一小部分“异常”或“有趣”示例。还有一种情况是，将异常（感兴趣）示例错误分类为正常示例的成本通常远高于反向错误的成本。大多数（正常）类别的欠采样被认为是提高类别对少数类别敏感性的一种好方法。本文表明，我们对少数（异常）类进行过采样和对多数（正常）类进行欠采样的方法相结合，可以实现比仅对多数类进行欠抽样更好的分类性能（在ROC空间中）。本文还表明，与改变Ripper中的损失率或Naive Bayes中的类先验相比，我们对少数类进行过采样和对多数类进行欠采样的方法相结合，可以实现更好的分类器性能（在ROC空间中）。我们对少数群体类进行过采样的方法包括创建合成的少数群体类示例。实验使用C4.5、开膛手和Naive Bayes分类器进行。该方法使用受试者工作特性曲线下面积（AUC）和ROC凸包策略进行评估。

1已在其他申请中报道（Provost&Fawcett，2001）。有人试图处理欺诈电话（Fawcett和Provost，1996）、电信管理（Ezawa、Singh和Norton，1996）和文本分类（Lewis和Catlett，1994；Dumais、Platt、Heckerman和Sahami，1998；Mladini´c和Grobelnik，1999；Lewis和Ringuette，1994；Cohen，1995a）以及卫星图像中的石油泄漏检测（Kubat、Holte和Matwin，1998）等领域的不平衡数据集。

机器学习算法的性能通常使用预测准确性来评估。然而，当数据不平衡和/或不同错误的成本差异显著时，这是不合适的。例如，将主图像中的像素分类视为可能的癌症（Woods，Doss，Bowyer，Solka，Priebe，&Kegelmeyer，1993）。典型的乳房X光摄影数据集可能包含98%的正常像素和2%的异常像素。一个简单的默认策略，即猜测多数类，将给出98%的预测准确率。然而，应用程序的性质要求在少数类中有相当高的正确检测率，并允许在多数类中有较小的错误率，以实现这一点。在这种情况下，简单的预测准确性显然是不合适的。受试者工作特性（ROC）曲线是一种标准技术，用于总结在真阳性和假阳性错误率之间的一系列权衡中的更出色性能（Swets，1988）。曲线下面积（AUC）是公认的ROC曲线的传统性能指标（Duda，Hart，&Stork，2001；Bradley，1997；Lee，2000）。ROC凸包也可以用作识别潜在最优分类器的稳健方法（Provost&Fawcett，2001）。如果一条线穿过凸包上的一个点，那么没有其他具有相同斜率的线穿过另一个具有更大真正（TP）截距的点。因此，在任何与该斜率相关的分布假设下，该点的分类器都是最优的。

机器学习社区通过两种方式解决了阶级失衡的问题。一种是将不同的成本分配给训练实例（Pazzani，Merz，Murphy，Ali，Hume，&Brunk，1994；多明戈斯，1999年）。另一种是通过对少数类进行过采样和/或对多数类进行欠采样来对原始数据集进行重新采样（Kubat&Matwin，1997；Japkowicz，2000；Lewis和Catlett，1994；Ling&Li，1998）。我们的方法（Chawla，Bowyer，Hall，&Kegelmeyer，2000）将多数类的欠采样与少数类的过采样的特殊形式相结合。使用各种数据集和C4.5决策树分类器（Quinlan，1992）、Ripper（Cohen，1995b）和Naive Bayes分类器进行的实验表明，与之前使用AUC或ROC凸包的其他重新采样、修改损失率和类先验方法相比，我们的方法有所改进。

第2节概述了绩效指标。第3节回顾了处理不平衡数据集的最密切相关的工作。第4节介绍了我们的方法的细节。第5节介绍了将我们的方法与其他重新采样方法进行比较的实验结果。第6节讨论了结果，并提出了未来工作的方向。

正确分类（真阴性），F P是错误分类为阳性（假阳性）的阴性示例的数量，F N是错误归类为阴性（假阴性）的阳性示例的数量；T P是正确分类的阳性示例（真阳性）的数量。

预测精度是通常与机器学习算法相关的性能指标，定义为精度=（T P+T N）/（T P+F P+T N+F N）。在数据集平衡和错误成本相等的情况下，使用错误率作为性能指标是合理的。错误率为1−准确度。在存在误差成本不等的不平衡数据集的情况下，更适合使用ROC曲线或其他类似技术（Ling&Li，1998；Drummond&Holte，2000；Provost&Fawcett，2001；Bradley，1997；Turney，1996）。

ROC曲线可以被认为代表TP和FP的相对成本的最佳决策边界族。在ROC曲线上，X轴表示%F P=F P/（T N+F P），Y轴表示%T P=T P/（T P+F N）。ROC曲线上的理想点是（0100），即所有阳性示例都被正确分类，没有阴性示例被错误分类为阳性。可以扫出ROC曲线的一种方法是通过操纵训练集中每个类的训练样本的平衡。图2显示了一个示例。线y=x表示随机猜测类的场景。ROC曲线下面积（AUC）是衡量更优秀表现的有用指标，因为它独立于所选的决策标准和先验概率。AUC比较可以建立类别之间的优势关系。如果ROC曲线相交，则总AUC是模型之间的平均比较（Lee，2000）。然而，对于某些特定的成本和类别分布，具有最大AUC的类别实际上可能是次优的。因此，我们还计算ROC凸包，因为位于ROC凸壳上的点可能是最优的（Provost，Fawcett，&Kohavi，1998；Provost和Fawcett（2001）。

相关工作提出了SHRINK系统，该系统将次要（阳性）和大多数（阴性）类别的重叠区域划分为阳性；它搜索“最佳正区域”（Kubat et al.，1998）。

Japkowicz（2000）讨论了数据集中不平衡的影响。她评估了三种策略：欠采样、重采样和基于识别的归纳方案。我们专注于她的采样方法。她对人工1D数据进行了实验，以便轻松测量和构建概念复杂性。考虑了两种重新采样方法。随机重采样包括随机对较小的类进行重采样，直到它由与多数类一样多的样本组成，而“集中重采样”包括仅对少数类和多数类之间边界上发生的少数示例进行重采样。考虑了随机欠采样，即随机对多数类样本进行欠采样，直到它们的数量与少数类样本的数量相匹配；重点欠采样涉及对距离更远的大多数类别样本进行欠采样。她指出，这两种采样方法都是有效的，她还观察到，使用复杂的采样技术在所考虑的领域没有任何明显的优势（Japkowicz，2000）。

市场分析问题（凌莉，1998）。在一个实验中，他们对大多数类别进行了低采样，并注意到当类别相等时，会获得最佳升力指数。在另一个实验，他们对正代表的样本进行了高采样（Ling&Li，1998）。（少数）例子与替换，以匹配负面（多数）例子的数量与正面例子的数量。过采样和欠采样的组合并没有显著改善升力指数。然而，我们的过采样方法与他们的不同。

Solberg和Solberg（1996）考虑了SAR图像中浮油分类中数据集不平衡的问题。他们使用过采样和欠采样技术来改进油片的分类。他们的训练数据分布有42个切片和2471个外观相似的切片，给出了外观相似的先验概率0.98。这种不平衡将导致学习者（在没有任何适当的损失函数或修改先验的方法的情况下）以对许多浮油样本进行错误分类为代价，对几乎所有看起来相似的样本进行正确分类（Solberg&Solberg，1996）。为了克服这种不平衡问题，他们从浮油中过度采样（替换）了100个样本，并从非浮油类中随机采样了100个样品，以创建一个具有相同概率的新数据集。他们在这个平衡的数据集上学习了一个分类树，并在误差估计的留一法中，在浮油上实现了14%的误差率；在外观相似的情况下，他们实现了4%的错误率（Solberg&Solberg，1996）。

另一种与我们的工作类似的方法是多明戈斯（1999）的方法。他将“元主”方法与多数欠采样和少数过采样进行了比较。他发现元成本比任何一种都有所提高，并且欠采样比平均过采样更可取。基于错误的分类器对成本敏感。对每个例子的每个类别的概率进行了估计，并根据分类错误成本对例子进行了最佳重新标记。示例的重新标记扩展了决策空间，因为它创建了新的样本，分类器可以从中学习（Domingos，1999）。

在不平衡数据集上训练的前馈神经网络可能无法在类之间进行足够的区分（DeRouin，Brown，Fausett，&Schneider，1991）。作者提出，神经网络的学习率应适应数据中类表示的统计。他们根据提供给神经网络进行训练的样本比例计算出一个注意力因子。网络元素的学习率根据注意力因素进行了调整。他们在人工生成的训练集和真实世界的训练集上进行了实验，这两个训练集都有多个（两个以上）课程。他们将此与复制少数类样本以平衡用于训练的数据集的方法进行了比较。少数民族类别的分类准确率得到了提高。

Lewis和Catlett（1994）研究了监督学习的异质不确定性抽样。这种方法对于训练具有不确定类的样本非常有用。训练样本分两个阶段递增标记，不确定实例传递到下一阶段。他们对C4.5进行了修改，以包括用于确定叶片类别值的损失率。通过与LR/（LR+1）的概率阈值进行比较来确定类值，其中LR是损失率（Lewis和Catlett，1994）。

也就是说，构建反映页面中单词出现的特征向量。通常，在文本分类中，有趣类别的实例很少。在信息检索问题中，这种对负类的过度表示可能会导致评估分类器性能的问题。由于错误率不是偏斜数据集的良好指标，因此信息检索中算法的分类性能通常通过精度和召回率来衡量：

Mladini´c和Grobelnik（1999）提出了一种特征子集选择方法来处理IR域中的不平衡类分布。他们试验了各种特征选择方法，发现当与Naive Bayes分类器相结合时，优势比（van Rijsbergen，Harper，&Porter，1981）在他们的领域中表现最好。比值比是一种概率度量，用于根据文档与正类（少数类）的相关性对文档进行排序。另一方面，一个词的信息增益不关注特定的目标类别；它是为每个类按单词计算的。在不平衡的文本数据集中（假设98%到99%是负类），大多数特征将与负类相关联。与文本分类的信息增益相比，比值比在其度量中包含了目标类信息，从而提供了更好的结果。Provost和Fawcett（1997）引入了ROC凸包方法来估计不平衡数据集的分类性能。他们指出，阶级分配不平等和错误成本不平等的问题是相关的，并且几乎没有做什么工作来解决这两个问题（Provost&Fawcett，2001）。在ROC凸包方法中，ROC空间用于将分类性能与类别和成本分布信息分离。

总结文献，对多数类的抽样不足比对少数类的抽样过多能够建立更好的分类器。在之前的工作中，两者的结合并不会导致分类器优于仅使用欠采样构建的分类器。然而，少数类别的过度抽样是通过替换原始数据进行抽样来完成的。我们的方法使用了不同的过采样方法。

先前的研究（Ling&Li，1998；Japkowicz，2000）已经讨论了过度抽样和替换，并指出这并不能显著提高少数族裔的识别率。我们根据特征空间中的决策区域来解释潜在的影响。从本质上讲，随着少数类被过采样的数量不断增加，其效果是将特征空间中相似但更具体的区域识别为少数类的决策区域。这种对决策树的影响可以从图3中的图中理解。

图3:a）构建决策树后，三个少数类样本（用“+”表示）所在的决策区域。该判定区域由实线矩形表示。b） 同一数据集的所选少数类样本的放大视图。小的实线矩形显示了决策区域，这是对具有复制的少数类进行过采样的结果。c） 同一数据集的所选少数类样本的放大视图。虚线显示了使用合成生成对少数类进行过采样后的决策区域。

图3中的绘图数据取自乳腺造影数据集1（Woods等人，1993）。图中少数类样本用+表示，多数类样本用o表示。在图3（a）中，实线矩形表示的区域是多数类决策区域。然而，它包含三个少数类样本，用“+”表示为假阴性。如果我们复制少数类，少数类的决策区域将变得非常特殊，并将在决策树中引起新的分裂。这将导致更多的终端节点（离开），因为学习算法试图学习少数类的越来越多的特定区域；本质上，过度适应。少数阶级的复制不会导致其决策边界扩散到多数阶级区域。因此，在图3（b）中，先前在多数类决策区域中的三个样本现在具有非常特殊的决策区域。

我们提出了一种过采样方法，通过创建“合成”示例对少数类进行过采样，而不是通过替换进行过采样。这种方法的灵感来源于一种在手写字符识别中被证明是成功的技术（Ha&Bunke，1997）。他们通过对真实数据执行某些操作来创建额外的训练数据。在他们的情况下，像旋转和偏斜这样的操作是干扰训练数据的自然方式。我们通过在“特征空间”而不是“数据空间”中操作，以不太特定于应用程序的方式生成合成示例。通过取每个少数类样本并沿着连接任意/所有k个少数类最近邻居的线段引入合成示例，对少数类进行过采样。根据所需的过采样量，从k个最近邻居中随机选择邻居。我们的实施目前使用了五个最近的邻居。例如，如果所需的过采样量为200%，则只选择五个最近邻域中的两个邻域，并在每个邻域的方向上生成一个样本。合成样本的生成方式如下：取所考虑的特征向量（样本）与其最近邻居之间的差异。将该差异乘以0到1之间的随机数，并将其添加到所考虑的特征向量中。这导致沿着两个特定特征之间的线段选择一个随机点。这种方法有效地迫使少数群体的决策区域变得更加普遍。下一页的算法SMOTE是SMOTE的伪代码。表4.2显示了随机合成样本的计算示例。过采样量是系统的一个参数，可以为不同的群体生成一系列ROC曲线，并进行ROC分析。

在少数类中，用于10倍交叉验证的训练集。少数民族阶层的抽样比例分别为其原始规模的100%、200%、300%、400%和500%。图表显示，在较高复制度下使用替换的少数过采样的树大小远大于SMOTE的树大小，并且在较高复制程度下使用替换技术对少数过采样进行的少数类识别不如SMOTE。

考虑一个样本（6,4），并设（4,3）为其最近邻居。（6,4）是识别k个最近邻居的样本。（4,3）是它的k近邻之一。设：f1 1=6 f2 1=4 f2 1-f1 1=-2 f1 2=4 f2 2=3 f2 2-f1 2=-1新样本将生成为（f1’，f2’）=（6,4）+rand（0-1）\*（-2，-1）rand（0-2）生成0到1之间的随机数。

通过从多数阶级人群中随机移除样本，直到少数阶级成为多数阶级的特定百分比，从而对多数阶级进行抽样不足。这迫使学习者经历不同程度的欠采样，而在较高程度的欠抽样情况下，少数族裔在训练集中的比例更大。在描述我们的实验时，我们的术语是这样的：如果我们以200%的比例对多数类进行抽样不足，这意味着修改后的数据集将包含少数类元素的数量是多数类元素的两倍；也就是说，如果少数类有50个样本，而多数类有200个样本，并且我们低于200%的样本多数，那么多数类最终将有25个样本。通过将欠抽样和过抽样相结合，学习者最初对消极（多数）类的偏见被逆转为有利于积极（少数）类。分类器是在受“吸烟”少数类和欠采样多数类干扰的数据集上学习的。

ROC生成用于比较使用C4.5的SMOTE和欠采样，以及使用C4.5和Naive bayes的SMOTE。使用AUC和ROC凸包评估性能。

2.开膛手：我们使用开膛手（Cohen，1995b）作为基本分类，将SMOTE和欠采样的各种组合与普通欠采样进行了比较。我们还将开膛手损失率（Cohen&Singer，1996；Lewis&Catlett，1994）从0.9变为0.001（作为一种改变误分类成本的方法），并将这种变化的影响与SMOTE和欠采样的组合进行了比较。通过将损失率从0.9降低到0.001，我们能够为少数族裔建立一套规则。

3.朴素贝叶斯分类器：通过改变少数类的先验，可以使朴素贝叶斯分类器2对成本敏感。我们将少数类的先验从多数类的1倍改变到50倍，并与C4.5的SMOTE和欠采样组合进行比较。

这些不同的学习算法允许将SMOTE与一些可以直接处理分类错误成本的方法进行比较。%FP和%TP在每种数据组合的10倍交叉验证运行中取平均值。通过计算五个最近邻居并生成合成示例，对少数类示例进行过采样。使用梯形规则计算AUC。我们为每个ROC曲线外推了一个额外的点TP=100%和FP=100%。我们还计算了ROC凸包，以确定最佳分类器，因为位于外壳上的点可能是最佳分类器（Provost&Fawcett，2001）。

我们在九个不同的数据集上进行了实验。表5.2总结了这些数据集。这些数据集的大小和类别比例差异很大，因此为SMOTE提供了不同的领域。按照增加不平衡的顺序，它们是：

1.Pima Indian Diabetes（Blake&Merz，1998）有2类768个样本。该数据用于确定亚利桑那州凤凰城附近人群中的糖尿病阳性病例。阳性样本的数量只有268个。对糖尿病病例的检测具有良好的敏感性将是该类别的一个理想特征。

2.Phoneme数据集来自ELENA项目3。数据集的目的是区分鼻音（0类）和口音（1类）。共有5个功能。类别分布为类别0中的3818个样本和类别1中的1586个样本。

3.成人数据集（Blake&Merz，1998）有48842个样本，其中11687个样本属于少数群体。该数据集具有6个连续特征和8个标称特征。在此数据集上评估了SMOTE和SMOTE-NC（见第6.1节）算法。对于SMOTE，我们提取了连续特征，并生成了一个仅具有连续特征的新数据集。

4.E状态数据4（Hall，Mohney，&Kier，1991）由来自国家癌症研究所的酵母抗癌症药物筛选的一系列化合物的电拓扑状态描述符组成。NCI酵母抗癌药物筛选的E状态描述符由Tripos，股份有限公司生成。Brie fley，一系列约60000种化合物在给定浓度下针对一系列6种酵母菌株进行了测试。该试验是仅在一种浓度下进行高通量筛选，因此结果会受到污染等。当暴露于给定化合物时，测量酵母菌株的生长抑制作用（相对于酵母在中性溶剂中的生长）。活性类别要么是活性的——至少有一种单一酵母菌株被抑制超过70%，要么是非活性的——没有一种酵母菌株被禁止超过70%。该数据集有53220个样本和6351个活性化合物样本。

5.Satimage数据集（Blake&Merz，1998）最初有6个类。我们选择最小的班级作为少数族裔班级，并将其他班级分解为一个班级，如中所述（Provost等人，1998）。这给了我们一个倾斜的2类数据集，有5809个多数类样本和626个少数类样本。

样品。尽管如此，SMOTE技术也可以通过指定SMOTE用于哪个类来应用于多类问题。然而，在本文中，我们专注于2-类问题，以明确地表示正类和负类。

8.乳腺造影数据集（Woods等人，1993）有11183个样本和260个钙化点。如果我们将预测准确度视为这种情况下分类器优度的衡量标准，当每个样本都被标记为非钙化时，默认准确度将为97.68%。但是，分类器最好能正确预测大多数钙化。

9.Can数据集是使用AVTAR（Chawla&Hall，1999）版本的Mustafa可视化工具5从Can ExodusII数据生成的。罐子被压碎的部分被标记为“非常有趣”，罐子的其余部分被标记成“未知”。生成了一个443872个样本的数据集，其中8360个样本被标记为”非常有趣“。

SMOTE的ROC曲线是通过使用C4.5或Ripper为一系列修改后的训练数据集中的每一个创建一个分类器来生成的。给定的ROC曲线是通过首先对少数类进行特定程度的过采样，然后对多数类进行递增程度的欠采样来生成曲线上的连续点。欠采样量与普通欠采样量相同。因此，数据集的每个ROC曲线上的每个对应点代表相同数量的多数类样本。不同的ROC曲线是从不同水平的少数群体过度采样开始产生的。ROC曲线也是通过将Ripper中的损失率从0.9变为0.001，并将Naive Bayes分类器的少数类先验从原始分布变为多数类的50倍来生成的。

图7：电话号码。SMOTE-C4.5、Under-C4.5和朴素贝叶斯的比较。在ROC空间中，SMOTE-C4.5比Naive Bayes和Under-C4.5占主导地位。SMOTE-C4.5分类器是潜在的最佳分类器。

图9至图23显示了使用三个分类器获得的九个数据集的实验ROC曲线。将多数类平原欠采样的ROC曲线（Ling&Li，1998；Japkowicz，2000；Kubat&Matwin，1997；Provost&Fawcett，2001）与我们将合成少数类过采样（SMOTE）与多数类欠采样相结合的方法进行了比较。平面欠采样曲线标记为“under”，SMOTE和欠采样组合ROC曲线标记为为“SMOTE”。根据数据集的大小和相对不平衡，创建了一条到五条SMOTE和欠采样曲线。我们只在图中显示了SMOTE与欠采样和普通欠采样曲线相结合的最佳结果。C4.5中的SMOTE ROC曲线也与使用朴素贝叶斯分类器（标记为“朴素贝叶斯”）改变少数类先验获得的ROC曲线进行了比较。还比较了使用开膛手生成的“SMOTE”、“Under”和“Loss Ratio”ROC曲线。对于给定的ROC曲线族，生成ROC凸包（Provost&Fawcett，2001）。ROC凸包是使用Graham算法生成的（O’Rourke，1998）。作为参考，我们在图19中显示了使用少数过采样通过复制获得的ROC曲线。

图8：电话号码。SMOTE裂土器、欠裂土器和修改裂土器损失率的比较。在ROC空间中，SMOTE Ripper占主导地位，而非Under Rippers和Loss Ratio。更多的SMOTE开膛手级位于ROC凸面船体上。

图10：Pima印第安人糖尿病。SMOTE裂土器、欠裂土器和修改裂土器损失率的比较。在ROC空间中，SMOTE Ripper占主导地位，而非Under Rippers和Loss Ratio。

抽样或少数类过度抽样。少数族裔的抽样比例分别为50%、100%、200%、300%、400%、500%。大多数类别的样本不足，分别为10%、15%、25%、50%、75%、100%、125%、150%、175%、200%、300%、400%、500%、600%、700%、800%、1000%和2000%。多数类欠采样和少数类过采样的数量取决于数据集的大小和类的比例。例如，考虑图17中乳房X光摄影数据集的ROC曲线。有三条曲线——一条是普通多数类欠采样曲线，其中欠采样范围以不同的间隔在5%和2000%之间变化，一条是SMOTE和多数类欠抽样的组合曲线，一条为朴素贝叶斯曲线——还有一条ROC凸包曲线。图17所示的ROC曲线适用于少数族裔，采样率为400%。SMOTE-ROC曲线上的每个点都代表（合成）过采样和欠采样的组合，欠采样的量与普通欠采样的范围相同。为了更好地理解ROC图，我们在附录a中显示了其中一个数据集的不同ROC曲线集。

对于Can数据集，由于数据集的结构性质，我们不得不在比其他数据集更小的程度上进行SMOTE。对于Can数据集，在网格几何中已经建立了一个结构邻域，因此SMOTE可以导致创建位于表面下的邻域（因此不感兴趣），因为我们关注的是物理变量的特征空间，而不是结构信息。

图11：讽刺。SMOTE-C4.5、Under-C4.5和朴素贝叶斯的比较。Naive Bayes和SMOTE-C4.5的ROC曲线存在重叠；然而，在较高的TP处，来自SMOTE-C4.5的更多点位于ROC凸包上。

图12：讽刺。SMOTE裂土器、欠裂土器和修改裂土器损失率的比较。烟雾开膛手主宰着ROC空间。ROC凸包主要由SMOTE Ripper的点构成。

图13：森林覆盖率。SMOTE-C4.5、Under-C4.5和朴素贝叶斯的比较。SMOTE-C4.5和Under-C4.5 ROC曲线彼此非常接近。然而，来自SMOTE-C4.5 ROC曲线的更多点位于ROC凸包上，从而建立了优势。

ROC曲线下面积（AUC）是使用梯形规则的形式计算的。给定ROC曲线的最左下角是分类器在原始数据上的表现。最右上角的点始终为（100%，100%）。如果曲线没有自然终止于该点，则会添加该点。这对于在相同的%FP范围内比较AUC是必要的。

图14：森林覆盖率。SMOTE裂土器、欠裂土器和修改裂土器损失率的比较。SMOTE Ripper展现了在中华民国空间的统治地位。SMOTE Ripper曲线的更多点位于ROC凸包上。

图15：油。SMOTE-C4.5、Under-C4.5和朴素贝叶斯的比较。尽管SMOTE-C4.5和Under-C4.5 ROC曲线在点处相交，但来自SMOTE-C4.5%曲线的更多点位于ROC凸包上。

图16：油。SMOTE裂土器、欠裂土器和修改裂土器损失率的比较。“下裂土器”和“SMOTE Ripper”曲线相交，“下裂土器”曲线中的更多点位于ROC凸包上。

图17：乳腺造影。SMOTE-C4.5、Under-C4.5和朴素贝叶斯的比较。SMOTE-C4.5和Under-C4.5曲线在ROC空间中相交；然而，由于ROC凸包上的点的数量，SMOTE-C4.5具有更多潜在的最优分类器。

图19：通过SMOTE对少数类示例进行过采样和通过Mammography数据集的复制对少数类实例进行过采样的比较。

（a） SMOTE-C4.5、Under-C4.5和朴素贝叶斯的比较。SMOTE-C4.5和Under-C4.5曲线在ROC空间中相交；然而，根据ROC凸包上的点数量，SMOTE-C4.5具有更具潜力的最优分类器。

图21:E状态。SMOTE裂土器、欠裂土器和修改裂土器损失率的比较。根据ROC凸包上的点数，SMOTE Ripper有更多潜在的最佳分类器。

图23：可以。SMOTE裂土器、欠裂土器和修改裂土器损失率的比较。SMOTE Ripper和Under Ripper ROC曲线在大部分ROC空间重叠。

Provost（2000）建议，应始终将简单地改变决策阈值视为更复杂方法的替代方案。在C4.5的情况下，这意味着改变决策树叶子处的决策阈值。例如，一个叶子可以将示例分类为少数类，即使该叶子上超过50%的训练示例代表多数类。我们通过将C4.5决策树学习器的叶子上的决策阈值设置为0.5、0.45、0.42、0.4、0.35、0.32、0.3、0.27、0.25、0.22、0.2、0.17、0.15、0.12、0.1、0.05、0.0来进行实验。我们在Phoneme数据集上进行了实验。图24显示了通过调整对少数族裔的偏见，将SMOTE和欠采样组合与C4.5学习进行比较。该图显示，SMOTE和欠采样组合ROC曲线在整个值范围内占主导地位。

然而，由于SHRINK方法和SMOTE方法的数据点不同，因此无法直接进行比较。与片面选择相比，SMOTE没有明显的改进。

在这一研究领域，有几个主题需要进一步考虑。自动自适应选择最近邻居的数量将是有价值的。创建合成邻居的不同策略可能能够提高性能。此外，选择最近的邻居，重点关注分类错误的例子，可以提高性能。少数类样本可能具有多数类样本作为其最近邻居，而不是少数类样本。这种拥挤可能有助于重新绘制有利于少数群体的决策面。除了这些主题之外，以下小节还讨论了SMOTE的两种可能扩展，以及SMOTE在信息检索中的应用。

虽然我们的SMOTE方法目前不处理具有所有标称特征的数据集，但它被推广为处理连续特征和标称特征的混合数据集。我们将这种方法称为合成少数过采样技术标称连续[SMTE-NC]。我们在UCI存储库的成人数据集上测试了这种方法。下面描述SMOTE-NC算法。

1.中值计算：计算少数类的所有连续特征的标准差的中值。如果样本与其潜在的最近邻居之间的标称特征存在差异，则该中值包含在欧几里得距离计算中。我们使用中值来惩罚标称特征的差异，惩罚量与连续特征值的典型差异有关。

2.最近邻计算：使用连续特征空间计算识别k个最近邻的特征向量（少数类样本）与其他特征向量（少数类样本）之间的欧几里得距离。对于所考虑的特征向量与其潜在最近邻居之间的每个不同标称特征，在欧几里得距离计算中包括先前计算的标准偏差的中值。表2展示了一个例子。

F1=1 2 3 A B C[设这是我们计算最近邻居的样本]F2=4 6 5 A D E F3=3 5 6 A B K因此，F2和F1之间的欧几里得距离为：Eucl=sqrt[（4-1）2+（6-2）2+（5-3）2+Med2+Med2]Med是幂级数的连续特征的标准偏差的中值。特征编号5:B的中值项包含两次→D和6:C→E、 其区别于两个特征向量：F1和F2。

3.填充合成样本：使用与前面描述的SMOTE相同的方法创建新的合成少数类样本的连续特征。标称特征被赋予出现在大多数k近邻中的值。

这里报道的SMOTE-NC实验与SMOTE的实验设置相同，只是我们只检查一个数据集。成人数据集的SMOTE-NC与我们的典型结果不同：如图26和27所示，它的表现比基于AUC的普通欠采样差。我们只提取了连续特征，以分离SMOTE和SMOTE-NC对该数据集的影响，并确定这种奇怪现象是否是由于我们对标称特征的处理。如图28所示，即使只有连续特征应用于成人数据集的SMOTE，也无法实现比普通欠采样更好的性能。一些少数类连续特征具有非常高的方差，因此，少数类样本的合成生成可能与多数类空间重叠，从而导致比普通欠采样更多的误报。当我们在大于50%的度数吸烟时，AUC测量值的降低也支持了这一假设。SMOTE的程度越高，数据集中的少数类样本越多，因此与多数类决策空间的重叠越大。

图27：成人。SMOTE裂土器、欠裂土器的比较，以及修改裂土器中的损失Ra-tio。SMOTE Ripper和Under Ripper ROC曲线在大部分ROC空间重叠。

在上述等式中，V1和V2是两个相应的特征值。C1是特征值V1出现的总次数，C1i是类别i的特征值V1的出现次数。类似的约定也可以应用于C2i和C2。k是一个常数，通常设置为1。该方程用于计算给定特征向量集中每个标称特征的值差矩阵。方程1给出了一组固定的有限值上的几何距离（Cost&Salzberg，1993）。Cost和Salzberg修改的VDM省略了Stan fill和Waltz在δ计算中包含的权重项，这有助于使δ对称。两个特征向量之间的距离∆由以下公式给出：

r＝1产生曼哈顿距离，r＝2产生欧几里得距离（Cost&Salzberg，1993）。wx和wy是修改后的VDM中的示例权重。对于新示例（特征向量），wy＝1，并且wx是对更可靠示例（特征矢量）的偏差，并且被计算为特征矢量的使用次数与特征矢量的正确使用次数的比率；因此，更精确的特征向量将具有wx≈1。对于SMOTE-N，我们可以忽略方程2中的这些权重，因为SMOTE-N不直接用于分类目的。然而，我们可以重新定义这些权重，使少数类特征向量的权重更接近多数类特征向量；因此，使这些少数类特征看起来离所考虑的特征向量更远。由于我们更感兴趣的是形成少数类的更广泛但准确的区域，因此可以使用权重来避免沿着更接近多数类的邻居进行填充。为了生成新的少数类特征向量，我们可以通过考虑特征向量的多数票及其k个最近邻居来创建新的集合特征值。表6.2显示了创建合成特征向量的示例。

我们正在研究SMOTE在信息检索（IR）中的应用。IR问题具有过多的功能和潜在的许多类别。在将给定文档或网页转换为单词袋格式后，SMOTE必须与特征选择算法结合使用。

与SMOTE的一个有趣的比较是Naive Bayes和Odds比率的组合。比值比关注目标类，并根据文档与目标类或正类的相关性对文档进行排名。SMOTE还通过创建目标类的更多示例来关注该类。

结果表明，SMOTE方法可以提高少数类分类器的准确性。SMOTE为过采样提供了一种新的方法。SMOTE和欠采样的组合比普通欠采样性能更好。SMOTE在各种数据集上进行了测试，训练集中存在不同程度的不平衡和不同数量的数据，从而提供了一个多样化的测试平台。基于ROC空间中的支配地位，SMOTE和欠采样的组合也比Ripper中的不同损失率或Naive Bayes分类器中的不同类先验表现更好：这些方法可以直接处理偏斜的类分布。SMOTE强制进行重点学习，并引入了对少数族裔阶层的偏见。只有在Pima（偏差最小的数据集）中，Naive Bayes分类器的性能才比SMOTE-C4.5好。此外，只有在Oil（机油）数据集中，Under Ripper（裂土器下方）的性能才比SMOTE Rippler（烟雾裂土器）好。对于Can数据集，SMOTE类和欠类ROC曲线在ROC空间中重叠。对于所有其他数据集，SMOTE classi fier的性能优于Under classi ifer、Loss Ratio和Naive Bayes。在总共进行的48个实验中，SMOTE classi fier并非仅在4个实验中表现最佳。

解释为什么合成少数过采样可以提高性能，而替换的少数过采样则不太简单。与引入合成示例相比，当通过复制（带替换的采样）进行少数过采样时，考虑对特征空间中决策区域的影响。在复制的情况下，随着区域中少数群体样本的复制，导致少数群体分类决策的决策区域实际上会变得更小、更具体。这与所需效果相反。我们的合成过采样方法使分类器建立更大的决策区域，其中包含附近的少数类点。同样的原因可能适用于为什么SMOTE比Ripper的损失率和Naive Bayes表现更好；尽管如此，这些方法仍在从数据集中提供的信息中学习，尽管有不同的成本信息。SMOTE提供了更多相关的少数群体类样本供学习，从而允许学习者划分更广泛的决策区域，从而扩大少数群体类的覆盖范围。

DE-AC04-76DO00789。我们感谢Robert Holte提供了他们论文中使用的漏油数据集。我们还感谢Foster Provost澄清了他使用Satimage数据集的方法。我们还要感谢匿名评审员提出的各种富有见地的意见和建议。

下图显示了石油数据集的不同ROC曲线集。图29（a）显示了Oil数据集的ROC曲线，如正文所示；图29（b）显示了没有ROC凸包的ROC曲线；图29（c）显示了使用和不使用SMOTE获得的两个凸包。图29（c）中虚线和星形所示的ROC凸包是通过将Under-C4.5和Naive Bayes纳入ROC曲线族来计算的。图29（c）中实线和小圆所示的ROC凸包是通过在ROC曲线族中包括500个SMOTE-C4.5、Under-C4.5和Naive Bayes来计算的。具有SMOTE的ROC凸包在没有SMOTE时占主导地位，因此SMOTE-C4.5提供了更优化的分类器。

（a） SMOTE-C4.5、Under-C4.5、Naive Bayes的ROC曲线及其ROC凸包。（b） SMOTE-C4.5、Under-C4.5和Naive Bayes的ROC曲线。（c） ROC凸面外壳，带和不带SMOTE。