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

1.2 课题研究目的和意义

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，即数据中某一类别的样例数远远小于其它类别。这就会对我们的数据研究和分析造成一定的困难，令得出的结论缺乏可信度。对少数类样本进行扩充，增加其样例数目，就是一种非常有效的方法。然而传统的数据集扩充方法，仅仅可以针对某一类型的数据集进行扩充，或者缺乏可解释性，并且效果有限。因此，研究一种可解释性强，并且广泛适用于不同数据集的有效的数据扩充方法是非常必要的。

1. 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。根据数据量的多少和数据不平衡程度，可以将数据不平衡问题分为以下几类：大数据。针对类别不平衡问题的解决方案主要分为以下三类:数据层面、算法层面以及混合的方法。

数据层面，主要通过样本采样技术对数据集进行预处理，如对多数类样例进行下采样、对少数类样例进行上采样，或者同时使用这两种方法，从而达到不同类别之间样本数量比例的相对平衡。Ling等人提出了随机上采样和随机降采样算法，其算法思想是目前最简单的样本采样技术。随后，由于随机过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，Chawla等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。HAN等在SMOTE的基础上，提出了Borderline-Smote算法，解决了SMOTE算法生成样本重叠的问题，使生成的样本更有效。近年来，样本采样算法大多也是基于SMOTE方法的改进。

算法层面，主要是通过对已有的分类算法进行改进，以弱化分类器受多数类样本的影响，增加少数类样本的重要性。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、基于专家决策的主动学习以及主要用于异常样本检测中的单类学习等。参考文献

混合的方法，结合数据层面和算法层面的方法，集成方法就是其中最重要的一种。集成方法是提高分类器性能的一种常用方法，将此方法用在非平衡问题中，也可以显著的提高分类的性能。参考文献

1. 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）达到一定的比例。为了防止生成过多的少数类样本，人为控制两种类别数据数据的比例。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，通过集成学习将这K个分类器的结果集成，得到最终的预测结果。本论文拟研究的主要内容包括：

1、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不是一个有意义的图像，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，训练生成模型用于生成少数类样本。

2、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，本论文拟采样一个指标（如类内散度最大化）来评价生成样本的多样性，从而可以有效的扩充少数类样本增加有效的数据信息。

3、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。

4、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，得到分类器性能的变化，从而找到一个合适的平衡比例。

5、对于数据集类别不平衡比例问题，目前还没有一个明确的定义。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，研究在网络模型不变的情况下，对于不同的数据集，当少数类样本数量达到多少时就无法有效训练生成器。

3.2 创新点

1、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。

2、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。

3、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、通过实验的方法，

四、研究方案及进度安排，预期达到的目标

4.1 研究方案

4.2 研究进度安排

4.3 预期达到的目标

五、为完成课题已具备和所需的条件

5.1 已具备的条件

5.2 所需的条件

六、预计研究过程中可能遇到的困难和问题以及解决的措施

1. 主要参考文献

【1】Ling C X, Li C. Data mining for direct marketing: Problems and solutions[C]//Kdd. 1998, 98: 73-79.

【2】Chawla N V , Bowyer K W , Hall L O , et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1):321-357.