非平衡数据训练方法概述*)

张 琦 吴 斌 王 柏
(北京邮电大学计算机科学与技术学院通信软件工程中心 北京 100876)

摘 要 现实世界中数据分类的应用通常会遇到数据非平衡的问题,即数据中的一类样本在数量上远多于另一类,例如欺诈检测和文本分类问题等。其中少数类的样本通常具有巨大的影响力和价值,是我们主要关心的对象,称为正类,另一类则称为负类。正类样本与负类样本可能数量上相差极大,这给训练非平衡数据提出了挑战。传统机器训练算法可能会产生偏向多数类的结果,因而对于正类来说,预测的性能可能会很差。本文分析了导致非平衡数据分类性能差的多方面原因,并针对这些原因列出了多种解决方法。

关键词 非平衡数据,小析取项,元学习

A Survey on Imbalanced Data Learning Method

ZHANG Qi WU Bin WANG Bai
(Telecommunications Software Engineering Centre, School of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract Many real world applications involve learning from imbalanced sets problems. The class imbalance problem corresponds to domains for which one class is represented by a large number of examples while the other is represented by only a few, such as fraud detection and text classification. We often concentrate on rare cases, which are often of great interest and great value, called the positive cases. Others are called the negative cases. The number of examples of majority class is much higher than the others, which presents an important challenge to the traditional machine learning community. Traditional machine learning algorithms may be biased towards the majority class, thus producing poor predictive accuracy over the minority class. This paper gave several reason of bad performance in rare cases classification and methods for addressing the problems.

Keywords Imbalanced data, Small disjuncts, Meta-learning

1 介绍

数据挖掘的分类方法在现实世界的多个领域有着重要应用,是预测的重要方法,其典型应用有识别信用卡交易欺诈、预测电信设备的故障、从卫星图像检测油井喷发和电信领域客户的流失预测等。但是在现实世界的数据中,通常数据集中标号不同的两类样本的数量是不等的,甚至有着极大的差别,即数据集中的两类是高度倾斜的或者说是非平衡的。相关的例子很多,通过卫星图像检测油井喷发的数据集就是非平衡数据的一个好例子,因为 937 张卫星图像中只有 41 张包含浮油,我们可以说包含浮油的图像是少数类样本。

然而在数据挖掘中,通常少数类样本才是我们首要关心的,例如识别信用卡交易欺诈中的欺诈用户,样本的这一类被称为正类样本,其余则被称为负类样本。由于数量上的严重倾斜,分类算法对非平衡数据集进行分类的性能不尽人意,因为少数类样本通常比普通样本难以识别,而且大多数数据挖掘算法对于处理少数类样本有很大困难。当对非平衡数据集进行有指导的训练时,其训练算法通常对多数类样本会产生很高的预测准确率,但是对少数类样本的预测准确性却很差。通常情况下多数类样本远多于少数类样本,这意味着对所有样本进行预测,可以在不预测出任何少数类样本的情况下得到很高的正确率。诸如决策树归纳系统或者多层感知器等典型分类器被设计为使整体准确率最高,而不考虑每个类的相对分布情况,非平衡数据给这类典型的分类器提出了挑战。这些分类器在关注于将多数类样本尽量分类准确时,倾向于忽视少数类,因此传统算法对于解决非平衡数据的分类问题的能力有限。

本文深入分析了导致非平衡数据集分类结果性能低下的原因,深入地探讨了一些可能的原因,如不恰当的评估度量、不当的归纳偏移以及噪声数据等问题。针对这些原因,我们总结了多种的解决方法,致力于改进我们关心的少数类样本的分类性能。

2 非平衡数据集分类困难的原因

2.1 不当评估度量

评估度量在数据挖掘中至关重要,它用于指导数据挖掘算法并评估数据挖掘结果,例如在决策树算法进行分类的任务中,信息增益可以用于指导决策树的构建,而准确率将用于评估最终建成的决策树。如果上述评估度量不能充分评估少数类样本,则分类算法就可能对少数类样本处理不当。分类的准确率是被正确分类的样本占数据集样本总数的比例,是分类任务中最经常使用的评估度量,它在对少数类进行度量

*)本课题得到国家自然科学基金(60402011)资助。张 琦 硕士研究生,研究方向:数据仓库与数据挖掘。吴 斌 副教授,研究方向:数据挖掘,智能信息处理。王 柏 教授,博士生导师,研究方向:分布式计算。

时的缺点是显而易见的。相对于多数类来说，少数类样本对准确率的影响力较小，因为通常情况下多数类样本远多于少数类样本，这意味着对所有样本进行分类，可以在不识别出任何少数类样本的情况下得到很高的正确率。

通过实验，对 26 个数据集进行训练，平均结果显示，识别少数类样本的分类规则的错误率是识别多数类样本的规则的 2～3 倍，更重要的是少数类样本被预测出来的可能性要远远低于多数类样本，即分类中少数类样本有着远低于多数类的查准率和查全率[1]。我们通过观察发现，在极端倾斜分布的数据中，少数类的查全率可能是 0——根本没有产生少数类的分类规则。

举例说明评估度量如何指导数据挖掘系统使用的算法。我们只分析一些用于指导决策树构建的典型的度量。大多数决策树采用自上而下的方式增长，测试条件被反复选择，产生决策树的新的分支和层次。测试选择函数用于识别和评估潜在测试条件，通常采取这样的方式来评判一次测试的结果好坏——判断每个分支的纯度，然后计算总的纯度值。分支的纯度值是由遵从分支的样本的数目来决定的。这些度量（例如信息增益）倾向于选择这样的测试：在一个平衡树里大多数样本的纯度有所增长，而不是选择对于一个较小的数据集来说有着很高的纯度但对于其它的大多数样本来说纯度却很低的测试。通过测试的每个分类规则都要被赋予分类标号，标号的分配通常是基于对错误进行计算的一个阈值（通常是 50%），即用查准率来决定分类标号，问题是一个单独的高纯度分支可能识别有用的少数类样本，但是由于其涵盖的样本数量较少而导致这个分支可能不被分配标号，识别少数类的分类规则也就被忽略了。

在关联规则挖掘系统中使用支持度和可信度两个度量来指导寻找关联规则。支持度表示包含这些关联的记录的数量，而可信度表示找到关联的次数的比率。通常，关联规则挖掘系统只寻找具有支持度大于最小规定值的规则，这样大量地寻找规则的空间会被修剪。从效率的角度出发，支持率的最小规定值不可能设得太低以识别很少见的关联，因而少数类样本具有的关联规则就可能不被发现。

这些不恰当的度量导致了数据挖掘中少数类样本识别的困难，最终这些度量将被更复杂地、能更好地评估我们重点关心的正类样本的度量所取代。这些新的度量将在本文的 3.1 部分讨论。

2.2 缺少数据：绝对缺少和相对缺少

2.2.1 *绝对缺少* 非平衡数据挖掘的根本问题是我们重点关心的正类数据数量较少，以至于在少数类内部难以发现规律。研究表明，少数类具有比普通类高得多的分类错误率。深入研究还显示，少数类样本导致了训练得到的分类器中存在小析取项。具体地说，学习系统通常生成包含几种析取项的概念定义。每种析取项是最初概念的子概念的连接性的定义。一个析取项的范围取决于它正确分类的训练样本的数量，当这个数量很小的时候就被称为“小析取项”。实验表明，小析取项具有比大析取项高得多的错误率。基于 30 组真实数据集的研究表明，这些数据集训练得到的分类器所产生的绝大部分错误集中在小析取项[5]。

小析取项易于产生错误是数据缺少的直接结果，因此，理解为什么小析取项这么易于产生错误有助于理解为什么少数类样本分类困难。一种说法认为，部分小析取项不代表少数类样本或者异常样本，而代表诸如噪声等其它的数据。因此，只有有意义的小析取项应该被保留。大多数分类器归纳系统都有一些防止过度拟合的方法来去除那些被认为没有意义的子概念（如析取项）。基于统计的性能测试被用于防止过度拟合。只涵盖少数样本的析取项不会通过此类测试。问题是小析取项的性能不能得到可靠的评估，有意义与无意义的小析取项混杂在一起，而有意义的小析取项可能被去除。基于经验，我们知道去除所有小析取项的策略将导致整体错误率的上升，因而并非一个好的方法。当只有少数几个样本时，错误评估技术也是不可靠的，因此也会遇到同样的问题。这些方法对大析取项很有效，因为统计性能测试和错误率评估技术给出了对大析取项来说相对可靠的评估结果，但是对小析取项来说却并非如此[1]。

事实上小析取项本质上并非比大析取项更容易产生错误。使小析取项更容易产生错误的原因是分类器的偏移、属性噪声的干扰、缺失属性、类噪声和训练集在少数类样本上的数量等。

为了改进分类器的性能，关注小析取项问题可能比关注非平衡类问题本身更有效。实验结果显示，与使用非平衡类问题的标准或者改进的解决方法相比，把小析取项纳入考虑范畴的方法得到了更好的性能。

2.2.2 *相对缺少* 有时候样本在绝对数量上并不少，但是相对于其它类的样本来说所占的比例很小。为什么分类算法难以识别少数类样本呢？原因之一就是使用贪心搜索试探法难以发现少数类样本，而更全局性的方法通常又不容易处理。贪心搜索试探法在以下方面对处理少数类样本有困难：一是少数类样本可能需要许多条件的结合，因此孤立地检查任何条件可能无法提供足够的信息或指导。非少数类也可能遇到这样的问题，但是这种问题对少数类的影响要更大一些。参见本文 2.3 数据分裂部分。

举一个相对缺少在关联规则挖掘中的例子，食品加工机和平底锅在超市中购买的量相对较少，因此尽管二者经常被消费者一起购买，发现二者的关联还是困难的。为了发现这种关联，支持度的最小阈值就要设得很低，这样将导致关联的爆炸，因为经常出现的样本之间的关联数量将会是庞大的。这些随机的同现将埋没少数类样本之间的有意义的关联，这也是数据相对缺少带来的问题。

2.3 数据分裂

许多数据挖掘算法采用分治法，即最初的问题被分解为越来越小的问题，导致的结果就是样本空间被分解为越来越小的部分。决策树算法就是这种分治法的典型例子，在决策树算法中，所有的数据被重复地分解为越来越小的部分。这个过程可能会导致数据分裂。数据的规律只能在每个单独的数据块中找到，这些数据块却只包含了较少的数据，因而一些跨越数据块的规律可能因此丢失，这就是数据分裂问题。这在对少数类样本进行分类时问题尤为突出。因此，所有的迭代式的分治法在分析少数类样本上都会遇到困难。不使用分治法的数据挖掘算法可能会更适合非平衡数据的分类。这类方法在本文第 3 节有描述。

2.4 不当的归纳偏移

将特定样本一般化或分类器的归纳都需要一种额外的偏移。没有这种偏移，归纳步骤就不能实现，训练也就不能完成。因此数据挖掘系统的偏移对其性能来说是至关重要的。许多训练系统利用偏移来实现分类器的通用化并避免过度拟合。这种偏移可能对数据挖掘系统训练少数类样本的能力有

不好的影响。

假设一个具有最大通用偏移的训练系统，当系统决定创建一个涵盖某些训练样本的析取项时，它会选择满足这些样本但不满足其它样本的最通用的条件集。这与最大特定偏移相反，最大特定偏移会将所有可能的满足训练样本的条件添加进去。最大通用偏移对于大析取项来说很有效，但是对小析取项却相反，这导致了前文提到过的小析取项问题。因此最大通用偏移对少数类样本来说是不合适的，它可能会过于概括化，导致了产生预测少数类样本的错误率高的分类规则，其后这种规则就被修剪掉了[1]。大多数处理小析取项的方法都是通过调整训练系统的偏移来实现的。

归纳偏移本身也会倾向于对少数类样本不利，在不确定的情况下，许多归纳系统会产生有利于多数类的偏移。

2.5 噪声

噪声数据对少数类样本的影响大于普通类。因为少数类样本数量较少，少量的噪声就可以影响被训练的子概念。这样训练系统不能区分特殊（少数类）样本和噪声，如果训练系统减小其通用性，将会导致不希望得到的结果，即将噪声数据也包含进来，因此噪声数据使防止过度拟合技术（例如修剪）成为必需，以便能够减少噪声引入的小析取项，但这样导致的结果就是一些"真"的少数类样本没有被训练。如果这些少数类样本很重要，训练系统应该能够调整偏移使它们包含在内，尽管这样会导致一些不想要的结果。

3 几种解决非平衡数据分类困难的方法

3.1 使用更恰当的评估度量

在数据挖掘中引入评估度量可以更好地指导搜索过程，并更好地估计数据挖掘的最终结果。

相对于少数类来说，查准率将更多的权重放在了多数类上，这使分类器很难对少数类样本进行很好的分类，因此非平衡数据的分类需要使用另外的度量。最通用的是使用 ROC 分析和 ROC 曲线下区域（AUC）来评估整体的分类性能。AUC 不会对任何类给予更多的权重，因此它并不会向不利于少数类的方向偏移。ROC 曲线，像查准率－查全率曲线一样，也能用于不同程度的折中在引入额外的错误的正类样本的代价和增加正确分类的正类样本的收益之间进行折中。ROC 分析可以用于多数处理少数类样本的系统。

许多系统使用查准率和查全率的变形来对数据挖掘过程和结果进行评估，例如几何平均值（查准率与查全率的乘积的平方根），还有 F 度量，F 度量是参数化的，可以调整查准率和查全率二者的相对权重（F1 是按二者权值相等计算，F2 是以查全率乘以 2 计算）。查准率和查全率两者都是针对正类定义的，因而在使用这些度量时少数类样本得到较好的评估。使用遗传算法的数据挖掘系统——Timeweaver 就是使用 F 度量来评估每次迭代中进化出的分类规则是否合适，F 度量的参数可以控制查准率和查全率的相对权重，这个权重周期性的变化来确保不同的分类规则集的进化，这样一部分规则具有较高查准率而另一部分规则具有较高的查全率，最终的目标是得到查准率和查全率都比较高的结果[12]。

使用查准率来决定是否对规则给予标记是十分不可靠的，如果规则涵盖了较少的训练数据则不会被赋予标记。因此我们使用一些新的度量来对少数类和小析取项的关联规则的准确率做更好的评估。一个度量是 Laplace 评估。这个度量的标准定义是 $(p+1)/(p+n+2)$，其中 p 是分类规则涵盖的正类样本数，n 是分类规则涵盖的负类样本数。这种度量使查准率评估偏向 1/2，但是在样本数量增长的情况下其重要性就会下降。处理少数类样本和小析取项的更复杂的错误评估度量是 Quinlan 提出的，这种度量将类的分布考虑进去来改进小析取项的查准率评估。这种方法使用了更具代表性的一个样本集，而不是用整个训练集来计算类的分布。这个样本集只使用"近似"小析取项的训练样本，就是在析取项中最多不满足一个条件的样本。Quinlan 的实验结果表明，在数据挖掘中使用这种评估，分类的性能会得到改善，对严重倾斜的数据的效果尤其显著[1]。

3.2 非贪心搜索技术

前面已经提过为什么贪心搜索技术处理少数类样本的性能不高，本文这部分主要描述专门处理少数类问题的非贪心搜索方法。

第一个方法使用了遗传算法。遗传算法是全局的搜索技术，使用的是候选解的总数而不是某个单独的解，并使用随机算子来指导搜索的过程。这些特征使遗传算法能够处理属性的交互并防止停留在局部最优上，这也是遗传算法适合处理少数类样本的原因。

决策树训练算法几乎都是使用贪心搜索算法的。为了处理这些贪心的、登山搜索算法，我们使用了 Brute 方法。Brute 方法使用穷举的深度限制搜索，以寻找精确的规则，尽管有的规则只包含少数几个训练样本。相对于其它算法而言，Brute 方法的性能相当好，可以发现别的算法所不能发现的"金块"信息，但是它产生的规则的长度需要受限。

关联规则挖掘系统通常采用穷举搜索算法，因此理论上可以找到少数类样本之间的关联。问题是支持度的最小值如果设得太小以发现少数类样本之间的关联的话，这些算法得到的规则的规模就是不可控制的了，因此这种穷举搜索算法也是无法发现少数类样本之间的关联的。

3.3 使用更合适的归纳偏移

大多数数据挖掘系统使用偏移来进行通用化处理。一个通用偏移对普通样本来说是有好处的，但是对少数类样本则是不合适的，甚至会导致少数类样本完全被忽略。

改进非平衡数据挖掘的性能的途径之一就是选择更合适的偏移。已有的最简单的方法就是使用统计权值测试或错误计算技术来去除一些小析取项，希望能够只去除训练不当的小析取项。但是这种方法不仅降低了少数类的分类性能，也降低了总体的分类性能，因此只能采用更为复杂的方法。这些方法对少数类样本的影响不能直接度量，判断这些方法的效果只能通过小析取项的性能的改善，以上是基于少数类样本都以小析取项的形式出现的假设。Holte，Acker 和 Porter 改变了已有的训练系统的偏移—CN2，以使偏移更具体，即为得到的小析取项计算更具体的偏移值，而不是对所有的析取项都使用 CN2 的最大通用偏移值。使用特殊偏移的最大值得到的结果显示小析取项的性能得到了改善，但是大析取项的性能有所降低，导致了整体性能的下降。这种对偏移的改变过于极端，于是改为使用选择性的特殊偏移。这带来了少数类样本分类性能的更大的改善，但是没有改善整体分类的准确率[1]。

Ting 使用了一个相似的方法，也是用特殊偏移的最大值，但是采取了一些方法来避免这种偏移使大析取项的性能降低。基本方法就是首先使用 C4.5 决策树训练器来判定一个样本是涵盖于大析取项还是小析取项，如果是大析取项则

使用 C4.5 对样本进行分类，否则使用一个基于样本的训练器对样本进行分类。文[1]研究的结果是乐观的，也表明混和法可以改善小析取项的性能，但是这种改善并非确定的。

另一项研究使用了近似混和法。也是使用 C4.5 来判定样本是属于大析取项还是小析取项，归于小析取项的样本接下来被输入到基于遗传算法的训练系统，这个训练系统产生落入每个单独的小析取项的样本的分类规则，而归于大析取项的样本使用 C4.5 进行分类。

总的来说研究者们采用了不同的方法尝试选择一种适宜的偏移，以便在小析取项的问题上得到更好的性能。这些方法只是混和法的成效，原因可能是由于研究者还是使用整体分类的准确率来评估方法的性能，而不是集中关注小析取项的性能。从偏移方面着手处理少数类问题是有希望的，值得更深入的研究。

3.4 只训练少数类

分类的典型方法就是从两类数据中选取样本，然后产生用以区分它们的模型。这是最常规的方法，大多数研究者认为只有在训练集包含两类数据时才能建模。此外，对许多机器训练算法如决策树、朴素贝叶斯或者多层感知器等来说，如果训练集数据不是包含来自两类的样本，算法是无法运行的。但事实上对于许多应用来说，得到少数类数据的样本是困难的，在这种情况下，对含有单一类的数据进行训练是唯一可能的解决办法。此外，在数据严重不平衡的情况下，例如欺诈检测和信息过滤，有指导的训练算法需要对数据进行某种形式的平衡。其中忽略大部分负类样本仅学习正类样本本身，是更快速的方法，但并不是一个有保证的方法。

Hippo 就是使用这种基于识别方法的数据挖掘系统，它使用神经网络算法，只训练正类(少数类)样本，以此发现正类样本的识别模式，而像传统方法那样发现非正类和负类样本之间的区别。支持矢量机(SVM)也可以用这种方法训练少数类样本，并取得了成效[3]。只训练少数类的系统也可以使用来自所有类的样本。Brute 、Shrink 和 Ripper 就是三个这样的数据挖掘系统。Brute 方法已经被用于查找波音飞机制造过程的错误，它主要关注找到错误(正类样本)和预测错误的规则的性能，而不关心负类的准确率。Shrink 使用相似的方法从卫星雷达图片中检测油井喷发。Ripper 使用分治法，迭代的产生规则来涵盖那些没有被涵盖的样本。每条规则的增长都是通过不断添加条件直到没有负类样本涵盖在内。这样可以为每一类都产生规则，无论是多数类还是少数类。有了这样的架构，就可以直接为少数类训练规则。

不是所有产生的分类规则都会被使用，大多数数据挖掘系统都会对每条规则产生一个质量估计，通常是基于训练集数据对规则的准确率评估。分类规则就可以按照这种质量度量来排序，用户可以从 n 条规则中选择 m 条最好的，$m \leqslant n$。改变 m 的值就可以产生准确率/涵盖率的曲线，依据这个曲线可以根据需要选择特定的点。使用这种方法数据挖掘系统就不需要在训练之前决定停止产生正类预测规则的点。

支持矢量机(SVM)的极限重新平衡方法是这类方法中的一种，即完全忽略多数类的样本。实验结果显示，分类性能的确通过 SVM 的极限重新平衡方法得到了改善[3]。

噪声、特征的稀少和少数类所占的低比例，所有这些性质导致了使用单一类数据进行训练具有更好的性能。实验表明，处理严重非平衡的并具有高维度噪声特征空间的数据时，对正类样本进行单一类的训练是一种具有很好的健壮性的分类技术。这种技术可以作为过分特征选择的替换方法，在使用非线性内核时得到的结果很好，这是在特征空间水平直接进行的特征选择所做不到的。

3.5 分割数据

处理非平衡数据的一种方法就是通过小心的分割数据来降低数据不平衡的程度。有效地分割数据就会让初始的数据挖掘问题分割为子问题。

假设某目标事件是很少见的，只有 0.001%。那么可能可以把数据分为两块，R1 和 R2，在 R1 中目标事件有 20%，而在 R2 中只有 0.0001%。这样对 R1 的处理就可以避免数据极端不平衡的问题，而对 R2 可以完全不考虑，因为 R1 涵盖了目标事件的绝大部分，所以这样的结果可能是可以接受的。

3.6 为少数类样本制定不同的统计标准

关联规则挖掘在少数类样本上的问题前面已经提过，为了避免关联规则的数量的爆炸，最低支持度的值－minsup 不能定得太低，这导致了少数类样本之间的关联被埋没。这种问题可以通过这样的方法解决，即根据产生关联的对象出现的频率不同而指定不同的最低支持度。

3.7 代价敏感训练

处理非平衡数据挖掘问题的一种方法就是代价敏感训练。由于正确识别正类样本的价值远超过正确识别负类样本的价值，因此可以为负类样本误判为正类样本和正类样本误判为负类样本两种错误指定不同的代价，这可以使分类器向有利于少数类样本正确分类的方向偏移。

这种方法的问题在于用于判断指定代价的信息不易得到。部分原因在于代价通常取决于多个条件，不容易权衡，因此更实用的做法是只预测少数类并产生最佳正类预测规则的排序表[16]。这样在数据挖掘完成后可以判断阈值取多少合适。大多数的数据挖掘系统可以直接处理代价敏感方法，代价信息可以传递到数据挖掘算法中。

3.8 采样

处理非平衡数据的最常用方法就是采样。采样方法的基本思想就是通过改变训练数据的分布来消除或减小数据的不平衡。

3.8.1 *基本采样方法* 基本采样方法包括 under-sampling 和 over-sampling。under-sampling 是通过减少多数类样本的数量来平衡两类样本，而 over-sampling 则是通过复制少数类样本来完成。这两种方法都能减小数据整体的不平衡程度，但是都有一些缺点。under-sampling 忽略了潜在的有用的多数类样本，因而可能降低分类器的性能。而 over-sampling 引入了额外的训练数据，这会延长建立分类器所需时间，更不好的是它对样本进行精确的复制会导致过度拟合。在极端的情况下，分类器会产生只涵盖一个被复制多次的样本的规则。同时这种方法也没有解决本文前面提到的数据缺乏问题。因此 under-sampling 可能是更好的方法。当然如果是使用人工数据进行研究就不会有上述问题出现。

3.8.2 *高级采样方法* 高级采样方法将 under-sampling 和 over-sampling 巧妙地结合起来。一种 under-sampling 方法就是只去除相对其它样本来说是冗余的多数类样本，或者去除与少数类样本接近的多数类样本，认为它们是噪声数据。SMOTE (Synthetic Minority Over-sampling TEchnique)则引入新的非重复的人造少数类样本[19]。这种方法增加了通用性，而不是像精确复制样本一样会引起过度拟合。

一种混和式专家方法将许多分类器的结果联合起来，每个分类器都是通过使用不同的 under-sampling 和 over-sampling 的比例的方法得到的。这样还是不能确定哪种采样方法更好，最好的方法就是根据不同的领域灵活变化。结果显示混和式专家方法性能较好，在文本分类问题上查准率和查全率都超过其它的方法。

另一种方法是先确定一种性能较好的类分布比例，然后按照这种比例采样，产生多个训练集。每个训练集包含所有的正类样本和负类样本的一个子集。但每个多数类样本要保证在训练集中至少出现一次，这样不会有数据遗漏。对每个训练集分别进行训练，然后用元学习的方法将得到的多个分类器合成一个分类器。这种方法可以用于任何算法，但是需要预先知道一种较好的类分布比例，这可以通过一些准备性的工作得到，不过增加了训练所需时间。

3.9 元学习

元学习的概念是由 Prodromidis 等人于 2000 年首先提出的，该方法采用集成学习的方式来生成最终的全局预测模型（即元分类器）。该方法的基本思想是从已经获得的知识中再进行学习，从而得到最终的数据模式。

基本分类器输出的集成方式有三种：

投票：绝对（相对）多数投票，加权投票。

决策：指定特殊的“决策者”，当各基本分类器的输出无法达成一致时，采用“决策者”的输出。

结合：使用相关的先验与领域知识指导各输出的综合方法。

元学习的优点是在基本训练阶段，可以自主地选择合适的训练算法来生成独立的基本分类器；在元学习阶段，由于系统可灵活采用各种集成策略，因此最终生成的元分类器具有较高的预测精度。

在非平衡数据的分类问题上，由于元学习的每种基本算法都有独到之处，因而某种基本算法会对某类特定数据样本比其余的基本算法更为有效。从不同的算法中选择最佳分类器然后将它们预测的结果合并，要比单一使用某种方法的效果要好，这些方法包括从同一算法得到多个分类器、合并每个算法的结果、从所有分类器中选择最好的等。

3.10 其它方法

3.10.1 *推进算法* 推进算法，如 AdaBoost，就是迭代的给训练集数据的分布加权值。在每次迭代中，推进算法会增加没有正确分类的样本的权值，减小正确分类的样本的权值。这使训练系统在下次迭代中更关注于分类错误的样本。少数类样本更容易产生错误，所以有理由相信推进算法可以改进少数类样本的性能。因为这种算法有效地改变了训练数据的分布，我们可以认为它是高级采样方法的一种。

AdaBoost 的权值更新规则可以设为代价敏感的，所以分类错误的少数类样本比多数类样本被赋予了更高的权值，这就是 Adacost 分类系统。依据经验，这种方法会得到比 AdaBoost 低的错误分类代价。

还有一种算法也使用推进方法处理非平衡数据问题—SMOTEBoost[19]。这种算法认识到推进算法也可能遭遇 over-sampling 的问题，例如过度拟合，这是因为推进算法倾向于赋予少数类样本更大的权值，这等于复制了部分少数类样本。因此 SMOTEBoost 方法并不是通过更新每类样本的权值来改变训练数据的分布，而是通过使用 SMOTE 算法来添加新的少数类样本。经验显示，这种方法将得到比 Adacost 更高的 F 度量。

推进算法并不确保分类的性能一定会得到改善，分类的性能与所用的基本训练算法密切相关，如果基本训练系统可以有效地在查准率与查全率之间权衡，那么推进算法可以显著的改善它们的性能[13]。

3.10.2 *两阶段规则归纳* 大多数分类系统都试图同时使查准率和查全率这两个存在竞争的度量提高，这是比较困难的。PNrule 使用两阶段规则归纳来分别关注这两个度量。

第一阶段，如果不能发现高查准率的规则，但这些规则具有相对来说较高的查全率，那么这种低查准率的规则也将被接受，即第一阶段关注于查全率。第二阶段主要优化查准率，这是通过训练并识别第一阶段得到的规则中分类错误的正类样本来实现的。第二阶段使第一阶段允许对小析取项敏感，而第二阶段允许分类错误的正类样本被分组聚在一起，解决了前文提过的数据分裂的问题。

实验结果显示，在简单问题上 PNrule 得到的结果和其它训练系统差不多，但是在引入更复杂的概念时，其它的训练系统的性能都会下降而 PNrule 方法则保持了较高的性能。结果显示 PNrule 比 AdaBoost 得到的性能要好，特别在少数类问题上。

3.10.3 *DataBoost-IM 方法*[2] 在过去的几年中，集成法是一种前景看好的技术，它可以改进弱分类算法的性能。分类器的集成包括一系列独立训练得到分类器，这些分类器被集成后用于分类。推进算法是一种集成法，这种方法通过关注难以分类的困难样本来改进弱分类器的性能。推进算法产生一系列的分类器，这些分类器的输出结果用加权投票等方法集成起来作为模型的最终预测结果。在每个步骤中，基于早期分类器的性能，训练样本被重新加权和选择。这样将产生一部分有着较低权重的“容易”样本和一部分有着高权重的“困难”样本。在每次迭代中，推进算法试图产生新的分类器，这些分类器能更好地预测以前的分类器预测得很差的样本。这是通过集中对困难样本进行分类来实现的。近期的研究显示，推进算法适用于很大范围内的问题，并取得了很好的效果。

DataBoost-IM 方法是一种新的方法，将推进算法和基于整体学习算法结合，并结合产生人工数据来改进针对非平衡数据集的分类器的预测性能。在 DataBoost-IM 方法中，来源于多数类和少数类的困难样本在推进算法的执行中都被识别出来。然后困难样本用于分别产生多数类和少数类的人工样本。人工数据被加入最初的训练集，在新的训练集中，类的分布和不同类之间的总权重被重新平衡。

DataBoost-IM 方法用 F 度量、G 度量和整体准确性这些度量评估，并与十七个使用决策树作为基本分类器的非平衡数据集的分类结果进行对比。结果显示，与基本分类器、标准基准点推进算法和三个针对非平衡数据集的基于推进数据的高级算法相比，DataBoost-IM 的方法的效果很好，这种方法没有因为数据不平衡的原因而牺牲任何一个类，多数类和少数类二者都产生了很好的预测结果。

这种方法和以前的方法在以下方面不同：首先，分别从少数类和多数类中识别困难样本和产生人工样本；其次，生成的人工样本向困难样本的方向有信息偏移，这些困难样本是下个分类器组件在推进过程中需要关注的。这就是说，它为多数类和少数类一起提供了附加的知识，防止了推进算法过分关注那些困难样本；第三，在新的训练集中，类的分布被重新

平衡，以便减少训练算法向多数类的偏移，即使用来自多数类的数量被削减的样本和少数类的样本，以确保训练集中两类都有代表；第四，在新的训练集中不同类的总权重被重新平衡，使推进算法不只关注困难样本，也关注少数类的样本。这样，多数类和少数类两者的预测都得到了改善。

3.10.4 *特征选择* 在文本分类中应用了一些特征选择度量，其中信息增益(information gain-IG)、χ^2(chi-square)、相关系数(correlation coefficient-CC)和优势率(odds ratios-OR)是最有效的。CC 和 OR 是单边度量，而 IG 和 CHI 是双边度量。使用单边度量来选择特征，只能选出表现出"是成员"的特征，而使用双边度量结合了表现出"是成员"(正特征)和"不是成员"的特征(负特征)。单边度量是不考虑负特征的，而负特征却是非常有价值的。但双边度量不能保证两种特征是最优结合的，尤其对非平衡数据来说。

双边度量通过简单的忽略正负特征的标志和不将它们的值进行对比，隐含地将正负特征合并。然而在非平衡数据中，正特征的值和负特征的值是不可比的，而且不同的性能度量会给正负特征加上不同的权重，因而两种特征的最优比率也是取决于性能度量的，所以双边度量并不能保证正负特征是最优结合的。而且单边和双边度量都不允许对两种特征的合并进行控制。

最近提出了一种新的特征选择架构，在这个架构中，正负特征独立选择，然后显式地合并。在架构中几种标准方法被统一起来，并根据数据特征和性能度量为每一类提出了一系列最优化合并正负特征的新方法。实验表明根据非平衡数据本身，选择接近最优的方式将正负特征结合，具有巨大的潜在和实际的优点[9]。

3.10.5 *BMPM 方法*[10] Biased Minimax Probability Machine (BMPM)方法与以往的方法不同，并不采用采样、调整决策树阈值或调整代价值的方法，而是基于拓展的 Mimimax Probability Machine (MPM)算法的一种新方法。这种方法的模型通过直接控制数据分类在最差情况下的真实查准率来建立有偏移的分类器。在给定可靠的平均值和方差估计的情况下，这种模型通过直接控制真实查准率的下限来建立分类的边界，因而可以对非平衡数据做系统化的严格的处理，与三种算法——朴素贝叶斯、K-最近邻居方法和 C4.5 决策树方法相比，BMPM 的效果最佳。

总结 本文分析了非平衡数据集中我们通常主要关注的少数类样本难以正确分类的原因，列出了一系列针对这些原因的算法解决方法，其中的一些方法已经在医疗诊断、欺诈检测等领域得到了应用，与单纯使用传统机器训练算法相比，其预测的性能得到了改善。

但是这些方法并非适用于每个应用领域，它们对非平衡数据集的分类性能改进的效果也是不确定的，因而需要结合应用领域的特点和专业经验来选择合适的方法，并进行多次实验、反馈、修正和改进，才能得到最优效果。

参 考 文 献

1 Weiss G M. Mining with Rarity: A Unifying Framework. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

2 Guo Hongyu, Viktor Herna L. Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

3 Raskutti B, Kowalczyk A. Extreme Rebalancing for SVMs: a case study. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

4 Batista G E A P A, Prati R C, Monard M C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

5 Jo T, Japkowicz N. Class Imbalance versus Small Disjuncts. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

6 Phua C, Alahakoon D, Lee V. Minority Report in Fraud Detection: Classification of Skewed Data. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004, 6(1)

7 Petrushin V A, Kao A, Khan L. The 4th Intl. Workshop on Multimedia Data Mining(MDM/KDD2003), Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

8 Dolores del Castillo M, Serrano José Ignacio. A Multistrategy Approach for Digital Text Categorization from Imbalanced Documents. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

9 Zheng Zhaohui, Wu Xiaoyun, Srihari Rohini. Feature Selection for Text Categorization on Imbalanced Data. Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2004,6(1)

10 Huang K, et al. Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine, Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proc. of the 2004 IEEE Computer Society Conf. on ,2004,2: Ⅱ-558～Ⅱ-563

11 Huang Kaizhu, Yang Haiqin, King I, Lyu M R. Data mining from Extreme Data Sets. Very Large and/or Very Skewed Data Sets. Computer Vision and Pattern Recognition, 2004. CVPR 2004. In: Proc. of the 2004 IEEE Computer Society Conf. on, 2004, 2: II-558～II-563

12 Joshi M V. On Evaluating Performance Of Classifiers for Rare Classes, Data Mining. In: 2002. ICDM 2002. In: Proc. 2002 IEEE Intl. Conf. on ,Dec. 2002. 641～644

13 Joshi M V, Agarwal R C, Kumar V. Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?. In: Proc. of the eighth ACM SIGKDD intl. conf. on Knowledge discovery and data mining

14 Chen C, Liaw A, Breiman L. Using Random Forest to Learn Imbalanced data. http: // stat-www. berkeley. edu/users/chenchao/666. pdf

15 Provost F. Machine Learning from Imbalanced Data Sets. http: // pages. stern. nyu. edu/～fprovost/Papers/skew. pdf

16 Maloof M A. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. http: // www. site. uottawa. ca/～nat/Workshop2003/maloof-icml03-wids. pdf

17 Chawla N V. C4. 5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure. http: // www. site. uottawa. ca/～ nat/Workshop2003/chawla. pdf

18 Laurikkala J. Improving Identification of Difficult Small Classes By Balancing Class Distribution. http: // www. cs. uta. fi/reports/pdf/A-2001-2. pdf

19 Chawlal N V, et al. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), 107～119

20 Visa S, Ralescu A. Experiments in Guided Class Rebalance based on Class Structure. http: // morden. csee. usf. edu/avatar/publications/SMOTEBoost7. pdf