计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：　基于熵的过采样框架
作者：　张念蓬，吴旭，朱强
网络首发日期：　2020-08-31
引用格式：　张念蓬，吴旭，朱强．基于熵的过采样框架．计算机工程与应用．
https://kns.cnki.net/kcms/detail/11.2127.TP.20200831.1516.010.html

网络首发时间：2020-08-31 16:11:48
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20200831.1516.010.html

基于熵的过采样框架

张念蓬，吴　旭，朱　强

西安电子科技大学 数学与统计学院，西安 710071

摘　要：数据挖掘与机器学习技术日益趋向成熟并且被广泛应用于实际问题的处理中，但该领域仍面临着诸多挑战，如不平衡数据集分类问题。利用过采样技术处理这类问题时，通常只考虑数量的不平衡，而不考虑数据分布是否平衡。本文利用信息熵度量数据集的局部密度信息，从分布上考虑数据集的不平衡程度，并提出了基于熵的危险集的概念和它的三种使用策略，即基于熵的危险集过采样算法、基于熵的安全集过采样算法和基于熵的自适应过采样算法。竞争性的实验结果表明，这些算法可以有效提升经典过采样算法的性能，为进一步利用信息熵理论研究不平衡数据集提供了成功的实践经验。

关键词：数据挖掘；不平衡数据；数据分类；数据分布；信息熵

文献标志码：A　　中图分类号：TP391　　doi：10.3778/j.issn.1002-8331.2005-0317

张念蓬，吴旭，朱强. 基于熵的过采样框架.计算机工程与应用

ZHANG Nianpeng, WU Xu, ZHU Qiang. Entropy-based Oversampling Framework. Computer Engineering and Applications

Entropy-based Oversampling Framework

ZHANG Nianpeng, WU Xu, ZHU Qiang

School of Mathematics and Statistics, Xidian University, Xi’an 710071, China

Abstract: Although data mining technology has gradually matured and is being dealt with in a wide range of practical problems, the field still faces many challenges, such as the problem of imbalanced datasets classification. When oversampling technology is used to deal with such problems, usually only the imbalance of the number is considered, not whether the data distribution is balanced. This paper uses information entropy to measure the local density information of the dataset, which considers the imbalance of the dataset from its distribution. In addition, we also proposes the concept of dangerous set and its three usage strategies, namely Entropy-based Dangerous set oversampling algorithm, Entropy-based safe set oversampling algorithm and Entropy-based adaptive oversampling algorithm. Experimental results show that these algorithms can effectively improve the performance of classic oversampling algorithms. For follow-up studies how to use entropy information theory processing imbanlanced data provides a successful experience.

Key words: Data mining; imbalanced dataset; data classification; data distribution; information entropy

基金项目：国家自然科学基金面上项目（No.61672025）。

作者简介：张念蓬(1994-)，男，硕士，主要研究领域为图论，数据挖掘，E-mail: 1558072113@qq.com；吴旭（1993-）男，硕士，主要研究领域为数据挖掘，机器学习；朱强（1976-）男，博士，教授，主要研究领域为图论、互连网络、多处理器系统的可靠性和故障诊断。

1 引言

数据挖掘是一种在海量数据中寻找即时的、有价值的信息的技术[1]。经过近些年的发展，数据挖掘已经形成了很多行之有效的模型和算法，它们主要集中在分类、聚类、关联分析等方面。其中，分类也被称为有监督学习，这类算法需要对数据的特征和类标签进行分析处理，得到不同的特征组合与类标签之间存在的判别规律，并将这些规律以知识的形式保存下来。当需要为新的数据判定类别时，分类算法能利用之前学得的知识为其贴上预测标签。

尽管数据挖掘与机器学习技术日益趋向成熟并且被广泛应用于实际问题的处理中，但该领域仍面临着诸多挑战，如不平衡数据集分类问题。顾名思义，不平衡数据集中至少有一类数据的数量明显多于或少于其他类的数据数量[2]。这类问题应用十分广泛，如 VIP 用户流失的检测[3]、欺诈交易识别[4]、医疗诊断[5]、银行破产预测和企业信用评估[6]等。

经典的机器学习算法和模型通常是基于“数据集是平衡的”这一假设建立的，若直接将它们应用在不平衡数据集上，性能会大幅下降。机器学习算法中的一个重要目标是最小化经验误差，即一个分类模型的目标是最小化总体分类误差，而少数类的分类结果对于总体来说影响是很小的。而且不平衡度越大，少数类对总体分类误差的影响越小。因此，分类器会通过主动保护多数类实例的方法来提升模型的整体性能，而忽视了对少数类实例的预测，甚至会将大量少数类实例误判为多数类。这样显然是不合理的。在不平衡数据的分类过程中，少数类数据的价值通常要大于多数类数据，而且随着不平衡度的增加，少数类数据的价值会越来越高。例如，在医疗诊断的过程中，将癌症患者误诊为健康的代价远高于将健康的人误诊为癌症患者的代价，该病人很可能会因此错过最佳的治疗时间，这带来的后果是非常可怕的。

2 不平衡数据的处理手段

用于提高不平衡数据集分类性能的技术整体上可以被分为两类：算法级方法和数据级方法。

算法级方法包括改进经典算法、代价敏感方法和分类器集成。修正分类算法以处理不平衡问题的策略是算法级技术[2]。代价敏感方法则是为不同的数据类型提供不同的错误分类代价。分类器集成是需要训练多个不同的弱分类模型，并按照特定的方式将弱分类模型组合起来，由它们的共同决策来预测数据的类标签，从而提高数据预测的准确性[7]。

数据级方法可以看作是一种独立于分类器的技术，用于重新平衡数据分布，使标准算法以用户的目标为中心[8]。特别地，数据级方法可以分为欠采样多数类实例[9]和过采样少数类实例[10]两种方案。欠采样方法通过减少多数类实例的数量来创建原始不平衡数据集的平衡子集。过采样方法通过增加少数类数据实例的个数来平衡数据集。Chawla 等人提出一种基于线性插值的过采样算法 SMOTE[11]。SMOTE 算法的主要思想是随机选取一些少数类实例作为种子，并选取种子的 k 个最近邻中的一个或多个少数类实例，与其结合为邻居对适应合成过采样方法(ADASYN)[10]、边界 SMOTE 算法(borderline)[12]、安全级 SMOTE

算法(safe)[13]等。

过采样技术通常是处理不平衡数据集的首选方法。传统的衡量类不平衡的指标是不平衡率 IR，即多数类数据的数量与少数类数据的数量之比。IR 反映了数据集在数量上的不平衡程度，但没有度量分布上的不平衡程度。即使数据集是数量平衡的，类分布的不平衡仍然可能存在[14]。此外，少数类集合的分类准确性与信息实例的数量有关，而与少数类实例的数量无关[15]。

因此，衡量少数类与多数类之间数据分布的不平衡程度是重要的。本文利用信息熵度量数据集的局部密度信息，从分布上考虑数据集的不平衡程度，并提出了基于熵的危险集的概念和它的三种使用策略，即基于熵的危险集过采样算法(EDgS)、基于熵的安全集过采样算法(ESS)和基于熵的自适应过采样算法(EAS)。基于熵的过采样框架具体分为三个部分，首先介绍了数据集熵差的具体计算方法和危险集的概念，这一部分是该框架的基础和起点；其次介绍了危险集的三种使用策略，分别是在危险集上过采样、在危险集的补集上过采样和自适应的过采样，这三种策略的侧重点不同，特点和优势也各不相同，适用于不同分布的数据集；最后，本文在算法中加入了生成实例的检测机制，若生成实例能通过检测，则该实例可以在数据分布的意义下平衡数据集，反之，该实例不具备平衡数据分布的能力，将其删掉即可。

3 信息熵的介绍

一个集合 D 的信息熵的计算公式如下：

$$E(D) = -\sum_{i=1}^{|c|} p_i \cdot \log_2 p_i \quad (1)$$

其中 p_i 通常为第 i 条数据的概率，本文用基于距离的局部密度在整体密度中的权重代替。众所周知，熵可以度量数据分布的不确定性。因此，本文利用熵差来度量数据集分布的不平衡程度，这与以往的 IR 完全不同。

在图 1 中，可以清楚地看到使用熵差(ED)的优点。这两个数据集具有不同的 ED 和相同的 IR。对于图 1(a)，两个类之间没有重叠区域，并具有清晰的分类边界，这使得任何一个简单的分类器都能很容易地完成识别；图 1 (b)则完全不同。显然，IR 无法区分这两个分布不同的数据集。总之，这些少数类的代表性实例是研究少数类分布的关键。以往的研究表明，固定 IR 时，少数类中的代表性实例越多，分类器的分类性能越好[14][15]。因此，用 IR 作为测量不平衡度的唯一指标是不合适的。

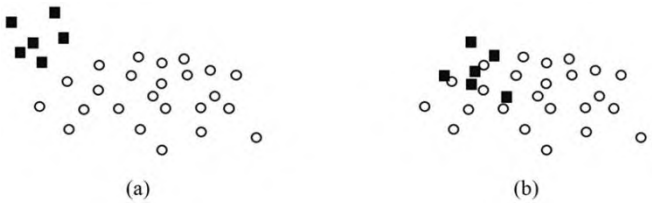

图 1　ED 相同、IR 不同的两个数据集

Fig.1　Two data sets with the same ED and different IR

熵通常用来度量数据分布的不确定性，它可以看作是信息分布的反义词。换句话说，数据分布的随机性越强，它包含的信息就越少[16]。对于不平衡数据来说，更分散的类内分布和更少的数据量将意味着更高的熵。在这种情况下，熵被引入到输入空间中作为数据分布的度量方式。

另外，本文基于信息熵将少数类数据集分为危险集和安全集。如果一个少数类实例属于危险集，则表示这个实例周围的少数类分布比较稀疏，在这些区域过采样，可以有效扩大数据集中少数类的范围，反之则表示实例周围的少数类分布比较密集，在这些区域过采样，会降低错分多数类实例的风险。

4 基于熵的过采样框架

本章节的主要内容是基于熵的过采样框架，具体可以分为以下三个部分。第一部分是数据集的熵的计算方法和计算过程中涉及到的统计量的含义，并在此基础上形成基于熵的危险集，讨论了危险集的意义。第二部分为危险集的使用策略和不同的使用策略所对应的含义，并给出不同策略对应的具体算法流程。第三部分通过实验验证了算法的有效性。

4.1 基于熵的危险集

这一小节介绍熵差的具体计算过程，并形成相应的算法流程。

给定一个训练数据集 D，包含实例 $X=\{x_i \mid x \in R^n, i=1,2,\cdots,m\}$，实例所属类别为 $C=\{c_l \mid l=1,2\}$，相应的实例数量表示为 m_1, m_2。数据集 D 中的任意两个实例表示为 $x_i=(x_{i1}, x_{i2}, \ldots, x_{in})$ 和 $x_j=(x_{j1}, x_{j2}, \ldots, x_{jn})$，这两个实例的距离计算公式通常定义为欧氏距离，如下：

$$d(x_i, x_j)=\sqrt{\sum_{k=1}^{n} | x_{ik}-x_{jk} |^2} \quad (2)$$

我们使用公式(3)为给定数据集的第 i 个实例定义一个基于密度的实例位置统计量：

$$\mu_k(x_i)=\frac{1}{k}\sum_{t=1}^{k} sim_t(x_i, Q_k(x_i)) \quad (3)$$

其中 $Q_k(x_i)$ 表示 x_i 的 k 近邻集合，$sim(\cdot,\cdot)$ 为相似度度量公式，通常使用欧氏距离。因此，$\mu_k(x_i)$ 是一个局部密度度量公式，用于测量 x_i 距离其 k 近邻的平均距离，同时表达了实例 x_i 附近的密度信息。第 i 个样本的基于密度的类位统计量由下式给出：

$$\omega_i=\frac{\mu_k(x_i)}{\sum_{x_j \in c_l} \mu_k(x_j)} \quad (4)$$

式中，ω_i 是 x_i 在 c_l 总密度度量中的比例。因此，每个实例的类内密度可以通过基于密度的类位统计来测量。x_i 附近的密度越高，$\mu_k(x_i)$ 和 ω_i 就越小。换句话说，ω_i 的大小反映了 x_i 的类内密度。

$$E_l=-\frac{1}{N_l}\sum_{x_i \in c_l} \omega_i \cdot \log_2 \omega_i \quad (5)$$

每一类的熵由公式(5)计算。令 c_1 和 c_2 分别代表少数类和多数类。容易得到 $E_1 \geq E_2 > 0$。众所周知，熵是由信息量的多少和信息对称性决定的。实验结果表明，在不平衡数据集上，多数类和少数类的熵的大小通常依赖于

信息量的多少。也就是说，少数类的类内熵通常大于多数类的类内熵。在此基础上，信息对称性影响类内熵的大小。为了度量数据集分布的不平衡程度，本文提出了一种新的度量方法：

$$ED = \theta = E_1 - E_2 \qquad (6)$$

另外，本文将少数类实例按 ω_i 排序，截取较大的一半，用来形成危险集 D_g 。由此将少数类数据集分为危险集和安全集。如果一个少数类实例属于危险集，则表示这个实例周围的少数类分布比较稀疏，在这些区域过采样，可以有效扩大数据集中少数类的范围，但是也会提高错分多数类实例的风险；反之则表示这个实例周围的少数类分布比较密集，在这些区域过采样，虽然生成的实例的多样性有所下降，但同时也会降低错分多数类实例的风险。

数据集的基于熵的危险集算法(EDg)的具体细节见表 1。

表 1　EDg 算法

Table 1　EDg algorithm

输入：数据集 D ，邻居数 k

输出：危险集 D_g ，熵差 ED ，多数类集合的熵 E_{maj} ，少数类集合 D_{min}

1. 将 D 按标签分为集合 D_1 和 D_2 ，初始化 $D_g = \varnothing$
2. 使用等式(2)计算 D 中所有实例 x_j 的 k 近邻并形成集合 $Q(x_j)$
3. 使用等式(3)计算 D 中所有实例 x_j 的实例位置统计量 $\mu_k(x_i)$
4. 使用等式(4)计算 D 中所有实例 x_j 的类位统计量 ω_j
5. 使用等式(5)计算 D_1 、 D_2 集合的熵 E_1 、 E_2
6. 计算熵差 $\theta = E_1 - E_2$
7. 如果 $\theta > 0$ ，则
8. 将 C_1 中所有的实例按 ω 的大小排序，并将后一半实例并入 D_g
9. 返回 D_g ， $ED = \theta$ ， $E_{maj} = E_2$ ， $D_{\min} = D_1$
10. 否则
11. 将 C_2 中所有的实例按 ω 的大小排序，并将后一半实例并入 D_g
12. 返回 D_g ， $ED = -\theta$ ， $E_{maj} = E_1$ ， $D_{\min} = D_2$

4.2 危险集的使用策略

EDg 算法为每个少数类实例计算出基于密度的类位统计量，也就是数据分布意义下的权重，权重越大，说明该实例周围的类内分布越稀疏。因此，本节提出三种基于熵的过采样策略，分别是在危险集上过采样、在安全集上过采样和自适应的过采样策略。这三种过采样的策略在合成过程中都采用线性插值的办法，只是在选取种子对时有所不同。

基于熵的危险集过采样算法(EDgS)首先利用 EDg 算法求出危险集；其次在危险集上随机的选择种子对，并使用公式(7)实现线性插值；

$$x_{new} = x_p + \left(x_q - x_p\right) \times \delta \quad (7)$$

其中 $\delta \in U[0,1]$，是均匀分布的随机数。最后检测整个数据集中 ED 的变化，若 $\Delta ED < 0$，则说明新实例在数据分布上平衡了数据集，是有价值的，应该保留。否则，删除新生成的实例。这样生成的新实例不仅可以在数据分布上平衡数据集，也可以有效扩大数据集中少数类的范围和多样性。表 2 详细描述了 EDgS 的实现过程。

基于熵的安全集过采样算法(ESS)首先利用 EDg 算法求出危险集，在 $C_{\min}$ 上求 D_g 的补集，得到安全集 D_s；其次在 D_s 上随机的选择种子对，并使用公式(7)实现线性插值；其余的步骤与 EDgS 算法相同。但相较于 EDgS 算法，ESS 算法生成的新实例的多样性会有所下降，错分多数类实例的风险也会显著降低。表 3 详细描述了 ESS 的实现过程。

表 2　EDgS 算法

Table 2　EDgS algorithm

输入：不平衡数据集 D，近邻数 k，算法 EDg

输出：过采样后的数据集 D

1. $D_g, \theta, E_{\text{maj}}, D_{\min} = EDg(D, \text{k})$
2. While $\theta > 0$：
3. 在 D_g 中随机选择实例 x_p，并在 $Q(x_p)$ 中随机选择 x_q
4. 将 x_p, x_q 带入式子(4-6)中生成新实例 x_{new}
5. 使用(3)(4)和(5)计算 $D_{\min} \cup \{x_{new}\}$ 的熵 e_1
6. 如果 $e_1 < E_{\min}$，则
7. $E_{\min} = e_1, D_{\min} = D_{\min} \cup \{x_{new}\}, D = D \cup \{x_{new}\}$
8. $\theta = E_{\min} - E_{maj}$

表 3　ESS 算法

Table 3　ESS algorithm

输入：不平衡数据集 D ，近邻数 k，算法 EDg

输出：过采样后的数据集 D

1. $D_g, \theta, E_{\mathrm{maj}}, D_{\mathrm{min}} = EDg(D, \mathrm{k})$
2. $D_s = D_{\mathrm{min}} - D_g$
3. While $\theta > 0$ ：
4. 在 D_s 中随机选择实例 x_p ，并在 $Q(x_p)$ 中随机选择 x_q
5. 将 x_p, x_q 带入式子(7)中生成新实例 x_{new}
6. 使用(3)(4)和(5)计算 $D_{\mathrm{min}} \cup \{x_{new}\}$ 的熵 e_1
7. 如果 $e_1 < E_{\mathrm{min}}$ ，则
8. $E_{\mathrm{min}} = e_1, D_{\mathrm{min}} = D_{\mathrm{min}} \cup \{x_{new}\}, D = D \cup \{x_{new}\}$
9. $\theta = E_{\mathrm{min}} - E_{maj}$

基于熵的自适应过采样算法(EAS)首先为每个少数类实例赋权，权重为 ω_i ；然后在考虑权重的基础上随机选择少数类实例 x_p ，在 $Q(x_p)$ 中随机选择 x_q ，并使用公式(7)实现线性插值；其余的步骤与 EDgS 算法相同。

与 EDgS 算法和 ESS 算法相比，EAS 算法可以有效增加生成的少数类数据的多样性，减小错分多数类实例的风险。表 4 详细描述了 EAS 的实现过程。

本节利用危险集的思想，给出了一个基于熵的过采样策略的框架，并在此框架下得到 EDgS、ESS 和 EAS 算法，这三个算法在理论上各有侧重。如 EDgS 在危险集上生成新实例，会显著增加少数类数据的多样性；ESS 在安全集上生成新实例，更加注重生成实例的安全性；EAS 则在整个少数类数据集上自适应的生成少数类，是前两种算法折中的结果。

4.3　实验结果及分析

为验证提出的算法的有效性，本章选取来自 UCI[17]和 KEEL-dataset repository[18]中的 6 个二分类数据集进行实验仿真，它们的详细介绍见表 5。每个数据集分别通过 7 种过采样算法(SMOTE，borderline，EDgS，safe，ESS，ADASYN，EAS)进行处理，且

选择 SVM 作为基分类器。评价指标选择 AUC 和召回率，因为 AUC 能客观地反映分类器对不平衡数据集的综合预测能力，召回率能反映出分类器对少数类实例的分类准确度。显然，AUC 和召回率的值越大，算法的性能就越好。

表 6 和表 7 分别列出了 8 个算法在 6 个数据集上的 AUC 和召回率的得分和排名的详细信息。

表 4　EAS 算法

Table 4　EAS algorithms

输入：不平衡数据集 D ，近邻数 k

输出：过采样后的数据集 D

1. 将 D 按标签分为集合 D_1 和 D_2
2. 使用(3)(4)和(5)计算 D 中所有实例的 ω_j ， D_1 、 D_2 集合的熵 E_1 、 E_2
3. 如果 $E_1 - E_2 > 0$ ，则
4. $\theta = E_1 - E_2$ ， $E_{maj} = E_2$ ， $E_{\min} = E_1$ ， $D_{\min} = D_1$
5. 否则
6. $\theta = E_2 - E_1$ ， $E_{maj} = E_1$ ， $E_{\min} = E_2$ ， $D_{\min} = D_2$
7. While $\theta > 0$ ：
8. 为 $D_{\min}$ 中的所有实例 x_i 赋权，权重为 ω_i
9. 在 $D_{\min}$ 中考虑权重地随机选择 x_p ，并在 $Q(x_p)$ 中随机选择 x_q
10. 将 x_p, x_q 带入式子(7)中生成新实例 x_{new}
11. 使用(3)(4)和(5)计算 $D_{\min} \cup \{x_{new}\}$ 的熵 e_1 和所有少数类实例的新权重 ω_{new}
12. 如果 $e_1 < E_{\min}$ ，则
13. $\omega_i = \omega_{i,new}(x_i \in D_{\min})$
14. $E_{\min} = e_1, D_{\min} = D_{\min} \cup \{x_{new}\}, \theta = E_{\min} - E_{maj}, D = D \cup \{x_{new}\}$

表 5　二分类数据集的描述信息

Table 5　Description information of binary classification dataset

编号	数据集	样本数	特征数	少数类数据数量	IR	ED
1	abalone17vs78910	2338	7	58	39.3	0.09
2	blocks0	5472	10	559	8.8	0.008
3	ecoli0vs1	220	7	77	1.9	0.03
4	german	1000	24	300	2.3	0.013
5	glass1	214	9	76	1.8	0.027
6	yeast0359vs78	506	8	50	9.1	0.092

表 6　8 个算法在 6 个数据集上的 AUC 得分和排名

Table 6　Shows the AUC scores and rankings of eight algorithms on six data sets

数据集	SVM	SMOTE	Borderline	EDgS	safe	ESS	ADASYN	EAS
abalone17vs78910	0.842(7)	0.915(5)	0.921(3)	0.930(2)	0.822(8)	0.846(6)	0.918(4)	**0.932(1)**
alocks0	0.985(4)	0.982(5)	0.979(6)	0.987(2)	0.977(7)	0.986(3)	0.972(8)	**0.990(1)**
ecoli0vs1	**0.996(3)**	**0.996(3)**	0.994(7.5)	0.995(6)	**0.996(3)**	0.994(7.5)	**0.996(3)**	**0.996(3)**
german	**0.784(1.5)**	0.769(3)	0.766(4.5)	0.776(4.5)	0.770(6)	**0.784(1.5)**	0.754(8)	0.761(7)
glass1	0.784(7)	0.785(6)	**0.815(1)**	0.802(4.5)	0.776(8)	0.809(2)	0.802(4.5)	0.807(3)
yeast0359vs78	0.700(7)	0.719(5)	0.741(4)	0.751(3)	**0.761(1)**	0.754(2)	0.686(8)	0.708(6)
平均排名	4.92(6)	4.50(5)	4.33(4)	3.67(2.5)	5.5(7)	3.67(2.5)	5.92(8)	**3.5(1)**

表 7　8 个算法在 6 个数据集上的召回率得分和排名

Table 7　shows the recall rate scores and rankings of 8 algorithms on 6 data sets

数据集	SVM	SMOTE	Borderline	EDgS	safe	ESS	ADASYN	EAS
abalone17vs78910	0.015(8)	0.718(7)	0.730(6)	0.892 (3)	0.875(5)	0.884(4)	0.901(2)	**0.905(1)**
alocks0	0.736(8)	0.946(5)	0.960(4)	0.977(2)	0.885(7)	0.904(6)	0.963(3)	**0.972(1)**
ecoli0vs1	0.961(7)	**0.976(1.5)**	0.974(4)	0.969(5)	0.959(8)	0.962(6)	**0.976(1.5)**	0.975(3)
german	0.389(8)	0.616(4)	0.631(2)	**0.645(1)**	0.505(7)	0.622(3)	0.564(6)	0.612(5)
glass1	0.442(8)	0.713(6.5)	0.793(3)	0.839(2)	0.713(6.5)	0.753(4)	0.733(5)	**0.844(1)**
yeast0359vs78	0.188(8)	0.573(5)	0.535(6)	0.647(4)	0.246(7)	0.688(2)	0.660(3)	**0.697(1)**
平均排名	7.83(8)	4.83(7)	4.17(4.5)	2.83(2)	6.75(6)	4.17(4.5)	3.42(3)	**2(1)**

对于基于线性插值的算法来说，borderline 和 EDgS 都是在危险集上进行过采样，safe 和 ESS 都是在安全集上进行过采样，ADASYN 和 EAS 都是在整个少数类数据集上进行自适应的过采样。因此，将上述算法两两之间进行对比是比较合理的。可以看出，本章提出的 EDgS、ESS 和 EAS 的 AUC 得分均强于 borderline、safe 和 ADASYN。特别是 EAS 算法，在对 ADASYN 算法进行提升的同时，也在多个数据集上取得了很好的名次，如数据集 abalone17vs78910、alocks0 和 ecoli0vs1。这体现了本文提出的算法在综合预测能力上的优势。

不平衡数据分类问题中少数类实例通常更加珍贵，因此少数类被正确分类的比例是很重要的。我们的算法在召回率得分上显示出非常强的竞争力。用于实验的 6 个数据集中，基于熵差的过采样算法只在 ecoli0vs1 上表现一般，这可能是由于该数据集的 ED 很小，而 IR 较大，相较于传统的不平衡度量方法，我们的算法不能很好的识别少数类和多数类。

5 总结与展望

本文利用熵信息来度量数据集的不平衡程度，为各种变量赋予实际意义，并给出用熵差计算数据分布不平衡度的具体方法；另外，利用熵信息计算出每个点周围的局部密度，得到了基于熵的危险集。随后给出了危险集的使用策略和对应的算法，即 EDgS、ESS 和 EAS 算法。实验证明，本文的研究内容可以有效提升经典过采样算法的性能。但不可否认的是，该理论和模型也存在一定的局限性，即对熵差较小的数据集的识别能力较差。针对这个问题，可以将 IR 和 ED 相结合，在利用 ED 检测数据分布的不平衡度的同时，使用 IR 来体现数据集数量上的不平衡度，从而进一步提高对数据集的综合识别能力。这也是我们接下来的研究内容和方向。

参考文献：

[1] Deng W H, Wang G Y, Xu J. Piecewise two-dimensional normal cloud representation for

time-series data mining[J]. Information Sciences, 2016, 374: 32-50.

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

[3] 汪明达，周俏丽，蔡东风.采用混合模型的电信领域用户流失预测[J].计算机工程与应用，2019, 55(24): 214-221.

Wang M D, Zhou Q L, Cai D F. User Churn Prediction in Telecom Domain Using Hybrid Model[J].Computer Engineering and Applications, 2019, 55(24): 214-221.

[4] 宋新平，丁永生，张革夫.集成分类法在财务欺诈风险识别中的应用[J].计算机工程与应用,2008, 44(34): 226-230.

Song X P, Ding Y S, Zhang G F.Application of integrated classification method in identifying risk of fraudulent financial report[J].Computer Engineering and Applications, 2008, 44(34): 226-230.

[5] 张涛.不平衡数据分类研究及在疾病诊断中的应用[J].黄河科技学院学报,2019, 021(5):15-22.

Zhang T. A study on imbalanced data classification and its application in disease diagnosis[J]. Journal of Huang He College, 2019, 021(5):15-22.

[6] 张涛，汪御寒，李凯等. 基于样本依赖代价矩阵的小微企业信用评估方法[J]. 同济大学学报:自然科学版, 2019, 48(1): 149-158 .

Zhang T, Wang Y H, Li K, et al. Credit Scoring of Small and Micro Enterprises Based on SampleDependent Cost Matrix[J]. Journal of Tong Ji University,2019, 48(1): 149-158.

[7] Galar M, Fernandez A, Barrenechea E, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, 42(4): 463-484.

[8] Yang J, Yu X, Xie Z Q, et al. A novel virtual sample generation method based on Gaussian distribution[J]. Knowledge-Based Systems, 2011, 24(6): 740-748.

[9] Lin W C, Tsai C F, Hu Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409: 17-26.

[10] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. 2008 IEEE international joint conference on neural networks, 2008: 1322-1328.

[11] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[12] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]. International conference on intelligent computing. Springer, Berlin, Heidelberg, 2005: 878-887.

[13] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: Safe-level-synthetic

minority over-sampling technique for handling the class imbalanced problem[C]. Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, 2009: 475-482.

[14] Japkowicz N, Stephen S. The class imbalance problem: A systematic study[J]. Intelligent data analysis, 2002, 6(5): 429-449.

[15] Tang B, He H. GIR-based ensemble sampling approaches for imbalanced learning[J]. Pattern Recognition, 2017, 71: 306-319.

[16] S. Kullback, N. York. Information theory and entropy[J]. Model Based Inference in the Life SciencesA Primer on Evidence, 2008: 51-82.

[17] Asuncion A, Newman D. UCI machine learning repository[J]. University of California, Irvine, School of Information and Computer Sciences, 2007, 9: 10-23.

[18] Alcalá-Fdez J, Fernández A, Luengo J, et al. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework[J]. Journal of Multiple-Valued Logic & Soft Computing, 2011, 17.