**一种基于聚类融合的不平衡数据欠抽样方法**

肖连杰1,2  郜梦蕊1,2 苏新宁1,2

（1.南京大学信息管理学院，南京 210023；

2.南京大学 江苏省数据工程与知识服务重点实验室，南京 210023）

**摘要:** 鉴于不平衡数据集分类准确性低的问题，提出基于聚类的不平衡数据欠抽样方法，利用python对uci不平衡数据集进行分类，该方法在查全率、查准率、f-measure、和auc结果的实验，证明了该方法的有效性。最后对文本进行分类。

**关键词:** 不平衡数据、聚类、分类、欠抽样、文本分类

**Imbalanced Data Under-Sampling Algorithm Based on Clustering Ensemble**

Xiao Lianjie1,2 and Su Xinning1,2

1. School of Information Management of Nanjing University, Nanjing 210023;

2.Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023)

**Abstract:**

**Keywords:** imbalance data; cluster; classification; under-sampling; text classification

**1 引言**

分类是机器学习中一个重要的研究方向。传统的分类算法能够很好的处理平衡数据集问题，并且将样本总体准确率作为评价标准。然而，在实践中数据的类分布往往是不平衡的，如医疗诊断[he]、软件缺陷预测[]、信用卡欺诈交易监测[]、从卫星图像检测石油泄漏等，在这些领域，我们更多的关心属于少数类的样本的分类结果。由于属于少数类的样本更容易被分类器错分，一旦错分将对分类器的准确率产生重要影响。以二分类问题为例，假如数据集中两类样本的样本数之比为1:99，如果以整体分类准确率为准则，则分类器更加倾向于样本数目更多的多数类，虽然计算得出的准确率达到了99%，但是这样的分类结果毫无意义。

传统分类算法通常假设训练数据集中的各类样本之间的分布是均衡的，因此设计的算法在平衡数据集上比在非平衡数据集上的分类性能好。但是，通常的数据集总是在数量和分布上是不均衡的，这种不平衡会影响分类器的性能。

传统分类器采用整体分类准确率和错误率为分类评价准则，但是实验证明这些分类并不是适用于不平衡分类。针对不平衡数据集分类，学者们提出适用于不平衡数据集的分类评价指标，常用的有G-mean，F-measure和ROC等。

**2 相关研究**

针对不平衡数据分类问题，国内外学者展开了大量研究工作，取得了丰硕的成果。目前，该领域的研究侧重点主要分为四个层面：基于数据预处理层面、基于算法改进层面、基于成本敏感层面和基于分类器融合层面。

基于数据预处理层面，主要对数据进行预处理进而改善数据集中各类样本之间数目的不平衡状态，最直接的办法是采用相应的数据采样技术：过抽样（Over-sampling）、欠抽样（Under-sampling）以及混合采样。

过抽样技术的目的是增加少数类样本数，使得增加后的少数类样本数量与多数类样本数量相同。Chawla[]提出的SMOTE方法最为经典，现有的许多过抽样方法都是在此基础上进行的优化。

欠抽样技术的目的是减少多数类样本数，使得减少后的多数类样本数量与少数类样本数量相同。

类样本数，从而降低或消除因类间样本数目不平衡对传统分类器性能的影响；

基于算法改进层面，算法的参数和结构会影响不平衡学习的性能，通过改进传统算法（与代价敏感学习相结合方法、与智能优化算法相结合的方法、与采样方法相结合的方法、与集成学习相结合的方法、调整自身参数以适应不平衡学习的方法）或设计新算法（随机森林算法、Adaboost算法）使之适应不平衡数据集的学习。

基于成本敏感层面

基于分类器融合层面

在评价标准方面，目前的以全局准确率为分类器的评价标准不再适用于不平衡学习，学者们提出了一系列旨在提高正类识别率的评价体系。

对于少数类事件，往往需要决策者能够及时发现并作出反应。少数类事件包括软件检测、自然灾害预测、癌症基因的表达、信用卡欺诈交易以及电信诈骗等。

在数据挖掘领域，对事件的发现属于预测问题，即分类问题。

以uci机器学习数据集中的某些非平衡数据集为研究对象，在非平衡数据集上如何进行重采样，使得非平衡数据集转化为平衡数据集，进而提高分类器性能。利用f-measure和g-mean方法对多组二分类不平衡uci数据集分类器的性能进行评估。

Pca是一种将高维数据转化为低维数据的最优线性压缩方法。实验证明，在不平衡数据的分类中，有效的利用特征提取方法不仅可以提高分类器的性能[[1]](#footnote-1)，而且还可以降低分类算法的复杂性并且减少冗余信息[[2]](#footnote-2)。

1. 不平衡数据分类方法

表1 不平衡数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 来源 | 样本数 | 属性数 | 属性类别 | 不平衡率 |
| Ionosphere | UCI | 351 | 34 |  | 1.786 |
| Glass0 | keel | 214 | 9 | Real | 2.06 |
| Vehicle1 | keel | 846 | 18 | real | 2.9 |
| Glass1 | keel | 214 | 9 | Real | 1.82 |
| Glass6 | keel | 214 | 9 | real | 6.38 |
| Segment0 | keel | 2308 | 19 | real | 6.02 |
| Vehicle0 | keel | 846 | 18 | integer | 3.25 |
| New-thyroid1 | keel | 215 | 5 | real /integer | 5.14 |
| Ecoli2 | keel | 336 | 7 | real | 5.46 |
| Ecoli3 | keel | 336 | 7 | real | 8.6 |
| Yeast3 | keel | 1484 | 8 | real | 8.1 |
| Ecoli1 | keel | 336 | 7 | real | 3.36 |
| Yeast-0-5-6-7-9\_vs\_4 | keel | 528 | 8 | real | 9.35 |
| Glass4 | keel | 214 | 9 | real | 15.47 |
| Yeast4 | keel | 1484 | 8 | real | 28.10 |
| Yeast6 | keel | 1484 | 8 | real | 41.40 |
| Poker-8-9\_vs\_6 | keel | 1485 | 10 | integer | 58.40 |
| Kddcup-land\_vs\_satan | keel | 1610 | 41 | Real/nominal | 75.67 |
| Kr-vs-k-zero\_vs\_fifteen | keel | 2193 | 6 | nominal | 80.22 |
| Abalone19 | keel | 4174 | 8 | Real/nominal | 129.44 |

1. 不平衡数据的应用

应急管理方面

突发事件能够在短时间内对生命、财产或者环境带来巨大损害。针对突发事件的不确定性，突发事件预测显得有意义并且充满挑战。目前的突发事件侦查主要集中在自然灾害，。。。预测台风，

1. 展望

对不平衡数据的多分类问题进行研究；对指标权重进行优化；

**3 基于聚类融合的不平衡数据欠抽样方法**

本文采用两种聚类策略对多数类数据进行欠抽样。k-means作为基本的聚类方法，其应用广泛、计算代价较小，因此作为本文的基线聚类方法。

两种聚类策略描述如下：

1. 利用k-means对多数类数据进行聚类，将k值设置为少数类的样本数量（即k=n），k个聚类中心所构成的样本代替原有的多数类数据集。因此，整个数据样本由数量相等的两类数据集组成，且为平衡数据集。
2. 在策略1中得到k个聚类中心，由于k个聚类中心并不全是原有多数类数据集中的数据点，为了能够使用原有多数类数据集中的数据，将距离每个聚类中心点最近的样本数据点提取出来，代替聚类中心点。这样产生了与少数类样本数量相同的样本集合。
3. 在策略1得到K个聚类中心，应用信息增益准则，计算每个特征的信息增益，并进行归一化，确定每个特征的权重，利用加权欧式距离，求出距离每个聚类中心点最近的样本数据点，代替聚类中心点，产生与少数类样本数量相同的样本集合。

**4 实验分析**

**4.1 实验数据集**

不平衡数据集

训练集

测试集

多数类

少数类

平衡数据集

数据重划分

评价

**4.2实验平台**

本实验平台使用操作系统为64位Mac，内存8G，处理器2 GHz Intel Core i5，编程语言为python 3.4，开发环境为Anaconda 5.5。

**5 结语**

研究表明，对于传统标准分类器，一个平衡的训练数据集能够比不平衡数据集提供更高的预测准确率以及更好的推广性能。11，67

|  |  |  |
| --- | --- | --- |
|  | 预测正类 | 预测反类 |
| 正类 | TP | FN |
| 反类 | FP | TN |

TN是指原本是反类，预测仍然是反类。根据混淆矩阵，定义整体准确率（OA）、正确率（TPR）、查全率（Recall）、小类准确率（MIA）、大类准确率（MAA）、查准率（Precision）、错误率（FPR）等一系列评价指标。

分类器的整体准确率

分类器的查全率，也称为小类准确率或者正准确率

分类器的大类准确率

分类器的G-mean标准 68

从G-mean的定义得出，该标准兼顾了大类准确率和小类准确率，比整体准确率更适合应用于不平衡数据集的分类，因此常常被用作不平衡数据集分类的评价标准。

F-measure

F-measure也是面向不平衡数据集的一个评价标准，下面给出一些相关定义：

分类器的查准率，也是小类样本的查准率

分类器的大类错误率

分类器的F-measure 11

这里b是一个调节Precision和Recall的系数，通常情况取值1。F-measure通常被用于不平衡分类性能的评价。相比于G-mean，该标准更佳注重对小类分类性能的评价。

**参考文献**

He H, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on knowledge and data engineering, 2009, 21(9): 1263-1284.

[1]毛国君, 段立娟. 数据挖掘原理与算法（第3版）[M]. 清华大学出版社, 2016.

[2]袁梅宇. 数据挖掘与机器学习WEKA应用技术与实践（第2版）[M]. 清华大学出版社, 2016.

1. Last M, Kandel A, Maimon O. Information-theoretic algorithm for feature selection[M]. Elsevier Science Inc. 2001. [↑](#footnote-ref-1)
2. Blum A L, Langley P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence, 1997, 97(1–2):245-271. [↑](#footnote-ref-2)