復旦大學

本科生课程论文

课程名称：	模式识别与机器学习	课程代码：	COMP130137.01
姓　　名：	熊中正	学　　号：	15302010038
学　　院：	软件学院	专　　业：	软件工程

复旦大学计算机科学与技术学院
2017～2018 学年第二学期期末考试试卷

■A 卷　　□B 卷　　□C 卷

课程名称： 模式识别与机器学习 **课程代码：** COMP130137.01

开课院系： 计算机科学与技术学院 **考试形式：** 课程论文

姓　名： 熊中正 **学　号：** 15302010038 **专　业：** 软件工程

声明：我已知悉学校对于考试纪律的严肃规定，将秉持诚实守信宗旨，严守考试纪律，不作弊，不剽窃；若有违反学校考试纪律的行为，自愿接受学校严肃处理。

学生（签名）：熊中正

2018 年　07 月　08 日

摘要

实际的分类问题往往都是不平衡分类问题，即样本数据分布不均衡，采用传统的分类方法，通常不能得到满意的分类效果。针对这一问题，许多新的机器学习算法被提出。本文介绍了当前常见的几种解决方案，并通过在真实数据集上的实验，探究比较了采样方法，集成学习方法，代价敏感学习方法三种方法的实际效果，并对实验结果作了一定的分析。

关键词：机器学习；不平衡数据分类；采样；集成学习；代价敏感学习

类别不平衡问题探究

一．引言

在机器学习中，类别不平衡问题指的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。例如在医院病例数据中，大部分的数据都是健康类型的数据，只有少部分数据对应的是疾病类型。数据在健康类别和疾病类别之间便产生了不平衡。传统的机器学习算法在处理类别不平衡问题时会遇到一些问题。传统的算法通常会偏向数量占优的类（Major Class），因为他们的损失函数会试图最优化相关数量，而并未将数据分布纳入考虑范围。为了解决类别不平衡问题，许多新的机器学习方法被提出。本文将介绍常见的三种处理类别不平衡问题的方法包括基于采样的方法，集成学习方法，以及代价敏感学习方法，并通过在给定数据集上的实验结果，具体分析这三种方法在处理相关问题的实际效果。本文的后续部分组织如下：第二部分是方法介绍，第三部分是实验所采用的评价标准和数据集介绍，第四部分是不同方法的实验结果比较分析，第五部分是结论。

二．方法介绍

在处理类别不平衡问题中通常有三种做法：采样、集成方法和代价敏感学习方法。

2.1 采样方法

采样方法是从数据集的角度解决不平衡问题。采样包括过采样（over-sampling）和欠采样（under-sampling）, 过采样是对数据量较少的一类数据（minor class）进行采样，增加其数据量，欠采样则是对主类进行采样，减少主类的数据量。采样方法旨在通过使数据的分布达到平衡来保证最后的训练效果。在本文中，关于过采样方法，我们探究了Random Sampling，SMOTE Sampling两种过采样方法的效果，关于欠采样方法，我们探究了ENN，CNN等多种欠采样方法在给定数据集上的实验效果。此外，我们还对欠采样和过采样相结合的采样方法进行了一定的实验。

2.1.1 过采样方法

Random Oversampling

该方法是最简单的过采样方法，它通过随机复制少数类的样本来达到增加少数类样本的目的。由于这种方法仅仅是对少数样本的简单复制，容易造成分类器的过拟合。

SMOTE Oversampling[1]

该方法建立在这样的假设之上：相距较近的正例之间的样本仍是正例。其主要思想是在相距较近的正例之间插入人造的正例。具体算法如下：对少数类的每一个样本X, 搜索其k个最近邻；然后随机选取这k个最近邻中的一个设为$\tilde{X}$，再在X与$\tilde{X}$之间进行随机线性插值, 构造出新的少数 类样本，即新样本$X_{new} = X + rand(0,1) * (X - \tilde{X})$，其中$rand(0,1)$表示区间$(0,1)$的一个随机数。若需要更多的虚拟样本，重复以上步骤即可。

2.1.2 欠采样方法

Random Under-sampling

随机欠采样方法是指在多数类中随机选取一些数据，使得在选取后的多类的数据量和少数类的数据量达到一个平衡。随机欠采样的方法实现简单，训练过程也比较快，但是这种做法可能会导致许多有用信息丢失，对最后的训练效果有所影响。

Edited Nearest Neighbor(ENN)[2]

在ENN方法中，如果Major Class中的一个样本不能被其K个最近邻使用KNN算法正确标识，那个这个样本将被移除。对Major Class中的每个样本执行这样的过程即可得到最后欠采样后的数据集。

Condensed Nearest Neighbor(CNN)[3]

CNN算法的核心思想是如果样本不能被某一集合正确分类，则将该样本加入该集合中。它的过程如下：首先将所有Minor Class的样本以及随机选取一个Major Class样本加入$\tilde{E}$中，将Major Class中剩下的样本加入E中。然后用$\tilde{E}$中的样本使用最近邻算法对E中的每个样本分类，将所有错分的样本加入到$\tilde{E}$中。最后的$\tilde{E}$作为采样后的集合。

2.1.3 过采样与欠采样相结合

SMOTE + ENN[4]

该方法同时结合了过采样和欠采样方法，它首先使用SMOTE方法对样本进行过采样，然后使用ENN方法对过采样后的数据集进行欠采样，两步采样后的数据集作为最终的数据集。

2.2 集成方法

集成方法是同时训练多个分类器，结合多个分类器的预测给出最后的评价。集成方法通常是和欠采样方法相结合。在欠采样方法中，通过减少主类的数据量使得主类和次类的数据达到均衡。数据的分布虽然达到了均衡，但是数据量缺可能减少很多，这就可能导致许多有用信息的丢失。使用欠采样后的数据集训练出的分类器效果可能较差。使用集成方法，结合多个欠采样方法训练出的弱分类器，可以在一定程度上实现较好的分类效果。此处，本文主要探究了EasyEnsemble[5]的效果，该算法的过程如下：

Algorithm 1 The `EasyEnsemble` algorithm.

{Input: A set of minority class examples $\mathcal{P}$, a set of majority class examples $\mathcal{N}$, $|\mathcal{P}| < |\mathcal{N}|$, the number of subsets T to sample from $\mathcal{N}$, and s_i, the number of iterations to train an AdaBoost ensemble H_i}
$i \Leftarrow 0$
repeat
  $i \Leftarrow i + 1$
  Randomly sample a subset $\mathcal{N}_i$ from $\mathcal{N}$, $|\mathcal{N}_i| = |\mathcal{P}|$.
  Learn H_i using $\mathcal{P}$ and $\mathcal{N}_i$. H_i is an AdaBoost ensemble with s_i weak classifiers $h_{i,j}$ and corresponding weights $\alpha_{i,j}$. The ensemble's threshold is θ_i, i.e.
  $H_i(x) = \text{sgn}\left(\sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \theta_i\right)$.
until $i = T$
Output: An ensemble:
$H(x) = \text{sgn}\left(\sum_{i=1}^{T}\sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \sum_{i=1}^{T} \theta_i\right)$.

2.3 代价敏感学习

最后一种方法是代价敏感学习。该方法通过修改代价函数对不同类的敏感性来消除类别不平衡的影响。在大部分不平衡分类问题中,稀有类是分类的重点，在这种情况下，正确识别出稀有类的样本比识别大类的样本更有价值。反过来说，错分稀有类的样本需要付出更大的代价。代价敏感学习赋予各个类别不同的错分代价，它能很好地解决不平衡分类问题。以两类问题为例，假设正类是稀有类，并具有更高的错分代价，则分类器在训练时,会对错分正类样本做更大的惩罚，迫使最终分类器对正类样本有更高的识别率。

三．评价标准和数据集

传统的评价标准通常是以准确率来作为评价指标。这一评价指标在处理类别不平衡问题时是不充分的。例如对于一个只有少数正样本的数据集，分类器如果将所有的样本都判为负样本，最后得到的正确率依旧很高。为了更公正的评价分类器的效果，我们选取[5]中提出的F-measure，G-mean以及AUC作为模型的评测指标。

在实验中我们使用了来自keel的三个二类别类别不平衡数据集。数据集的具体参数，以及样本分布如下：

数据集	样本数	小类样本占比
car	1728	3.76
yeast	514	9.92
wisconsin	683	34.97

四．实验结果比较分析

在这一部分，我们通过实验具体探究了上述提到的方法在类别不平衡数据集上的实际效果。我们使用了scikit-learn和imbalanced-learn[6]两个机器学习工具包帮助我们完成了实验。在实验过程中，我们使用了5-fold交叉验证方法，并取最后的平均值作为实验结果。我们选取了logistic regression分类器作为我们的Baseline方法。不同数据集的实验结果如下：

数据集	方法	F-measure	G-mean	AUC
	Baseline	0.10	0.09	0.90
	Random Over-Sample	0.30	0.87	0.94
	SMOTE	0.35	0.89	0.95

car	Random Under-Sample	0.20	0.79	0.87
	ENN	0.20	0.30	0.91
	CNN	0.17	0.27	0.91
	SMOTE + ENN	0.40	0.89	0.96
	Easy Ensemble	0.18	0.82	0.95
	Cost Weight	0.29	0.87	0.93

数据集	方法	F-measure	G-mean	AUC
yeast	Baseline	0.25	0.35	0.90
	Random Over-Sample	0.68	0.87	0.94
	SMOTE	0.70	0.88	0.93
	Random Under-Sample	0.70	0.87	0.90
	ENN	0.43	0.46	0.92
	CNN	0.75	0.85	0.90
	SMOTE + ENN	0.66	0.86	0.93
	Easy Ensemble	0.57	0.85	0.92
	Cost Weight	0.65	0.87	0.92

数据集	方法	F-measure	G-mean	AUC

wisconsin	Baseline	0.93	0.94	0.91
	Random Over-Sample	0.95	0.96	0.96
	SMOTE	0.95	0.97	0.96
	Random Under-Sample	0.95	0.95	0.91
	ENN	0.96	0.94	0.96
	CNN	0.95	0.97	0.95
	SMOTE + ENN	0.96	0.96	0.96
	Easy Ensemble	0.95	0.95	0.95
	Cost Weight	0.95	0.96	0.96

从实验结果我们可以发现：

1. Wisconsin数据集相较其他两个数据集，基本所有的方法都呈现出了非常好的效果。Yeast数据集的效果次之。Car数据集的实验结果是三组数据集中最差的一组。这一现象的产生可以从数据集的分布中得到解释。三组数据集中，Wisconsin数据集的数据分布最为均匀，小类样本占比达到30%以上，而Car数据集的数据分布最不均衡，小类样本仅占3%左右。数据集样本分布越均衡，对应分类器的效果自然越理想。

2. 从单个数据集来看，baseline方法的AUC虽然不差，但是该方法的F-measure和G-mean指标相对于使用了特殊方法处理的结果还是存在一定的差距（yeast数据集除外）。不同的方法提升的效果主要体现在F-measure和G-mean两个指标的提升上。AUC这一指标并未有太大改变。之所以会出现这种情况，主要是因为baseline的分类器将为数不多的正样本几乎全部错分为了负类。因此在混淆矩阵中TP一项非常低，所以导致在计算F-mean和G-measure时，两者计算后的数值较低。

3. under-sample方法相对于over-sample或者其他方法效果提升不不是很明显，有时会有效果下降的情况发生。从不同数据集的实验结果来看，under-sample的方法在Car数据集上的效果最差。前文也提到过，欠采样的方法通常会导致有效信息的丢失。而在Car数据集中，小类样本比例仅占3%左右，欠采样后数据量减少明显，可以猜想采样后的有效信息的丢失也是较为严重的。因此会出现欠采样效果较差的现象。

4. 从三个数据集的总体情况来看，SMOTE+ENN相结合的方法相对于其他的方法有着较好的效果。从三个数据集的实验结果可以看出，SMOTE方法有着不错的提升效果。而ENN方法在过采样的基础上，通过欠采样技术进一步去除了过采样带来的一些噪声数据。两者结合取得的提升效果有一定的合理性。

五．结论

类别不平衡问题是机器学习中非常常见的一大类问题。针对类别不平衡问题已经有了许多解决方案。本文在真实数据集上探究了采样，集成学习，代价敏感学习等三种常见的处理类别不平衡问题的方法在实际问题中的效果。通过实验，我们发现，过采样和欠采样相结合的采样方法（SMOTE + ENN）取得了不错的效果，其他的方法的效果提升并不是十分明显。当然，这些方法在不同的数据集上的效果也所有不同。在选择哪一种方法时还需要考虑数据集本身的一些特征。

除了本文所探究的三类方法之外，还有许多别的解决类别不平衡问题的方案。如可以在算法层面进行改进，选择对类别不平衡不太敏感的 SVM 分类器，或者选择一些新算法，例如 2014 年 Goh 和 Rubin 提出的 Box Drawing[7]相关算法。总而言之，在处理实际问题时，应该根据数据集的数据分布以及不同处理方法的特点综合考虑来确定最终的解决方案，以达到最优效果。

参考文献

[1] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*16 (2002): 321-357.

[2] Wilson, Dennis L. "Asymptotic properties of nearest neighbor rules using edited data." *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972): 408-421.

[3] Hart, Peter. "The condensed nearest neighbor rule (Corresp.)." *IEEE transactions on information theory* 14.3 (1968): 515-516.

[4] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." *ACM SIGKDD explorations newsletter* 6.1 (2004): 20-29.

[5] Liu, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou. "Exploratory undersampling for class-imbalance learning." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2009): 539-550.

[6] Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning." *The Journal of Machine Learning Research* 18.1 (2017): 559-563.

[7] Goh, Siong Thye, and Cynthia Rudin. "Box drawings for learning with imbalanced data." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014