**哈尔滨工业大学深圳研究生院**

**本科立项申请**

|  |  |  |  |
| --- | --- | --- | --- |
| **项目名称：** | 生成式模型在不平衡样本分类方法中的应用 | | |
| **依托单位：** | 哈尔滨工业大学深圳研究生院 | | |
| **项目负责人：** |  | **移动电话：** | 13798519353 |
| **电子邮箱：** | ckzhang812@gmail.com | **传 真：** |  |

# 一、立项背景（研究现状、趋势、研究意义等）

生成模型是目前机器学习中研究比较多的一个领域，在机器学习中，我们得到的数据通常假设其为独立同分布数据，生成样本则是指通过对其概率密度分布进行建模，并在该分布上进行采样的结果。但是由于数据的高维分布，使得我们难以通过建模方式来获取其分布，生成模型则是在不直接对数据分布进行建模的前提下，对数据样本进行扩充，或者是根据现有样本，合成一些新的样本。

生成式模型不仅在人工智能领域占有重要地位,生成方法本身也具有很大的研究价值。生成方法和判别方法是机器学习中监督学习方法的两个分支。生成式模型是生成方法学习得到的模型，生成方法涉及对数据的分布假设和分布参数学习, 并能够根据学习而来的模型采样出新的样本[1]。

不平衡数据集通常是指样本数据集的类别分布表现出不平衡的问题，数据集的不平衡问题会影响分类器的性能。当一些传统的方法面对不平衡问题时，往往不能取得很好的效果[2]，因为现有的成熟的分类器设计都是基于类分布大致均衡这一假设，用于训练的数据集是大致平衡的，以提高数据集的总体分类准确率为目标，忽略了少数类样本的重要性。所以当这些分类器应用到不平衡数据集的分类问题时，对于多数类的有较高的识别率，但少数类的识别率却很低。不平衡样本是生物信息数据的一个常见的问题，实际问题中也存在大量不平衡样本集的分类问题:信用卡欺诈行为检测，网络入侵检测，医学疾病诊断等。实际上不仅是类别数目的不平衡影响分类算法的结果，不同类的数据间的重合程度也会影响分类结果。而且数据集本身的概念复杂性、类重叠与类的不平衡分布的问题，使得不平衡数据的学习受到了广泛的关注和研究。

# 二、国内外研究现状和发展趋势

## 1．国内外研究现状

## 2．研究现状存在的核心问题和本项目的研究设想

* **研究现状存在的核心问题**

①

* **本项目的研究设想**

**参考文献**

[1] Esling P, Agon C. Time-series data mining[J]. ACM Computing Surveys (CSUR), 2012, 45(1): 12.

[2] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases[M]. ACM, 1994.

[3] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reduction for fast similarity search in large time series databases[J]. Knowledge and information Systems, 2001, 3(3): 263-286.

[4] Lin J, Vlachos M, Keogh E, et al. A MPAA-based iterative clustering algorithm augmented by nearest neighbors search for time-series data streams[M] //Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2005: 333-342.

[5] Korn F, Jagadish H V, Faloutsos C. Efficiently supporting ad hoc queries in large datasets of time sequences[J]. ACM SIGMOD Record, 1997, 26(2): 289-300.

[6] Lin J, Keogh E, Lonardi S, et al. A symbolic representation of time series, with implications for streaming algorithms[C]//Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM, 2003: 2-11.

[7] Shieh J, Keogh E. iSAX: indexing and mining terabyte sized time series[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 623-631.

[8] Keogh E J, Pazzani M J. An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback[C] //KDD. 1998, 98: 239-243.

[9] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration[J]. Data Mining and knowledge discovery, 2003, 7(4): 349-371.

[10] Gupta M, Gao J, Aggarwal C, et al. Outlier detection for temporal data[J]. Synthesis Lectures on Data Mining and Knowledge Discovery, 2014, 5(1): 1-129.

[11] Dasgupta D, Forrest S. Novelty detection in time series data using ideas from immunology [C] // Proceedings of the international conference on intelligent systems. 1996: 82-87.

[12] Ma J, Perkins S. Time-series novelty detection using one-class support vector machines[C]//Neural Networks, 2003. Proceedings of the International Joint Conference on. IEEE, 2003, 3: 1741-1745.

[13] Keogh E, Lin J, Fu A. Hot sax: Efficiently finding the most unusual time series subsequence[C] //Data mining, fifth IEEE international conference on. IEEE, 2005: 8 pp.

[14] Izakian H, Pedrycz W. Anomaly detection in time series data using a fuzzy c-means clustering[C] //IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint. IEEE, 2013: 1513-1518.

[15] Sivaraks H, Ratanamahatana C A. Robust and Accurate Anomaly Detection in ECG Artifacts Using Time Series Motif Discovery[J]. Computational and mathematical methods in medicine, 2015, 2015.

[16] 肖辉. 时间序列的相似性查询与异常检测[D]. 上海: 复旦大学, 2005.

[17] 詹艳艳, 陈晓云, 徐荣聪. 基于时间序列模式表示的异常检测算法[J]. 计算机应用研究, 2007, 24(11): 78-79.

[18] 杜洪波, 张颖. 基于 LLM 的时间序列异常子序列检测算法[J]. 沈阳工业大学学报, 2009, 31(3): 328-332.

[19] 汪斐. 在线的时间序列异常检测算法研究[D]. 中山大学, 2008.

[20] Li G, Bräysy O, Jiang L, et al. Finding time series discord based on bit representation clustering[J]. Knowledge-Based Systems, 2013, 54: 243-254.

[21] Agrawal R, Srikant R. Privacy Preserving Data Mining[C] //ACM Sigmod Record. ACM, 2000, 29 (2) :439-450

[22] Linden Y, Pinkas B. Privacy Preserving Data Mining[C]//Advances in Cryptology--CRYPTO 2000. Springer Berlin Heidelberg, 2000: 36-54.

[23] Yao A. Protocols for Secure Symposium on Foundations Computations[C]//Proceedings of Twenty-third IEEE of Computer Science. FOCS, 1982, 82: 160-164.

[24] Dolev D, Yao A C. On the Security of Public Key Protocols [J].IEEE Transactions on Information Theory, 1983, 29(2): 198-208.

[25] Yao A. How to Generate and Exchange Secrets[C]// Proceedings of the Foundations of Computer Science, 1986, 27th Annual Symposium on IEEE, 1986: 162-167.

[26] Liu X, Deng R, Choo K K R, et al. An Efficient Privacy-Preserving Outsourced Calculation Toolkits with Multiple Keys[J]. IEEE Transactions on Information Forensics & Security, 2016, 11(11):1-1.

[27] Abadi M, Feigenbaum J. Secure Circuit Evaluation [J].Journal of Cryptology, 1990, 2(1): 1-12.

[28] Goldreich O. Secure Multi-party Computation [J].Manuscript. Preliminary Version 1998:1-10.

[29] Oliveira S R M, Zaiane O R. Privacy Preserving Frequent Itemset Mining[C] //Proceedings of the IEEE International Conference on Privacy, Security and Data mining-Volume 14. Australian Computer Society, 2002: 43-54.

[30] Kantarcioglu M, Clifton C. Privately Computing a Distributed K-NN Classifier [M].Knowledge Discovery in Databases: PKDD 2004. Springer Berfin Heidelberg, 2004:279-290.

[31] Yu H, Jiang X, Vaidya J. Privacy Preserving SVM Using Nonlinear Kernels Horizontally Partitioned Data[C]//Proceedings of the 2006 ACM Applied Computing. ACM, 2006: 603-610.

[32] Evfimievski A, Srikant R, Agrawal R, et al. Privacy Preserving Mining of Association Rules[J]. Information Systems, 2004, 29(4): 343-364.

[33] Lin X, Clifton C, Zhu M. Privacy Preserving Clustering with Distributed EM Mixture Modeling[J]. Knowledge and Information Systems, 2005, 8(1): 68-81.

[34] Vaidya J, Kantarcioglu M, Clifton C. Privacy-Preserving Naive Bayes Classification[J].The VLDB JournalThe International Journal on Very Data Bases, 2008, 17(4): 879-898.

[35] Liu D, Bertino E, Yi X. Privacy of Outsourced k-mean Clustering[C]// ACM Symposium on Information, Computer and Communications Security. ACM, 2014:123-134.

[36] Lindell Y, Pinkas B. An Efficient Protocol for Secure Two-Party Computation in the Presence of Malicious Adversaries[J]. Journal of Cryptology, 2015, 28(2):312-350.

# 三、拟解决的关键科学问题和主要研究内容

## 研究内容

## 拟解决的关键问题

# 总体研究方案

## 学术思路

## 技术途径

## 可行性分析

# 项目计划目标

## 总体目标

本项目以基于多方计算下的时间序列异常检测为研究对象，通过设计安全多方协议使得算法能够将时间序列数据在加密后的密文条件下依然能够检测出异常数据。把时间序列异常检测、加密算法、以及安全多方协议等一系列内容整合成一个在安全多方下隐私保护的心电异常检测原型系统。

## 预期目标

**理论成果：**以高质量的学术论文5-6篇、研究生学位论文7-8人的形式提交。

**应用成果：**成果形式包括原型系统、相应软件以及关键技术的技术标准和专利申请。建立一个在安全多方计算下隐私保护的心电数据异常检测的原型系统，保证系统安全性、准确性。

**申请专利：**计划申报安全多方计算下时间序列异常检测算法的软件著作权，申请3项相对应的技术发明专利。

**人才培养：**培养时间序列研究和信息安全领域研究生7-8名。

**主要技术指标：**

* 安全性：确保在半诚实模型下，整个系统是安全可靠的，所有算法的中间结果，通讯协议的数据都不会导致隐私的泄露
* 准确性：针对给定的时间序列异常的定义，确保针对时间序列的检测时可以准确无误的将异常检测出来并返回客户端，尤其对于心电数据的检测，准确性的要求较高。
* 效率性：针对密文下的数据通常所需要的时间代价是非常大的，通过优化算法，保证系统的时间耗费满足异常检测的需求。

# 六、项目计划进度

## 1.计划进度

本项目拟在两年内完成，其研究的总体安排与进度如表1：

表1 项目计划进度

|  |  |
| --- | --- |
| 2017.10-2018.2 | 完成数据的采集工作，对数据的预处理部分，研究时间序列异常相关知识，包含针对具体的时间序列数据采取何种表达方式，对异常的定义，意见在明文下高效的时间序列异常检测算法的设计和实现。发表阶段性学术论文1-2篇，参加学术会议及讨论。经费拟使用所批准经费的30%左右。 |
| 2018.2-2018.9 | 针对现有的时间序列异常检测方法做出总结，简化算法流程，在算法层面上适应加密的需求。发表阶段性学术论文1篇，参加学术会议，经费拟使用所批准经费的20%左右。 |
| 2018.9-2019.2 | 针对安全多方计算的需求，根据前面工作中设计的时间序列异常检测算法设计安全多方协议，保证在半诚实模型下协议的安全可靠，测试不同的加密算法耗费资源的差异并选择最优加密算法，发表阶段性论文2-3篇。参加学术会议并讨论，经费拟使用所批准经费的30%左右。 |
| 2019.2-2019.10 | 根据协议实现异常检测算法，并针对不同的测试集进行测试，验证可行性后。运用至心电检测的原型系统中，反复验证，开发出安全、准确、快速的多方下隐私保护的心电检测系统。参加学术会议，对相关成果申请专利，邀请领域内专家交流讨论。经费拟使用所批准经费的20%左右。 |

## 2．工作条件

项目组依托哈尔滨工业大学深圳研究生院网络信息中心，所在的移动计算实验室拥有40多台高性能计算机，10台数据处理工作站，以及5台高性能服务器。本项目计划先在实验室使用数据处理工作站搭建小型分布式云平台，达到预期效果后，接入开放的云平台上，开展进一步的试验验证，并由此优化本项目的研究成果。

并且，哈尔滨工业大学深圳研究生院计算机学科已建成计算机应用中心、网络信息中心、智能计算等研究中心，每年投入的建设经费超过百万元，拥有一批Lenovo、Dell、HP等品牌的先进的试验设备和软件，可以满足项目研究的软硬件需要。另外，在文献资料方面，我校图书馆已经购买多个知名数据库镜像，可以查阅和下载历史和最新的国内外会议、期刊论文，了解最新的研究动态。本项目负责人及项目组成员跟国际同行有较密切联系，能够及时获得国际最新成果。

# 七、研究团队

## 项目负责人和核心研究人员简介

1. **项目负责人简介**
2. **项目组核心成员简介**