基于多样性数据生成和集成学习的两类非平衡大数据分类

1 课题来源及研究的目的和意义

1.1 课题来源

随着互联网和物联网的发展，数据正以前所未有的速度在增长，人类已经进入了大数据的时代。在这种环境下，研究人员可以基于这些数据进行统计分析，挖掘数据中蕴藏的有价值的信息。但在现实中，数据的质量很难得到保证，其中数据类别不平衡问题就是一种非常常见的情况。当用传统的分类方法去解决不平衡数据分类问题时，得到的结果就会有偏差。目前，对不平衡数据分类问题的研究是非常活跃的研究领域，根据Guo等[1]统计，在过去的10年内，总计192个期刊和会议共发表527篇关于非平衡问题的论文。这证明迄今为止，非平衡问题还是一个非常有价值的研究课题。

1.2 课题研究目的和意义

非平衡问题分为多类不平衡问题和两类不平衡问题[2]，本研究只考虑两类不平衡的情况，即数据中某一类别的样例数远远小于另一类别，并且把数量比较多的样本称为多数类样本或负类样本，数量较少的样本称为少数类样本或正类样本[3–5]。传统的分类方法，例如决策树、K近邻等方法，由于考虑算法在整个数据上的准确率，所以往往会忽视少数类样本。举个例子，如果数据集有10000个样本，而少数类样本只有100个，即使分类器把所有样本都分为多数类，也会达到99%的准确率。但是，对于少数类来说，这样的结果是不能接受的。这就会对我们的数据研究和分析造成一定的困难，令得出的结论出现偏差。因此，目前研究人员提出了许多方法解决二类不平衡问题。其中，对数据集进行重采样，增加少数类样例数目或者减少多数类样例数目，使数据集平衡化,就是一种非常有效的方法。Gracia等[6]分析了不同采样算法对分类性能的影响，实验表明上采样的结果往往比下采样要好，因为下采样可能会丢失一些重要的信息。然而大多数上采样方法，会导致原始数据集类别分布的改变，有可能导致分类器在训练过程中出现过拟合的情况。另外，在某些情况下，大多数上采样方法缺乏可解释性，并且效果有限。因此，研究一种可解释性强且有效的数据上采样方法是非常必要的。

2 国内外在该方向的研究现状及分析

数据类别不平衡问题仍然是目前研究的热点，主要是由于在现实生活中数据不平衡问题普遍存在，如医疗领域的肿瘤诊断，电商领域的恶意差评检测等。不同类别的样例数量相差较大，可能会严重影响标准分类模型的性能。根据数据不平衡程度，可以将数据不平衡问题分为以下几类：轻微不平衡（正负样本数量相差在一个数量级内），中度不平衡（正负样本数量相差在两个数量级以内）以及重度不平衡（正负样本数量相差超过两个数量级）。目前，对于数据对于类别不平衡问题，已有的文献中已经提出了不同的解决方法。主要分为以下三类[7]:数据层面、算法层面以及混合的方法。

数据层面，主要通过样本重采样技术对数据集进行预处理，主要是对多数类样例进行下采样、对少数类样例进行上采样，或者两者结合使用。从而达到不同类别之间样本数量比例的相对平衡。早在1998年，Ling等[8]就提出了随机上采样和随机下采样算法，其算法思想是目前最简单的样本采样技术。随后，由于随机上采样采取简单的复制样本策略来增加样本，这样容易产生模型过拟合问题。针对这一问题，研究者们提出了很多上采样方法。Chawla等[9]提出了少数类生成上采样方法SMOTE，该方法通过利用少数类样本及其邻域信息进行数据的上采样，扩充的样例和已有样例不相同，从而降低了过拟合的风险，但容易发生类间样本重叠的问题，并且引入了额外的噪声。HAN等[10]在SMOTE的基础上，提出了Borderline-Smote算法，它的主要思想是，在生成样本的过程中，如果少数类样本的k近邻中的全部样本为负类样本，则不进行样本生成，一定程度上解决了SMOTE算法类间样本

重叠（overlapping）的问题。研究人员基于SMOTE方法还提出了很多改进，比如MSMOTE[11]、B1-SMOTE和B2-SMOTE等[12]。参考文献[1, 13]对这些方法进行了很好的综述。

下采样是指减少多数类样本以达到数据集的平衡。近年来，对下采样方法也有大量研究成果。Chan等[14]提出了一种基于近邻的下采样方法ENN，该算法去除少数类样本最近的3个近邻样本中2个或者以上的类别不同的样本，但ENN算法去除的样本较少，不能很好的改善数据的平衡性。Kubat等[15]提出一种启发式下采样方法One-Sided Selection，用于去除在多数类边界线附近的样本或者噪声样本，该算法尽可能的留下具有代表性的样本，但是改变了原来的数据分布，可能会丢失一些重要信息。针对这一问题，Ha等[16]提出了一种基于遗传算法的下采样方法，能够保证数据分布信息的完整性。上采样和下采样方法，都是处理非平衡数据的重要手段，Estabrooks等[17]和Barandela等[18]建议同时使用上采样和下采样两种方法，这对于处理非平衡问题非常有效。

算法层面，主要是设计一个适用于非平衡数据的新分类算法，或对已有的分类算法进行改进，从而令这些算法可以处理非平衡数据。目前主要的方法有，对不同类别的样本赋予不同惩罚参数的代价敏感学习、集成方法以及主要用于异常样本检测中的单类学习等[7]。在这一层面，也有大量的研究成果。Breiman等[19]提出了一种基于bootstrap技术的集成方法，首先利用bootstrap从原始数据集中进行有放回抽样得到一系列新数据集，在这些数据上训练分类器，最后使用多数投票的方法或者加权投票的方法确定测试样本的类别，这种方式对于不稳定的系统可以提高其分类性能。对于传统的不平衡处理方法会改变原始数据分布的问题，Sun等[20]提出了一种带权重的集成方法，先将多数类样本划分成多个子集，每个子集都与少数类样本合并为一个平衡数据集，然后在这些平衡数据集上训练分类器，将各个分类器通过特定的集成规则组合，比较了不同集成规则对分类性能的影响。Li等[21]针对adaboost[22]方法在非平衡问题中，对于少数类样本性能无法保证的问题,提出一种基于KNN的改进方法K-Adaboost方法，使用KNN算法降低靠近少数类的多数类样本权重，令分类器更加重视少数类，提升adaboost在非平衡问题中的性能。参考文献[7]对目前比较常用的方法进行了很全面的综述。

混合的方法，结合数据层面和算法层面的方法，在一定程度上弥补了这两个层面方法的缺点，并且可以达到较好的分类性能。Chawla等[23]提出一种新的方法SMOTEBoost，在SMOTE的基础上结合Boosting[24]方法提高在非平衡数据集上的性能，与传统的Boosting技术赋予所有错误分类样本相同权重不同，SMOTEBoost先在少数类样本中合成人工样本，从而间接的改变了样本的更新权重，使分类器更加重视少数类样本。Rayhan等[25]提出了基于聚类的欠采样方法CUSBoost，首先把数据分为少数类样本和多数类样本，然后使用K-means算法对多数类样本进行聚类处理，并且从每个聚类中选择部分样本来组成平衡的数据。聚类方法帮助在多数类数据中选择了差异性更大的样本，实验证明该算法表现优于一般的集成学习算法。Cohena等[26]针对在医疗诊断中罕见阳性问题，结合了上采样和下采样两种方法，提出一种通过正则化参数调整SVM分类边界的算法。Yong等[27]基于k-means和遗传算法提出了适用于不平衡数据的采样方法，该方法首先将少数类样本通过k-means聚成k个簇，在每个簇中使用遗传算法生成新的样本。王等[28]在代价敏感的理论基础上，提出一种新的基于代价敏感集成学习的非平衡数据分类算法NIBoost，实验结果表明该方法在处理非平衡数据分类的问题上具有一定的优势。赵等[29]针对随机森林算法在非平衡数据集上表现的分类性能差的问题,提出一种新的过采样方法SCSMOTE，关键是找出少数类样本中合适的候选样本，基于这些候选样本生成新的样本，实现了对合成少数类样本质量的控制，有效提高了随机森林在非平衡数据集上的分类性能。

以上各个层面的方法中，数据层面的方法最为容易理解，该层面的方法通过数据重采样技术，达到平衡数据集的目的。最后得到的平衡数据集可以直接用于已有的分类模型，不需要对模型进行任何修改，具有很好的通用性。

3 主要研究内容及创新点

3.1 主要研究内容

本论文主要研究基于多样性数据生成和集成学习的两类非平衡大数据分类。

对于两类非平衡大数据分类的解决方案分为两个阶段，一是对少数类样本（正样本）进行上采样，增加少数类样本的数量，使其与多数类样本（负样本）达到一定的平衡比例。同时为了防止生成过多相似的少数类样本，提出一种度量生成样本多样的方法。然后将多数类样本划分成K个子集，各个子集分别与上采样的后的少数类样本构成K个相对平衡数据集，在这K个数据集上分别训练分类器，通过某种集成规则将这K个分类器集成。本论文拟研究的主要内容包括：

- 已有的样本采样方法仅仅适用于某些类型的数据集，而且生成的数据在一些情况下缺乏解释性。比如在图像数据上，通过SMOTE方法进行样本上采样，得到的新样本很可能是一张没有意义的图像，这些图像对我们训练分

类器价值不大。基于此种情况，本论文通过生成对抗网络或者变分自编码器，并设计相应的网络结构，训练生成模型用于生成有意义的少数类样本。

- 在对少数类样本进行上采样的过程中，已有的样本采样方法不能保证生成样本的多样性，导致生成的样本和已有的样本过于相似，不能增加有效的信息。为了保证生成样本的多样性，本论文拟采用一个指标（如类内散度最大化）来评价生成样本的多样性，并指导样本的生成。从而可以有效的扩充少数类样本，增加有效的数据信息。
- 为了保证不过多的对少数类样本上采样，导致数据的冗余，本论文将多数类样本划分为K个子集，令每个子集与上采样后的少数类样本构成一个相对平衡数据集，以减少上采样的数量。但是由于K的取值对于不同数据集来说可能不同，所以对K的选取需要讨论、实验，找到合适的取值。
- 如第3点所述，将多数样本划分为K个子集后，各个子集与上采样后的少数类样本构成一个相对平衡的数据集，但是我们无法知道当平衡比例达到多少时就可以有效的训练分类器。针对这一问题，本论文决定研究数据集在不同平衡比例下，分类器性能的变化，从而找到一个合适的平衡比例。
- 对于数据集类别不平衡比例问题，目前的研究较少。如果少数类样本过少，则无法有效的通过生成对抗网络或者变分自编码器训练生成器，本论文拟采用实验的方法，研究对于同一数据集，不平衡比例对分类性能的影响。

3.2 创新点

- 基于生成对抗网络或者变分自编码器对两类不平衡大数据进行数据生成。
- 通过集成的方法减少少数类样例上采样的数量，并提高分类器的性能。
- 提出一种评价生成样本多样性的指标，从而可以有效的增加数据信息，改善分类的效果。

4 研究方案及进度安排，预期达到的目标

4.1 研究方案

- 选择UCI数据库或者其它开放数据源中的数据集。
- 通过阅读大量的文献和资料，掌握生成对抗网络、变分自编码器以及集成算法的原理和实现方式。
- 学习开源深度学习框架Keras和TensorFlow，学习机器学习库sklearn， 并基于这些框架实现基于生成模型的非平衡分类算法。
- 实验分析，通过实验结果对所提出的方法进行分析，并与传统的上采样方法比较。
- 在导师的指导下，撰写毕业论文。

4.2 研究进度安排

- 2019年 04-05月：查阅相关文献资料，收集课题发展信息，并确定研究方法和目标。
- 2019年 06-09月：对提出的研究方法进行实验观察，分析其优缺点。
- 2019年 10-12月：收集材料，组织思想，撰写论文初稿。
- 2020年 01-03月：对论文初稿进行修改和整理，完成论文写作。
- 2020年 04-05月：准备毕业论文答辩。

4.3 预期达到的目标

设计出基于多样性数据生成和集成学习的两类非平衡大数据分类算法，并实现该算法。与已有的非平衡数据上采样方法进行充分的实验比较，分析所提方法的优点和不足。

5 为完成课题已具备和所需的条件

5.1 已具备的条件

1. 通过研究生课程的学习，已经对tensorflow和keras计算框架以及python编程有了比较深入的了解，为下面的研究工作奠定了实践基础。
2. 已经查找大量课题研究所需的文献材料，为理论证明做好准备。
3. 实验所需的环境已具备。

5.2 所需的条件

1. 选择或者设计合适的生成模型，从而提高生成样本的多样性和质量。
2. 基于生成模型的非平衡分类算法的实现，及其参数调优。
3. 与传统的上采样方法进行比较与分析。

6 预计研究过程中可能遇到的困难和问题以及解决的措施

遇到问题:

1. 生成模型结构的选择，参数的如何设置，选择什么样的训练方式。
2. 实验设计，需要记录哪些实验结果，通过什么方式与其它方法进行比较。

解决方法：

- 查阅相关的国内外文献，与老师和同学进行讨论。

参考文献

[1] HAIXIANG G, YIJING L, SHANG J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220–239.

[2] SUN Y, WONG A K, KAMEL M S. Classification of imbalanced data: A review[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(04): 687–719.

[3] WANG S, YAO X. Multiclass Imbalance Problems: Analysis and Potential Solutions[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2012, 42(4): 1119–1130.

[4] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge & Data Engineering, 2008(9): 1263–1284.

[5] VAN HULSE J, KHOSHGOFTAAR T. Knowledge discovery from imbalanced and noisy data[J]. Data & Knowledge Engineering, 2009, 68(12): 1513–1542.

[6] GARCÍA V, SÁNCHEZ J S, MOLLINEDA R A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance[J]. Knowledge-Based Systems, 2012, 25(1): 13–21.

[7] SPELMEN V S, PORKODI R. A Review on Handling Imbalanced Data[C] // 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). 2018: 1–11.

[8] LING C X, LI C. Data mining for direct marketing: Problems and solutions.[C] // Kdd: Vol 98. 1998: 73–79.

[9] 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.

[10] 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. 2005 : 878 – 887.

[11] HU S, LIANG Y, MA L, et al. MSMOTE: Improving classification performance when training data is imbalanced[C] // 2009 second international workshop on computer science and engineering : Vol 2. 2009 : 13 – 17.

[12] SÁEZ J A, LUENGO J, STEFANOWSKI J, et al. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291 : 184 – 203.

[13] ALI H, SALLEH M N M, SAEDUDIN R, et al. Imbalance class problems in data mining: a review[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2019, 14(3) : 1560 – 1571.

[14] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C] // Conference on Artificial Intelligence in Medicine in Europe. 2001 : 63 – 66.

[15] KUBAT M, MATWIN S, OTHERS. Addressing the curse of imbalanced training sets: one-sided selection[C] // Icml : Vol 97. 1997 : 179 – 186.

[16] HA J, LEE J-S. A new under-sampling method using genetic algorithm for imbalanced data classification[C] // Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. 2016 : 95.

[17] ESTABROOKS A, JO T, JAPKOWICZ N. A multiple resampling method for learning from imbalanced data sets[J]. Computational intelligence, 2004, 20(1) : 18 – 36.

[18] BARANDELA R, VALDOVINOS R M, SÁNCHEZ J S, et al. The imbalanced training sample problem: Under or over sampling?[C] // Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR). 2004 : 806 – 814.

[19] BREIMAN L. Bagging predictors[J]. Machine learning, 1996, 24(2) : 123 – 140.

[20] SUN Z, SONG Q, ZHU X, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5) : 1623 – 1637.

[21] LI K, XIE P, ZHAI J, et al. An improved adaboost algorithm for imbalanced data based on weighted KNN[C] // 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(. 2017 : 30 – 34.

[22] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of computer and system sciences, 1997, 55(1) : 119 – 139.

[23] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C] // European conference on principles of data mining and knowledge discovery. 2003 : 107 – 119.

[24] SCHAPIRE R E. The strength of weak learnability[J]. Machine learning, 1990, 5(2) : 197 – 227.

[25] RAYHAN F, AHMED S, MAHBUB A, et al. Cusboost: cluster-based under-sampling with boosting for imbalanced classification[C] // 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS). 2017 : 1 – 5.

[26] COHEN G, HILARIO M, SAX H, et al. Learning from imbalanced data in surveillance of nosocomial infection[J]. Artificial intelligence in medicine, 2006, 37(1) : 7 – 18.

[27] YONG Y. The research of imbalanced data set of sample sampling method based on K-means cluster and genetic algorithm[J]. Energy Procedia, 2012, 17 : 164 – 170.

[28] 王莉, 陈红梅, 王生武. 新的基于代价敏感集成学习的非平衡数据集分类方法NIBoost[J]. 计算机应用, 2019, 39 : 629 – 633.

[29] 赵锦阳, 卢会国, 蒋娟萍, et al. 一种非平衡数据分类的过采样随机森林算法[J]. 计算机应用与软件, 2019, 36: 255 – 261+316.