**南京审计大学**

**硕士研究生学位论文开题报告书**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 姓 名 | 杜玉坤 | | 性别 | 男 | 出生年月 | 1997.12.16 | 入学时间 | 2021.9 |
| 学 号 | MG2108129 | | 学科（类别） | | 统计学 | 研究方向 | 数据挖掘 | |
| 导师姓名 | | 苍玉权 | | | 职 称 | | 教授 | |
| 报告日期 | |  | | | 报告地点 | |  | |
| 论文选题 | | 一种基于自适应子空间插值的数据合成方法 | | | | | | |
| **（开题报告正文）**  **1、研究背景及意义**  在日常使用机器学习方法进行预测时,用于训练模型的数据经常会面临样本量不足、数据集观测误差较大等一系列问题。在大数据时代,可以通过使用数据合成的方法生成数据，它提供了一种经济高效的替代方案来收集和标记大量真实世界数据。此外，这些方法可以解决与真实世界数据相关的隐私问题，使其更加安全地共享和分析。  目前,代表性的数据合成方法可以分为三类。第一类包括插值、外推以及基于对抗网络深度学习等方法，这些方法会生成额外的数据样本，基于数据基础的分布进行数据合成,主要目的是提升数据的数量,可以提高数据集的质量和模型的泛化能力。第二类方法则是解决数据集不平衡的问题，例如SMOTE及其一些在SMOTE基础上扩展的方法。这些方法可以合成少数类样本，平衡数据集并提高模型的性能。最后，数据共享和研究需要对敏感数据进行隐私保护。合成数据可以帮助保护敏感信息，同时仍然实现数据共享和研究。通常，我们可以通过添加随机噪声来保护原始数据，生成用于共享和研究目的的合成数据，如差分隐私方法。总体而言，这些方法有助于优化数据在各种应用中的表示和使用。  对于某些实际任务，原始数据往往包含许多特征和未知的噪声,由于合成方法主要根据现有数据生成新数据，因此新数据的质量取决于原始数据的质量和数量。如果原始数据的质量较差或数量不足，则合成数据的质量可能会有限,可能会存在与真实的分布情况差异较大的问题。然而，对于受限规模和复杂噪声的数据集，缺乏有效的合成方法以扩展数据集的大小。  受分段线性插值和样条插值的启发，本研究提出了一种稳健而稳定的数据合成方法，命名为ASISO（Adaptive Subspace Interpolation for Sample Optimization ，该方法旨在自适应调整包含未知噪声的原数据集的样本大小和结构,其思想是将原始特征空间划分为多个子空间，每个子空间具有相等数量的样本，然后对相邻子空间中的样本进行线性插值。这种方法在两个方面实现了样本优化:首先，它可以自适应地调整数据集的大小,扩展后的数据通常与实际数据的误差非常小;其次，它可以调整样本的结构，显着减少具有大误差比例的样本的比例，从而减小数据噪声对于模型泛化的影响。  **2、相关研究现状**  (**1)数据噪声处理**  通过特征选择的方法，可以选择具有较高信息量和较低冗余的特征，减少高维数据的维度，提高模型的泛化能力和鲁棒性，如张乃明,张丽,卢维宏等提出的使用改进哈里斯鹰算法用于ＳＶＭ参数调整和特征子集选取，在降低特征维度的情况下达到了实现更高的分类准确率，实现同步优化效果。通过维度简约等算法，可以达到降低数据的维度，减少特征数目的效果，如刘文博等提出的基于t类加权核函数的主成分分析维度约简算法，该方法依据核矩阵特征值构造核函数权重，将多个核函数进行组合加权，进而达到了特征降维的目的。通过噪音过滤的方法，剔除数据中的噪音或异常样本，提高数据质量。如蒋远等人在检验数据质量报告的问题中，提出的基于生成模型的噪音过滤方法,达到了过滤与安全缺陷报告语义相似并且可能是噪音的非安全缺陷报告的效果。通过集成学习的方法，减小模型对噪音数据的敏感性，提高模型的鲁棒性，如[4]提出了一种基于RELM的时间序列数据加权集成分类方法，该方法首先针对时间序列数据中所含有的噪声,利用小波包变换方法对时间序列数据进行去噪处理,再使用一种基于粒子群优化PSO算法，对 RELM基分类器的权值进行优化，通过去噪，提高了模型的鲁棒性。使用统计模型或算法来建模噪音，并对数据进行去噪或噪音修复，如[5]提出一种鲁棒的深度学习去噪算法，该算法采用多层卷积神经网络估计噪声分布并且综合考虑地震数据的全局和局部信息，并利用残差学习策略提取噪声特征，并将两部分网络采用L1范数作为损失函数，增强网络模型的泛化能力，减少了模型的噪音；如[6]提出的一种基于改进多路径匹配追踪算法 (MMP)的高频超声信号稀疏去噪方法，该算法在迭代过程中设置阈值和引入剪枝操作,筛选误差较大的路径,减少迭代路径，通过增加少量的计算,实现信号数据重构精度的提高；  **(2)插值方法研究**  本文提出的方法是在插值思想的基础上进行改进,传统的插值方法是基于已知数据的函数值外推和预测点。对于给定的数据集，通常有两种情景。在两个已知的数据点之间进行插值，可根据距离进行选择，如最近邻插值。如果我们假设两个数据点之间的未知点与它们之间的直线相关，那么使用线性插补或者分段线性插补的方法，如具有非等距节点的分段线性插值，有学者提出的移位质量最大化的线性插值法。当需要对多个给定点进行插值时，常用的方法是通过构造基于这些给定点的高次多项式来预测未知点，其中包括拉格朗日插值和牛顿插值等方法。随着多项式的次数的增多，需要的数据就越多，需要插值次数的越也就越高，可能会存在模型过拟合的现象，导致插值结果偏离原函数，也称龙格现象。一种解决方案是使用局部多项式进行拟合，从而得到一个光滑函数。与使用高次多项式进行插值相比，这种方法可以使得模型具有更好的平滑性和数值稳定性，包括样条插值等。然而，由于上述方法需要对多个局部多项式进行拟合，该方法可能会导致相对较高的计算复杂度。  在使用插值方法扩展样本量时，节点的数量选择对插值结果的准确性和稳定性非常重要。一般情况下，我们可以使用等距节点，即在插值区间内均匀间隔的选择节点。另一种常用的方法是Chebyshev节点，它在插值区间内满足一定条件时，可以更好地进行函数的拟合。此外，插值的数量也会影响数据的稳定性和验证精度。确定适当的插值数量的一种方法是将不同插值数量下的训练模型与原始数据的模型进行比较。然而，这种方法可能导致模型的过拟合以及泛化能力较差。  **3、研究方案**  **(1)主要研究内容**  ASISO数据合成方法主要基于线性插值方法扩充样本并且提高样本的整体质量,其思想是将数据集原始特征空间划分为若干子空间，每个子空间包含相同数量的样本，然后对相邻子空间中的样本进行线性插值。该方法需要预先确定两个超参数（K 和 η）,参数K的直观解释是每个特征子空间中存在的样本数，参数η指线性插值时单位距离插入的样本数量。ASISO方法还涉及两个算法,即用于聚类的K-Space算法以及用于样本插值匹配的K-Match算法。  图表, 散点图  描述已自动生成  图1. ASLI方法优化过程  该方法流程图如图1所示。图1(a)中展示了多个含有噪声的样本点,其中f(x)为变量之间的真实函数关系; 首先使用K-Space算法对样本点进行聚类,如图1(b)所示;在图1(c)中,使用K-Match算法对相邻子空间的样本进行样本插值匹配;对于匹配的样本之间进行线性插值,如图1(d)所示.  对于给定数据集,其中,.我们假设数据是含有噪声的,设 与 的真实值为 ,。考虑我们假设是一个连续的函数, 表示 与 之间的真实函数关系。对于含噪数据, 可得:  其中, *,* 分别为与 的噪音,为误差项.可将公式(1)转换为一种普遍的表示:  令, 对于, 我们称  为特征最小点.  给定超参数k,我们使用无监督聚类K-Space算法对原始数据进行聚类.通过K-Space算法,可以讲原始特征空间分为n/k个特征子空间(注:对于n/k不为整数的情况,我们后文会进行探讨分析),每个子空间包含等量的k个样本,即, 原始数据集可以根据每个子空间包含的样本划分为多个子集,即, ,如图1(b)所示.对于两个相邻的子空间,由于是连续的, 因此我们可以假设在两个相邻子空间中,可以拟合为一个线性函数.公式(2) 可以转化为:  其中, 表示线性拟合误差.当子空间的测度以及相邻子空间的距离趋向于0时,有聚类之后,下一步需要对不同子空间对应的数据子集进行排序.  首先,需要计算各个簇的簇心,如下:  在集合中,我们定义为簇心距离特征最小点 最近的子集,定义为簇心距离最近的子集,为的簇心,,且.  对排序后的子集之间进行插值,序号相邻的子集为相邻的子空间对应的数据集, 仅在相邻的子空间之间进行插值.  当对相邻子空间之间进行插值时,插值规则如下:  1、进行插值的两个样本属于不同的相邻子空间;  2、每个样本都要进行插值;  3、每个样本只能进行一次插值.  根据以上规则,可选择的插值方案有 个,并不是所有的方案插值后都能达到很好的样本优化效果,因此我们通过K-Match样本匹配算法对相邻子空间当中的个样本进行线性插值匹配,该算法可以从个方案中快速且高效的选择优化效果较好的插值方案.  假设和都是连续的变量,给定另外一个超参数η,对相邻子空间与对应的样本进行线性插值,插入的样本数量为  .  以 and 为例, 插值生成的样本数据集为.插值公式为:  , (7)    ASLI方法的主要处理过程见算法1.   |  | | --- | | **算法1** ASLI算法 | | 输入: 原始数据集 , 超参数 和 η  输出: 优化后的数据集 | | 1: 通过K-Space算法对原始数据集进行聚类处理,得:, 其中  2: 计算每个簇的簇心:  3: 根据公式(5)和(6)对多个簇进行排序,可得:  4: 令 进行迭代循环  5: 通过K-Match算法对数据子集 和 的样本进行匹配,得:  6: 令 进行迭代循环  7: 通过公式(7)和(8)计算合成新样本  8: |   综上,对数据集进行ASLI处理,数据集应当满足的假设如下:  1、是一个连续的函数;  2、线性拟合误差 ;  3、 和 是连续的变量.  **(2)论文整体结构安排**  第一章：阐述选题的研究背景及意义，对国内外与研究内容相关的文献进行综述回顾，并说明论文的主要内容和框架结构。  第二章：介绍所涉及的相关概念及其理论基础;  第三章：介绍ASISO具体方法,引入相关定理并给出证明;  第四章：模拟实验,主要通过数据的模拟,检验超参数的选择策略,并且探究样本的优化效果;  第五章: 实证分析,结合多个机器学习方法,通过多个实际数据集来探究使用ASISO优化样本前后机器学习模型泛化提升的效果;  第六章: 进一步研究ASISO方法在图像数据的应用,结合Darknet、Resnet、VGG等深度学习图像特征提取模型对图像提取特征,然后给予特征数据进行数据合成;  第七章：本文的总结与展望。  **4、可行性分析**    本人研究生期间主要的研究方向为数据挖掘,机器学习与深度学习,对于相关的数据合成算法,深入研究了相关文献,并且积极与课题组成员深入探讨, 有针对性的对各种方法进行理论分析。本次研究已经发表见刊了学术论文一篇,审稿中的学术论文一篇。本次研究经过学校老师以及外界专家多轮审核,并且一直在不断完善。目前，论文的代码、理论以及相关实验都已经基本完成。  **5、本文创新点**  本文基于线性插值、样条插值以及KNN的思想,提出了一种全新的数据合成方法,该方法可以自适应的调整样本数量以及样本结构,并且该方法可以很好适用于含有大量未知噪声的数据集,可以起到很好的数据优化的效果:1、对原始含噪的数据集自适应地扩充了样本量,并且扩充生成的样本具有更小的噪声,降低了样本的平均噪声;2、自适应的调整样本结构,可以使得误差较大的样本量占比减小。  **6、研究进度及时间安排**  2022年6月——2022年9月，完成期刊论文撰写并投稿。  2022年9月——2022年10月，收集整理文献，增加模拟实验及实证分析。  2022年11月——2023年3月，进一步完善期刊论文,再次撰写一篇论文。  2023年4月——2023年9月，毕业论文撰写，完成论文初稿，并在老师的指导下进行修改。  **参考文献**  [1] Y. Lu, H. Wang, and W. Wei, “Machine Learning for Synthetic Data Generation: A Review,” 2023, doi: 10.48550/ARXIV.2302.04062.  [2] S. I. Nikolenko, *Synthetic Data for Deep Learning*, vol. 174. in Springer Optimization and Its Applications, vol. 174. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-75178-4.  [3] B. Mahesh, *Machine Learning Algorithms -A Review*. 2019. doi: 10.21275/ART20203995.  [4] M. Lepot, J.-B. Aubin, and F. H. L. R. Clemens, “Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment,” *Water*, vol. 9, no. 10, Art. no. 10, Oct. 2017, doi: 10.3390/w9100796.  [5] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, “A review of medical image data augmentation techniques for deep learning applications,” *J. Med. Imaging Radiat. Oncol.*, vol. 65, no. 5, pp. 545–563, Aug. 2021, doi: 10.1111/1754-9485.13261.  [6] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.  [7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.  [8] D. Dablain, B. Krawczyk, and N. V. Chawla, “DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2022, doi: 10.1109/TNNLS.2021.3136503.  [9] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” in *Advances in Intelligent Computing*, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887.  [10] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, “Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem,” in *Advances in Knowledge Discovery and Data Mining*, T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 475–482.  [11] T. Ha, T. K. Dang, T. T. Dang, T. A. Truong, and M. T. Nguyen, “Differential Privacy in Deep Learning: An Overview,” in *2019 International Conference on Advanced Computing and Applications (ACOMP)*, 2019, pp. 97–102. doi: 10.1109/ACOMP.2019.00022.  [12] D. Meng and F. De La Torre, “Robust matrix factorization with unknown noise,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1337–1344.  [13] I. J. Goodfellow *et al.*, “Generative adversarial networks (2014),” *ArXiv Prepr. ArXiv14062661*, vol. 1406, 2014.  [14] R. Sibson, *A Brief Description of Natural Neighbor Interpolation*. 1981, p. 36.  [15] E. W. Cheney and D. R. Kincaid, *Numerical Mathematics and Computing*. Cengage Learning, 2012.  [16] T. Blu, P. Thevenaz, and M. Unser, “Linear interpolation revitalized,” *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 710–719, May 2004, doi: 10.1109/TIP.2004.826093.  [17] J.-P. Berrut and L. N. Trefethen, “Barycentric Lagrange Interpolation,” *SIAM Rev.*, vol. 46, no. 3, pp. 501–517, Jan. 2004, doi: 10.1137/S0036144502417715.  [18] J. P. Musial, M. M. Verstraete, and N. Gobron, “Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series,” *Atmospheric Chem. Phys.*, vol. 11, no. 15, pp. 7905–7923, Aug. 2011, doi: 10.5194/acp-11-7905-2011.  [19] C. Runge, “Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten,” *Z. Für Math. Phys.*, vol. 46, no. 224–243, p. 20, 1901.  [20] C. A. Rabbath and D. Corriveau, “A comparison of piecewise cubic Hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics,” *Def. Technol.*, vol. 15, no. 5, pp. 741–757, Oct. 2019, doi: 10.1016/j.dt.2019.07.016.  [21] C. Habermann and F. Kindermann, “Multidimensional Spline Interpolation: Theory and Applications,” *Comput. Econ.*, vol. 30, no. 2, pp. 153–169, Sep. 2007, doi: 10.1007/s10614-007-9092-4.  [22] M. I. Ganzburg, “The Bernstein Constant and Polynomial Interpolation at the Chebyshev Nodes,” *J. Approx. Theory*, vol. 119, no. 2, pp. 193–213, Dec. 2002, doi: 10.1006/jath.2002.3729.  [23] W. S. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” *J. Am. Stat. Assoc.*, vol. 74, no. 368, pp. 829–836, 1979.  [24] D. D. Lichti, T. O. Chan, and D. Belton, “Linear regression with an observation distribution model,” *J. Geod.*, vol. 95, no. 2, p. 23, Feb. 2021, doi: 10.1007/s00190-021-01484-x.  [25] C. Liu, B. Li, Y. Vorobeychik, and A. Oprea, “Robust linear regression against training data poisoning,” in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 91–102.  [26] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: identifying density-based local outliers,” in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.  [27] Y. Guo, W. Wang, and X. Wang, “A Robust Linear Regression Feature Selection Method for Data Sets With Unknown Noise,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 31–44, 2023, doi: 10.1109/TKDE.2021.3076891. | | | | | | | | |