不平衡分类学习综述

陈晶*

(南京大学 计算机科学与技术系, 南京 210093)

Research on Class-Imbalance Learning

Chen Jing*

(Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

Abstract: There is a growing number of real-world applications showing characteristics of class-imbalance classification suffering from severe class distribution skews and underrepresented data. The complex characteristics of imbalanced datasets make it hard for mature and classical classifiers to take effect, thus require brand new principles, algorithms, and assessment metrics to learn from such datasets. Large amount of research work has proposed various pragmatic solutions to class-imbalance problems both academically and industrially over years. In this paper, we focus on the analysis of the nature of imbalanced learning, demonstration of state-of-art algorithms for class-imbalance problems and discussion of efficient assessment metrics to evaluate performance of such algorithms. Finally, major challenges and potential research directions are mentioned as well.

Key words: imbalanced learning; classification; sampling; cost-sensitive learning; ensembles; assessment metrics

摘 要: 大量实际应用中的分类任务呈现出不平衡分类问题的特征，不同类别在数据集上的分布高度不平衡，甚至有的类别样本量极少。不平衡数据集的复杂特性使得传统成熟的分类算法无法获得发挥优势，难以适应不平衡分类问题。为此，对不平衡分类问题的探讨需要全新的视角、算法以及评估指标。多年来，学术界和工业界提出并成功实践了许多富有创造性且高效实用的不平衡学习算法。本文将重点分析不平衡分类问题的性质，试图展现具有代表意义的不平衡学习算法，并讨论评估上述算法的有效指标，文章最后也给出了不平衡学习领域潜在的挑战与方向。

关键词: 不平衡学习; 分类; 采样方法; 代价敏感学习; 集成方法; 评估指标

1 引言

分类问题是数据挖掘领域主要任务之一，也是机器学习领域重要的研究方向。分类算法从训练集中学习得到分类函数，利用该函数预测未知样本的类别。经过长期研究与发展，目前有许多成熟的分类模型，在以数据类分布大致平衡为前提，也即不同类别样本数量大致相当的分类问题中，取得了可喜的效果。但是，上述对数据类分布大致平衡的假设是非常局限的。事实上，实际应用中的大量分类问题都是不平衡的，而传统分类算法难以适应并处理不平衡的分类问题，这引发了业界对不平衡学习热烈的探讨与研究。

关于不平衡分类问题的讨论与研究二十余年来一直非常活跃，过去的十几年里有多次最高学术级别的workshop 和会议都以该问题为主题开展，领域内诞生了大量富有创造力和巨大意义的成果。本文试图分析不平衡分类问题的性质，整理并综合学者们的研究以梳理该研究方向的主要进展。文章第 2 节深入分析、理解

* 作者简介:南京大学 计算机科学与技术系 2011 级学生

不平衡分类问题及其带来的挑战， 第 3 节概述地呈现并简单分析当前主流的各类不平衡学习算法，第 4 节讨论不平衡学习算法的评估标准，第 5 节描述该领域的研究发展趋势，最后作一总结。

2 理解不平衡分类问题

从本质上看，对类别分布不平衡的数据集进行分类是指，在分类问题中，某些类的样本数量远远多于其他类别的样本数量。严格讲，数据集中任两个不全等的类分布都可以看作是不平衡的分布；然而，业界通常认为只有当类在数据集上的分布极端不平衡，例如二元分类中达到 100:1，1000:1，甚至 10000:1 时[1]，才是真正意义上需要处理的不平衡分类问题。例如，寻找电信运行商的逃离客户[2]和从卫星图片对油井定位[3]等等。该方向上的大量研究工作集中在二元分类问题，本文将以二元分类为重点，对不平衡分类问题展开分析。

分类任务的长期发展产生了大量成熟且实用的分类算法，但这些算法难以直接应用于不平衡学习。传统方法以降低分类错误率$ErrorRate$作为分类模型的优化目标，不平衡分类问题的出现对这一标准提出质疑。典型地，考察二元分类癌症数据集，每条记录都属于“正例”或者“反例”，分别表征癌症患者和非癌症患者，实际中该数据集上反例的数量远远大于正例数量，为了方便解释，这里不妨假设反/正例比例为为99：1。一种朴素粗糙的方法，忽视少数类，直接把所有记录全部分为反例以降低分类错误率，事实上这种方法在该数据集上仅有1%的错误率，在$ErrorRate$评估下这种方法可谓表现超群。大多数传统分类器基于最小化分类错误率的目标通常也会采取类似的策略，倾向于忽视少数类，使得最终分类结果偏向于多数类。可是，实际问题中尤其是上述病情检测的场景下，误将癌症病人鉴别为健康是不可接受的。换言之，实际应用背景中的不平衡问题对少数类的误分是不可容忍、需要巨大代价的。因此，传统分类算法几乎无法适应不平衡分类问题。

不平衡分类问题对传统的分类算法提出了两点质疑：1. 传统分类算法在不平衡分类问题中能高度识别多数类，却几乎难以识别少数类，无法直接应用于新的问题之上；2. 仅仅实用分类错误率作为分类评估指标，无法从本质上代表不平衡学习算法的优劣。下面，本文将针对这两个问题分别展开讨论。

3 不平衡学习算法

不平衡分类问题的处理应用广泛，学者们在过去的二十余年展开大量深入的探索与研究，提出了许多具有里程碑意义的思路与学习算法[4]。从处理层面上看，这些算法大多从三个层面入手：1.数据层面；2.学习层面；3.数据与学习结合的层面[5]；从本质上看，Breiman et al.指出，几乎所有不平衡学习算法都在调整和控制四个核心影响因素[6]：训练集大小(training set size)、类优先级(class priors)、类别误分代价(cost of errors in different classes)和决策边界的设置(placement of decision boundaries)。这一节将讨论处理不平衡分类问题的几种主流思路，并分析最有代表意义的不平衡学习算法，以理解它们分别是如何通过控制上述四个因素进行不平衡学习的。

为了方便下文的刻画与分析，将不平衡分类问题数学化，抽象如下：

一般分类问题中，给定m个样本的数据集$S(|S|=m)\colon S=\{(x_i,y_i)\},i=1,\dots,m$，其中$x_i\in X$，X为n维特征空间；$y_i\in Y=\{1,\dots,C\}$表征$x_i$的类别，特别地，$C=2$时问题简化为二元分类。针对不平衡分类问题，定义S的少数类样本集$S_{min}\subset S$和多数类样本$S_{maj}\subset S$，有$S_{min}\cap S_{maj}=\emptyset$且$S_{min}\cup S_{maj}=S$。最后，定义S采样生成的新数据集为N，类似地，N上有N_{min}和N_{maj}。

3.1 Sampling采样法

不平衡学习困难的直观原因，是不同类别样本数量的悬殊差别。基于采样的算法着眼于对样本数量的调整，试图通过一些方法平衡原始数据集，然后在“平衡化”的新数据集上采用分类算法进行训练学习，是从数据层面处理不平衡分类问题。研究表明，sampling 方法能够有效处理不平衡数据集上的分类问题[7]。

3.1.1 Random Oversampling/Random Undersampling 随机向上／向下采样

顾名思义，random oversampling 在原数据集S的少数类S_{min}中随机选样产生集合N，复制N并将其中所有样本加入S更新数据集，$|S_{new}|=|S_{min}|+|S_{maj}|+|N|$；重复该过程可以任意调整S的类别比例。同理，random undersampling 在S中的多数类S_{maj}中随机选样产生集合N，将N从S中去除，$|S_{new}|=|S_{min}|+|S_{maj}|-|N|$，也能够调整S的类别比例。

Random oversampling 和 random undersampling 均通过改变少数类或多数类的样本数量以调整类别比例，是采样法最简单易懂的实现。但是，它们在实际应用中有着各自不可忽视的缺陷。Oversampling 通过复制添加少数类样本，导致分类器在学习过程中易产生过拟合现象，也即分类器过分依赖训练数据，在训练数据上分类表现良好，但是预测陌生数据时效果非常差[8]；另一方面，undersampling 方法随机去除多数类中的样本，容易丢失大量对训练分类器十分重要的训练数据，这样学习得到的分类器显然也是不够理想的。总的来说，随机采样算法在实际中应用并不多，但它的思想是非常宝贵的。

3.1.2 Synthetic Minority Oversampling Technique (SMOTE) 采样法

SMOTE 是一种 oversampling 算法，自提出后便在理论研究与实际应用中取得极佳表现[9]。它的基本思想是考察分析少数类样本的特征空间，合成并向数据集中增加人工数据。具体如图 1，给定k，对每个少数类样本$x_i \subset S_{min}$，计算x_i的k个近邻，随机选择k近邻中一个x_t，利用下式合成人工数据x_{new}并将其加入原数据集：

$$x_{new} = x_i + (x_t - x_i) \times \sigma \tag{1}$$

其中，$x_i \in S_{min}, x_t$为x_i近邻，$\sigma \in [0,1]$是随机数。

仔细分析，SMOTE 采样抛弃了 random oversampling 单纯复制少数类样本的做法，避免了分类器过拟合的问题。同时，SMOTE 提供了一种通过增加样本以平衡原数据集的有效方式，实践证明这种平衡方式成功地提高了分类性能。SMOTE 算法对用 sampling 方法解决不平衡分类问题的方向有重要意义，然而它仍然存在不少缺陷，更多基于 SMOTE 的算法由此提出。

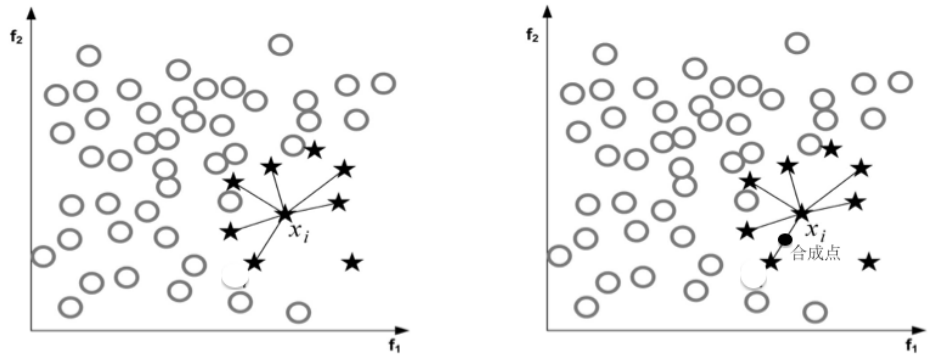

图 1 (a) (b)

图 1 (a)SMOTE 对少数类样本点x_i求近邻($k = 6$，且X为 2 维特征空间)；(b)对样本点x_i合成人工数据

3.1.3 基于 SMOTE 的采样算法

由于 SMOTE 采样对少数类中每一个样本都合成新样本点，这种做法忽视了被选定近邻与当前样本点可能的类别不同，有极大可能会出现新样本点类别重叠的问题，也即同一个合成的样本点在不同处理过程中被标记为不同的类别，造成 SMOTE 算法不容忽视的一个缺陷。针对这一问题，学者们提出了许多解决方案，典型如 BorderlineSMOTE 和 MSMOTE，这里简单分析 BorderlineSMOTE 规避上述问题的方式[10]。

在 BorderlineSMOTE 中，对少数类S_{min}的每个样本点x_i求k近邻点集，记作$S_{i-knn} \subset S$。若满足：

$$\frac{k}{2} < |S_{i-knn} \cap S_{maj}| < k \tag{2}$$

则将x_i加入DANGER集合，显然DANGER表征了少数类S_{min}中靠近类别界线(假设少数类与多数类的样本空间是有一定界限的)的所有样本点集。将DANGER集合作为 SMOTE 的输入，开始新样本点的合成算法。

不同于 SMOTE 对少数类所有样本进行样本点合成，BorderlineSMOTE 仅仅对靠近类别界线，也即在DANGER集合进行样本点合成。根据公式(2)中定义，选定阈值$\frac{k}{2}$和k，使得 BorderlineSMOTE 的处理几乎不会出现类别重叠问题，就此解决了单 SMOTE 的巨大缺陷。

3.1.4 informed undersampling 采样法

undersampling 思想也能够有效地处理不平衡分类问题。典型的 BalanceCascade 算法[11]，就是一种有指导的 undersampling 方法。

BalanceCascade 从多数类S_{maj}中选样产生N ($|N| = |S_{min}|$)，以$T = E \cup S_{min}$T 作为训练集学习得到分类器$C(1)$，对任意$x_i \in S_{maj}$，若x_i被$C(1)$正确分类，则$S_{maj} = S_{maj} - x_i$；迭代上述过程直到获得一组联合的分类器用于分类学习。

BalanceCascade 算法避免了 random undersampling 方法由于随机去除样本点导致的信息遗失；它采用迭代框架，每一步迭代仅仅去除那些已经被当前分类器正确分类、对后继分类器学习“没有帮助”的样本点，整个处理过程保留了所有有效的信息，在实验中取得了非常好的效果。

3.1.5 其他采样法

除了上述几种典型的采样算法，学者们在研究中还提出了大量高效的算法。带数据清洗技术的采样(如 SMOTE+Tomek Links)[12]，可以有效去除类别重叠问题；基于聚合的采样方法[13]也取得了优越的表现和性能。

3.2 cost-sensitive代价敏感学习方法

应用中的不平衡分类问题上，类别误判的代价天差地别，这间接决定了不平衡学习难以直接利用经典分类算法，故而对新分类学习算法提出要求。本质上讲，代价敏感学习方法着眼于学习算法的层面，详细考虑不同类别错判的代价。研究表明这一类算法能够很好地适应不平衡分类问题，在许多实际问题中的表现甚至大大超过前一节的 sampling 采样法[14]。

3.2.1 代价敏感学习框架

代价敏感分类学习方法的核心要素是代价矩阵，由类别误判的惩罚系数构成，不妨用$C(i,j)$表征将类别j误分为类别i的代价。特别地，对于二元分类，$C(+,-)$是把反例误判为正例的代价，$C(-,+)$是把正例误判为反例的代价[15]。基于上述代价矩阵，不平衡分类问题简化为：在当前数据集和代价矩阵下，使所有样本分类后的误分总代价最小的最优化问题。通常，认为识别正例的意义大于识别反例的意义，因此，把正例误判为反例的代价高于把反例误判为正例的代价，即$C(-,+) > C(+,-)$；另外，正确分类的惩罚为 0 ($C(+,+) = C(-,-) = 0$)。

3.2.2 代价敏感学习方法的实现方式

通常来说，基于代价敏感的学习算法有三种实现方式。1. 将误判代价作为原始数据集上的权重，然后用 sampling 方法调整数据集类别比例，根据该权重选择最优的类别比例；2. 引入集成方法，将代价最小化技术融入分类学习算法中，以期获得一个集成的分类模型；3. 直接将代价方程加入现有的分类学习算法进行修正，由于现有的分类学习算法非常多，所以这种实现方式没有统一的算法，对不同的分类学习算法有不同的修正方式。值得一提的 AdaCost 是一种典型的基于集成方法的代价敏感学习算法，将在 3.3.2 节详细讨论。

3.3 Ensemble集成方法

集成方法的基本思想是，利用算法从训练集学习得到一系列子分类器，然后利用这些子分类器的某种集成来提高分类准确率。常用的集成方法有 bagging 装袋、boosting 提升、random forest 随机森林等，其中 boosting 方法应用最为广泛[16]。

实际上，boosting 就是一个迭代过程，用来自适应地改变训练样本的分布，使子分类器更容易聚焦在某一类样本。通过 boosting，多个弱子分类器可集成为一个强大的分类器，因而在不平衡数据集上能获得有效的表现。AdaBoost 是 boosting 方法的典型代表，它对训练机数据的分布迭代加权，在每次迭代中增加错误分类的样本权重，减少正确分类的权重；由于普通分类算法在不平衡分类问题中，通常会误分少数类样本，这使得训练系统在下次迭代中更关注于少数类样本。

3.3.1 SMOTEBoost 算法

SMOTE 算法与 AdaBoost 算法集成得到的 SMOTEBoost 方法[17]，取得了巨大的成功。它在 AdaBoost 的每步迭代中进行 SMOTE，使每个后继的分类器更加聚焦于少数类。由于迭代得到的子分类器们实际学习产生于不同的样本，所以集成分类器会联合每个分类器的投票，最终决定样本的类别。SMOTEBoost 算法巧妙利用 AdaBoost 迭代框架，使得 SMOTE 自身取样类别重叠的问题大大减弱；同时，SMOTE 为 AdaBoost 提供“相对平衡化”的数据训练集，改善了在极度不平衡分布下，传统 AdaBoost 向多数类样本空间偏移的情况，使得 SMOTEBoost 算法呈现出整体优秀的性能。

3.3.2 AdaCost 算法

AdaCost 也是一种集成 AdaBoost 的代价敏感学习算法[18]。它利用代价矩阵，作为 AdaBoost 迭代框架中每一步权重调整的依据，同样修正了不平衡分类问题中，传统 AdaBoost 向多数类样本空间偏移的问题。研究表明，这种通过引入代价敏感来调整 AdaBoost 样本权重的方式，有效地改善 AdaBoost 性能，在不平衡分类问题的研究中起到了重要的作用。

3.4 其他重要算法

长期发展中，学者们还提出了许多独特而有效的不平衡学习算法[4]。如基于 kernel 的处理方法，可以在机器学习领域热门的 SVM 分类器基础上进行模型修改，包括加入 sampling 方法和直接修改 SVM 核函数等，借此改善 SVM 在不平衡分类问题中分隔超平面向多数类偏移的现象，提高模型性能。除了基于 kernel 的众多算法，还有主动学习、早期非常流行的 one-class 等大量优秀的算法。由于篇幅限制，对这些不平衡学习算法不再一一展开讨论。

另外，值得一提的是，尽管目前业界一般关注二元分类的不平衡学习问题，事实上多元分类的不平衡处理也至关重要。典型的如 AdaC2.M1[16]，与二元分类类似，它是一种利用 AdaBoost 的代价敏感学习算法，在实践中取得了高效的表现。除此之外，也有许多算法可以高效地处理多元不平衡分类问题，如基于代价敏感的神经网络以及一些集成算法。

4 不平衡学习算法的评估指标

随着不平衡学习算法的日益发展，如何正确有效地评估这些算法，成为制约该方向进一步发展的关键问题。如第 2 节所述，分类错误率作为传统分类学习算法的衡量标准，无法适用不平衡分类问题，因此必须制定有效的评估指标。

4.1.1 几个典型的评估指标

考虑二元分类问题，以少数类为 P(positive)，多数类为 N(negative)，那么根据分类器的预测类别和实际类别，存在如下混合矩阵(confusion matrix)，如表 1。

		实际类别	
		P	N
预测类别	P	TP (True Positives)	FP (False Positives)
	N	FN (False Negatives)	TN (TrueNegatives)

表 1 二元分类的混合矩阵

根据上述矩阵，传统的衡量标准$Accuracy$以及$ErrorRate$分别有如下定义，显然，二者都对数据集上类别分布的变化敏感：

$$Accuracy = \frac{TP+TN}{P+N};\quad ErrorRate = 1 - Accuracy \tag{3}$$

为了有效评估不平衡分类问题算法的效率，提出了如下几个新的指标[19]：

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F - Measure = \frac{(1+\beta)^2 \cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision} \tag{6}$$

$$G - mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{7}$$

Precision 表征分类预测的精确程度—在所有预测为 positive 的样本中，正确分类的比例；Recall 表征分类算法对 positive 类预测的完整性—啊所有实际 positive 的样本中，被正确预测的比例。Precision 对类别分布

的变化是敏感的，而 Recall 反之。根据上式，不平衡分类问题的目标简化为：在不影响 Precision 的前提下，尽可能提高 Recall。由于 Precision 与 Recall 的反向关系，实现上述问题显然需要一定权衡。

基于分析，F-Measure 是一种典型的权衡函数，它对数据集上类别分布的变化仍然敏感，其中参数β调整 Precision 和 Recall 在不同应用背景下的比重，整体描述了分类算法的表现。另一种权衡函数是 G-mean，同样可以刻画算法性能，且对类别分布的变化不敏感。

尽管上述的几个评估指标，与$Accuracy$和$ErrorRate$比较，描述能力有巨大的提高，但是它们只能刻画某一个分类器在某一种特定类别分布下的性能，能力仍然是有限的。

4.1.2 ROC(Receiver Operating Characteristics) curve 图

ROC 根据混乱矩阵，定义了TP_rate 与FP_rate

$$TP\_rate = \frac{TP}{P}$$
$$FP\_rate = \frac{FP}{N} \tag{8}$$

在 ROC 图中，FP_rate作横坐标，TP_rate为纵坐标，每一个点(FP_rate, TP_rate)都代表单一分类器在给定的类别分布数据集上的分类表现[20]。

当分类算法的输出是离散的类别值时，每个分类器会产生一对(TP_rate, FP_rate)，反映为 ROC 图上的一个点，如图 2 中各点。其中，A(0,1)代表最理想的分类器表现，分类器某次分类性能的优劣与其在 ROC 中对应点距 A 点远近成正比；除此之外，分布在对角线上的点，如 E，代表随机预测的分类器；而分布于 ROC 图右下方的点所代表分类器性能比随机预测更差，如 F；对其预测结果取反(G)所得对称的分类器却反而具有较好的分类预测性能。当分类器的输出一系列连续数值(多对(FP_rate, TP_rate))时，在 ROC 图中形成一条 ROC 曲线，如图 2 中 L1 与 L2。这种情况下，引入新的评估参数：$arear\ under\ the\ curve$(AUC)，即当前曲线右下方面积，来描述分类器的性能，AUC 越大性能越优。

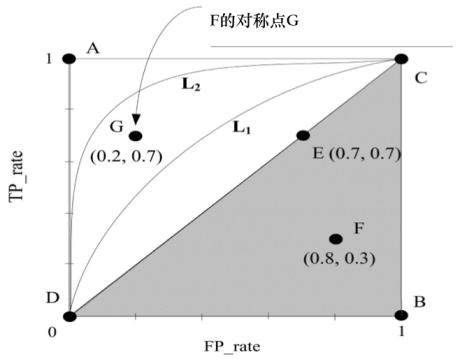

图 2 典型的 ROC curve 图

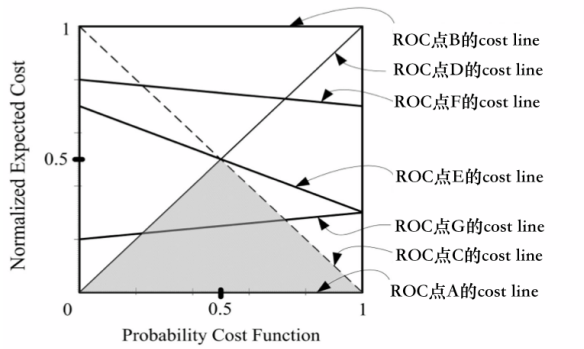

图 3 典型的 cost curve 图

4.1.3 PR(Precision-Recall) curve 图

不可否认，ROC 将分类器性能高效浓缩地整合于图像中，然而研究表明它在评估高度不平衡的数据集时，对分类算法性能的评估过优，而这一问题可以通过信息含量更大的 Precision-Recall(PR)曲线解决[19]。

PR 曲线以 recall 为横坐标，precision 为纵坐标，与 ROC 有极强的关联。PR 图几乎包括了 ROC 图所能表现的所有特征，而且较之于 ROC 提供了更丰富的信息，也更加适应不平衡分类问题背景。特别地，当一个分类器在不同类分布条件下的 FP 有巨大变化时，由于 N 数值较大，ROC 的FP_rate几乎不会有显著改变，也就无法捕捉这一种变化情况；相反，由于 PR 的 precision 考虑$TP + FP$，所 FP 的变化很容易从 PR 的 precision 展现出来。

由此可见，在高度不平衡分类问题中，PR 比 ROC 更能适应实际情况，被研究者们广泛采用为评估与分析的重要方式。

4.1.4 cost curve 图

ROC curve 功能强大，但是它无法蕴含类别误分的代价信息。于是，在 ROC 的基础上，学者提出了 cost curve，这是一个基于代价敏感的评估指标，能够在 ROC 的基础上有效囊括误分代价的信息，有着深远的理论与应用意义[21]。

总的来说，cost curve 以归一化代价(Normalized Expected Cost)为纵坐标，代价概率函数(Probability Cost Function)为横坐标，其中后者是正确分类正例概率的函数。Cost curve 与 ROC 有密切的联系，ROC 上对应的每一个点(FP_rate, TP_rate)都对应与 cost curve 中的一条线，如图 3：

$$E[C] = (1 - TP - FP) \times PCF(+) + FP \tag{9}$$

较之于 ROC 图，cost curve 图更加明确清晰，而且它还能呈现分类器在不同代价下表现性能。

5 潜在挑战和发展方向

现实应用面临着大量不平衡的分类问题，迫切的需求激发了对不平衡分类问题的研究和分析。随着新的不平衡学习算法不断被提出，这一领域的研究工作愈发成熟。但是，这并不意味着对不平衡分类问题的探索和研究已经到了尽头，相反，从体系化的角度来看，仍有大量的挑战和问题存在。

首先，目前大量处理不平衡分类问题的算法都是基于某一种特定的应用的背景，在这种情况下，虽然现行的分类器在表现上取得了巨大的改善，但却仍然难以从本质上理论地分析这种“改善”的程度，换言之，在现在的研究中缺乏足以分析不平衡分布下各个分类器的理论基础[4]。缺乏对问题本质的理论研究，使得许多至关重要的问题难以被回答，例如，采样方法的目的是平衡数据集，那么什么样的“平衡”才最利于分类学习呢？对不同类不平衡分布的情况，错误率是否存在一个界？这些问题严重阻碍了不平衡学习的体系化，使得对不平衡分类问题的研究缺少了有力的理论分析工具。

其次，缺乏对不同不平衡分类算法的统一评估标准。尽管上一节提出了几种分析不平衡分类算法的有效的工具与手段，但由于不平衡问题的特殊性，至今业界还达成一致，也即标准的评估手段，来有效地分析某一种算法在不同类分布下的性能，以及不同算法之间的对比。

6 总结

本文详细讨论了不平衡分类问题的产生与性质，并简单分析传统分类器在不平衡问题中的缺陷。然后，粗线条地介绍学术界处理该问题的几种重要思路和方向，对每一种思路，试图以一至两种具体的算法进行深入地学习与分析，以期对领域内流行的各种方法有一整体的认识和理解。最后，讨论能够适应不平衡分类问题的性能评估指标，完善对不平衡分类问题处理的整体流程。

致谢 感谢我的舍友对我的支持与耐心，容忍我深夜敲键盘改论文！

References:

[1] Weis G M, Provost F. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction[J]. Journal of Artificial Intelligence Research, 2003, 19:315-354.

[2] ZAWA K J, SINGH M, NORTON S W. Learning goal oriented Bayesian networks for telecommunications management[C]. Proc of the 13th International Conference on Machine Learning. San Fransis- co: Morgan Kaufmann, 1996: 139-147.

[3] M. Kubat, R.C. Holte, and S. Matwin. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning, vol. 30, no. 2/3, pp. 195-215, 1998.

[4] Haibo He, Edwardo A. Garcia. Learning From Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, september 2009.

[5] N.V. Chawla, N. Japkowicz, and A. Kolcz. Editorial: Special Issue on Learning from Imbalanced Data Sets. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 1-6, 2004.

[6] L.Breiman, J.Friedman, R.A.Olshen, C.J.Stone. Classification and Regression Tree. Boca Raton. FL: CRC Press, 1984.

[7] A. Estabrooks, T. Jo, and N. Japkowicz. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence, vol.20, pp. 18-36, 2004.

[8] R.C. Holte, L. Acker, and B.W. Porter. Concept Learning and the Problem of Small Disjuncts. Proc. Int'l J. Conf. Artificial Intelligence, pp. 813-818, 1989.

[9] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique[J]. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[10] H. Han, W.Y. Wang, and B.H. Mao. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Proc. Int'l Conf. Intelligent Computing, pp. 878-887, 2005.

[11] X.Y. Liu, J. Wu, and Z.H. Zhou. Exploratory Under Sampling for Class Imbalance Learning. Proc. Int'l Conf. Data Mining, pp. 965- 969, 2006.

[12] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20-29, 2004.

[13] T. Jo and N. Japkowicz. Class Imbalances versus Small Disjuncts. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 40-49, 2004.

[14] X.Y. Liu, Z.H. Zhou. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, vol: 18 (1) 2006, Page: 63 -77.

[15] C. Elkan. The Foundations of Cost-Sensitive Learning. Proc. Int'l Joint Conf. Artificial Intelligence, pp. 973-978, 2001.

[16] Mikel Galar, Alberto Fernaʹndez. Edurne Barrenechea, Humberto Bustince. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, And Cybernetics—PART C: Applications and Reviews, vol. 42, no. 4, July 2012.

[17] N.V. Chawla, A. Lazarevic, L.O. Hall, and K.W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 107-119, 2003.

[18] Wei Fan, Salvatore J. Stolfo, Junxin Zhang, Philip K. Chan. AdaCost: Misclassification Cost-sensitive Boosting. ICML, 2009.

[19] J. Davis and M. Goadrich. The Relationship between Precision- Recall and ROC Curves. Proc. Int'l Conf. Machine Learning, pp. 233-240, 2006.

[20] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical Report HPL-2003-4, HP Labs, 2003.

[21] R.C. Holte, C. Drummond. Cost Curves: An Improved Method for Visualizing Classifier Performance. Machine Learning, vol. 65, no. 1, pp. 95-130, 2006.