《模式识别与机器学习》课程实践：类别不平衡问题

截止日期：2018 年 7 月 10 日

1 问题描述

在现实的分类任务中，常常会面对**类别不平衡**（class imbalance）问题。类别不平衡问题的定义是：某些类别的数据量小于甚至远小于其他类别的数据量，类别的分布出现不均衡的状态。例如在欺诈交易识别问题中，绝大部分的交易是正常的，欺诈交易属于一小部分，属于“欺诈交易”类别和“正常交易”类别的数据就会出现不平衡。

不失一般性，考虑二分类问题，我们将比例大的类称为“大类”（majority class），将比例小的类称为“小类”（minority class）。在类别不平衡的情况下，大类和小类的比例会达到 10:1 甚至更高。当以准确度（Accuracy）为目标在类别不平衡数据上训练分类器时，分类器会更多地偏向于大类，更容易倾向于将样本分类为大类。如在大类和小类的比例为 10:1 的数据集上，分类器将所有样本分为大类，就可以达到 90% 的准确度，但是此时分类器对小类毫无分辨能力。在某些情景下（如欺诈检测），将小类准确分类是任务的主要目标，此时类别不平衡将会导致分类器模型的失效。

为了解决类别不平衡问题，许多不同的方法被提出，例如基于采样的方法，基于代价函数的方法，以及集成学习的方法 [1]。试采用或提出一种或几种方法，探究这些方法在类别不平衡数据上的效果。

2 数据集

提供三个数据集来训练模型以及测试效果，三个数据集分别是`car`、`yeast`、`wisconsin`，均为二类别数据集。三个数据集中，均存在不同程度的类别不平衡状况，数据集的基本情况见表1，

数据集	样本数	小类占比/%
car	1728	3.76
yeast	514	9.92
wisconsin	683	34.97

表 1: 数据集基本情况

数据集见附件中的`dataset` 文件夹，每个数据集目录下有四个文件，其中`*.txt` 和`*.pdf` 文件是对数据集的描述，`*.zip` 是全部数据，`*-5-fold.zip` 是按 5 折交叉验证划分好的数据。在训练及测试时，推荐使用 5 折交叉验证数据集。全部数据集都来自于 KEEL[1]。

3 评测指标

在类别不平衡问题中，我们遵循前人的工作，使用 F-measure、G-mean 以及 AUC（Area Under the ROC Curve) 作为模型的评测指标 [2]。

[1] `http://sci2s.ugr.es/keel/imbalanced.php?order=ir#sub20`

给定二分类问题的混淆矩阵（confusion matrix）如图1所示，F-measure、G-mean 评测指标按如下

		True class: Positive	True class: Negative
Prediction class	Positive	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

图 1: 混淆矩阵

公式计算，

$$
\begin{aligned}
&\text{FPR (False Positive Rate)} = \frac{FP}{FP+TN} \\
&\text{TPR (True Positive Rate)} = Acc_+ = \frac{TP}{TP+FN} \\
&\text{TNR (True Negative Rate)} = Acc_- = \frac{TN}{TN+FP} \\
&\text{G-mean} = \sqrt{Acc_+ \times Acc_-} \\
&\text{Precision} = \frac{TP}{TP+FP} \\
&\text{Recall} = \frac{TP}{TP+FN} = Acc_+ \\
&\text{F-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}
\end{aligned}
\tag{1}
$$

AUC 指标的定义是 ROC 曲线下面积，ROC 曲线描述了改变决策阈值（decision threshold）时模型在测试集上 TPR 和 FPR 的关系，将所有决策阈值下的（FPR，TPR）画在一条曲线上，就是 ROC 曲线，ROC 曲线下的面积就是 AUC 值，如图2。

评测时，在每个数据集上都计算相应的 F-measure、G-mean 以及 AUC，并将不同模型的评测指标进行比较。

4 其他说明

- 至少应包含一个 baseline 模型，baseline 为不使用任何处理类别不平衡方法的模型（如 logistc 回归、支持向量机、神经网络等）。并在 baseline 模型的基础上，尝试不同方法缓解类别不平衡问题。
- 本次课程实践结果以论文形式提交，内容至少应包括引言、方法介绍、实验结果分析、结论等部分，论文长度不应少于 4 页，建议不多于 10 页。课程实践占总成绩的 50%。
- 请在论文的第 1、2 页附上封面和**带本人签名的诚信声明**。并在截止日期之前将论文纸质版交至计算机楼 413 助教柯震处，每迟交 24 小时，扣论文成绩的 20%，扣完为止。
- 截止日期期间不在校的同学，请在离校前将封面和**带签名的诚信声明**提前交给助教，并于截止日期前在 elearning 上提交论文电子版，格式为 pdf，命名为`学号_姓名.pdf`。直接提交纸质版的同学，无需提交电子版。
- 如有其余事宜或疑问，请咨询助教zke17@fudan.edu.cn。

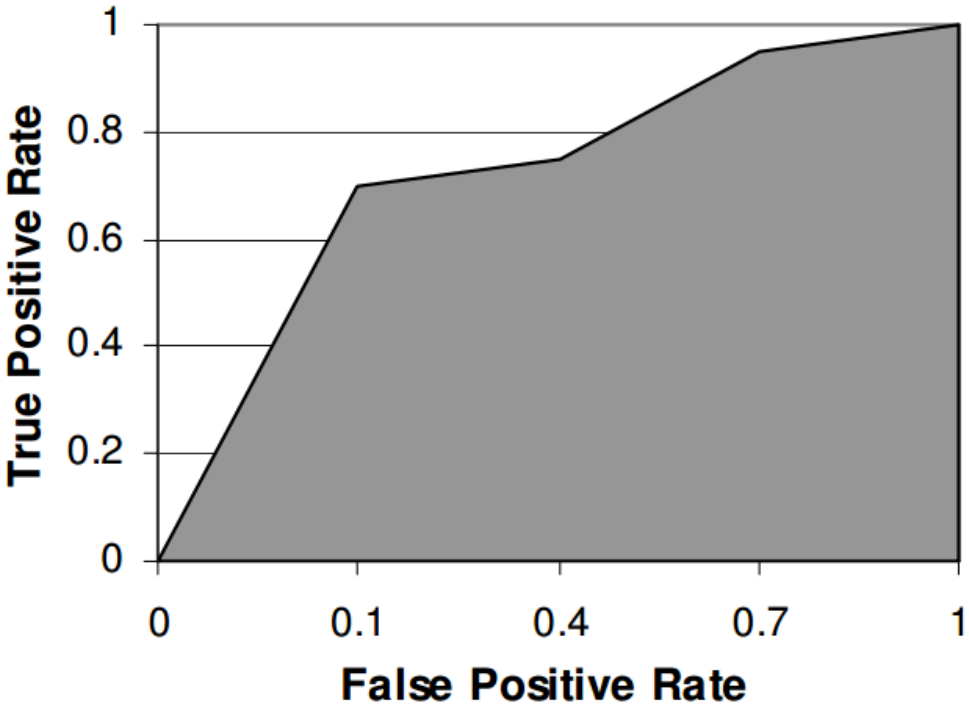

图 2: ROC 曲线

参考文献

[1] Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering.

[2] X. Y. Liu, J. Wu and Z. H. Zhou. 2009. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).