**哈尔滨工业大学**

**硕士学位论文开题报告**

**题 目：基于人工智能的数据隐私保护算法研究**

**院 （系） 计算机科学与技术**

**学 科 计算机科学与技术**

**导 师 何慧**

**研 究 生 郭睿**

**学 号 17S103173**

**开题报告日期 2018.09.11**

**研究生院制**

目录

[1. 课题来源及研究的背景和意义 4](#_Toc524358390)

[1.1 课题的来源 4](#_Toc524358391)

[1.2 课题研究的背景和意义 4](#_Toc524358392)

[2. 国内外在该方向的研究现状及分析 5](#_Toc524358393)

[2.1 国内外研究现状 5](#_Toc524358394)

[2.2 国内外文献综述的简析 6](#_Toc524358395)

[2.2.1 不同匿名原则的适用性简析 6](#_Toc524358396)

[2.2.2 常见的脱敏方法 7](#_Toc524358397)

[2.2.3 常用脱敏效果评测方法 8](#_Toc524358398)

[3. 主要研究内容 9](#_Toc524358399)

[4. 已完成的研究工作 10](#_Toc524358400)

[5. 研究方案及进度安排，预期达到的目标和取得的研究成果 10](#_Toc524358401)

[5.1 研究方案 10](#_Toc524358402)

[5.1.1 脱敏算法的研究 10](#_Toc524358403)

[5.1.2 基于人工智能的隐私保护模型的研究 10](#_Toc524358404)

[5.2 预期达到的目标和取得的研究成果 10](#_Toc524358405)

[5.3 进度安排 10](#_Toc524358406)

[6. 为完成课题已具备和所需的条件和经费 11](#_Toc524358407)

[7. 预计研究过程中可能遇到的困难和问题，以及解决的措施 11](#_Toc524358408)

[8. 主要参考文献 11](#_Toc524358409)

# 课题来源及研究的背景和意义

## 课题的来源

本课题来源于项目“面向互联网+的云服务系统安全防护技术（2017YFB0801800）”中的“多域云安全管控关键技术及系统（2017YFB0801801）”课题，以及导师在隐私保护算法方面的指导。

## 课题研究的背景和意义

随着人类社会进入大数据的时代，人们的各类信息被政府部门、企业组织甚至个人收集[1]，如医疗档案、社会调查、人口普查和商业数据等。这些海量的数据信息被用于进行数据挖掘和机器学习等研究[2]，来帮助政府制定相关政策或者企业创造商业价值，同时也为人们提供更加丰富、智能和便捷的生活方式。但随着企业收集的数据越来越多，数据外泄事件一再发生，用户敏感数据遭到泄露，企业信息受到严重威胁，信誉遭受损失。为此，企业积极投资于数据隐私和数据安全技术，将不可预见的安全成本最小化并减少风险损失[3]。

在大数据时代的背景下，由于数据信息的内容形式以及收集、发布和传播的方式变得多样化，信息安全问题涉及的范围更加广泛，并且研究的内容也更加复杂和多样[4]。其中，人们的隐私安全问题一直是信息安全问题的研究核心之一[5]。在大数据技术的不断推动之下，各个行业和政府部门之间的信息壁障逐渐被打破。数据的交换和共享成为了信息交流中越来越重要的活动，但是，这些交流的数据中包含了大量个人隐私和敏感信息。如果这些数据在没有经过隐私保护处理之前就对外进行发布或者交换，会非常容易造成用户的隐私泄露[6]。例如，在2006年，著名的因特网服务提供商美国在线发布了一份2GB的数据文件，其中包括了关于65万用户的约2千万个查询词条信息。尽管这份数据使用了随机数代替数据中用户的ID，但是当数据发布之后，两名纽约时报的记者仍然通过数据中的信息准确地找到并且采访了其中的某位用户[7]。

为了保护数据中的用户隐私信息，学者们提出了一种被称为隐私保护数据发布的技术[8]。该技术主要通过匿名或者添加噪音的方式对数据表中的信息进行一定程度的加密，将用户隐私暴露的概率降低在一定阈值以内，它可以帮助如社交网站、保险公司、医院和政府部门等在对外发布数据时，既不会泄露其中用户的隐私信息，还可以保证匿名数据的信息可利用性。

因此，对于产生和存储大规模数据的企业和组织而言，如何在提供正常服务的同时保证用户隐私信息不被泄露，在生产环境中动态地对用户敏感数据进行脱敏，在数据交换、共享及使用等过程中实现对敏感数据的定向、精准和彻底脱敏，达到数据安全、可信、受控使用的目标，是数据产生者和管理者亟待解决的技术问题。

而且随着非敏感数据的增多，攻击者完全有可能设计特定的推断攻击模型来从不敏感数据推测并且还原出不公开的敏感数据。在无时无刻不断产生数据的生产环境中，传统的诸如k-匿名化及差分隐私等静态数据保护方法在实时性和数据可用性上无法满足生产环境等的要求。故提出了基于人工智能的隐私数据保护模型，结合人工智能，动态地学习和防御针对数据隐私的推断攻击。

# 国内外在该方向的研究现状及分析

## 国内外研究现状

隐私保护问题是伴随着数据应用而提出的。在统计领域，隐私保护问题最先受到关注。当前，隐私保护的主要研究方向如表1所示。

表 1 隐私保护研究方向

|  |  |
| --- | --- |
| 研究方向 | 实例 |
| 通用的隐私保护技术 | Perturbation、Randomization、Swapping、Encryption |
| 面向数据挖掘的隐私保护技术 | Association Rule Mining、Classification、Clustering |
| 基于隐私保护的数据发布原则 | k-anonymity、l-diversity、m-Invariance、t-Closeness |
| 隐私保护算法 | Anonymized Publication、Anonymization with High Utility |

通用的隐私保护技术致力于在较低应用层次上保护数据的隐私，一般通过引入统计模型和概率模型来实现；而面向数据挖掘的隐私保护技术主要解决在高层数据应用中，如何根据不同数据挖掘操作的特性,实现对隐私的保护；基于隐私保护的数据发布原则是为了提供一种在各类应用可以通用的隐私保护方法，进而使得在此基础上设计的隐私保护算法也具通用性[9]。

在敏感数据保护算法方面，对称/非对称加密、同态加密、访问控制、安全审计和备份恢复等传统数据安全防护手段在用于敏感数据防护方面仍有欠缺，无法在不妨碍已有的数据处理、操作及分析过程的同时，实现对敏感数据的针对性保护。例如：传统加密技术由对称、非对称和散列算法构成，具有极高的安全强度，能够保证数据在传输过程中的机密性和完整性。但是，由于数据在使用时必须完全解密，对最终用户而言，敏感数据依然是明文，因而无法同时满足敏感数据安全性和可用性的需求。

目前，关于敏感信息保护问题的研究的前提主要基于攻击者对用户背景知识的掌握程度，在此条件下攻击者可以进行身份链接攻击、属性链接攻击、成员链接攻击等隐私攻击，因此，相关学者提出了K-匿名、L-多样性以及相关的原则[10]来保证脱敏后数据表的隐私安全，当一个数据表满足了某一匿名原则的条件时，数据表就会具备相应的隐私保护能力。其中具有代表性的k-anonymity匿名原则[11]可以有效地防止数据表中用户身份的泄露，在介绍该原则前，需要先明确数据表中可能会出现的属性类型。一般情况下，记录型数据表包含三种属性类型[12]：（1）标识符属性（ExplicitIdentifier），能够唯一或者在很大程度上识别用户身份的属性，并且在发布的数据中需要被移除；（2）准标识符属性（Quasi-Identifier，简称QI），作为用户的一般属性记录用户的非敏感信息，并且尽管单一项属性无法准确地识别用户的身份，但是当使用多项QI值作为条件在数据表中进行匹配时会有极大的概率辨认用户的真实身份；（3）敏感属性（Sensitive Attribute），记录了用户的敏感信息。例如，姓名属于标识符属性，年龄、性别和邮政编码属于QI属性，疾病属于敏感属性。如果匿名数据表遵循*k*-anonymity匿名原则，当攻击者使用目标用户的QI信息进行匹配时至少会得到*k*个无法区分的匹配个体。

差分隐私保护技术[13]通过匿名或者添加噪音的方式，使得当添加或删除数据表中某个用户的信息时不会对数据的分析结果产生明显的影响。通常情况下，差分隐私相比于一般的匿名保护技术具有更加可靠的保护效果，但是在不同环境下由噪音引起的误差失真更加难以控制[14]。

## 国内外文献综述的简析

### 不同匿名原则的适用性简析

1. k-Anonymity可以避免身份泄露，这是因为被k匿名化的数据集中的记录无法被映射回与原始数据集相对应的记录。

缺点：k-Anonymity不能保护敏感属性的泄露。比如：具有相同QI属性的k个病人，都患有相同的疾病，那只要确定了某个具有类似QI值的病人身份，就能推断出患有该病。

1. p-Sensitive k-Anonymity用于防止敏感属性的泄露，

定义：k>1 并且 p<=k,在满足k-Anonymity的前提下，具有相同QI值的记录至少有p种不同的敏感属性值。

缺点：该原则适用于不同敏感属性值出现频率相似的情况，如果不是这样，则会造成大量的数据丢失。比如：1000个病人中只有5个AIDS属性值为yes，其余为no，现在需要满足2-Sensitive k-Anonymity原则，那么最多只能划分为QI值相同的5组记录，这就意味着QI值需要进行粗粒度的泛化，从而造成信息的丢失。

1. l-Diversity解决k-Anonymity中存在的敏感属性泄露问题。

定义：每组具有相同QI值的记录，在每个敏感属性上至少有l种“well-represented”。对于“well-represented”，有以下几种定义方式：

1）不同值l-diversity。这种方式与l-sensitive k-anonymity相同。

2）熵l-diversity。对于一个组内特定敏感属性的熵的定义如下所示：

其中，p(G,c)是c在G中出现的概率，C是敏感属性值域，满足entropy l-diversity的数据集每个组G都有，。

3）Recursive (c,l)-diversity确保让最频繁的值出现的不要太多，让最不频繁的值出现的不要太少。m表示组G中敏感属性值的个数 表示G中第i个最频繁出现值得次数。如果，就说组G满足递归(c,l)-diversity，如果数据集中每个组都满足递归(c,l)-diversity，那么整个数据集就满足递归(c,l)-diversity。

缺点：难以且没必要实现，同2中例子。有以下两种攻击其无法防御：

——偏斜攻击Skewness attack。满足上述原则的组中，患有AIDS和未患有AIDS的病人数目相同，那么如果攻击者成功确认一位特殊病人属于这个组，那么就可以得到在这个5/1000的数据集中，该病人患有AIDS的概率有50%。

——相似性攻击Similarity attack。如果敏感属性值存在语义上的相似，隐私泄露还是会发生。比如在一个组中存在{lung cancer，liver cancer，stomach cancer}三种属性值，那攻击者可以推断出该组中的人患有癌症。

1. t-Closeness解决了l-diversity中固有的敏感属性泄露问题。

定义：每个组中敏感属性的分布和整个数据集属性的分布，之间距离小于阈值t。

因为两者分布接近，所以可以规避Skewness attack，此外，因为组中的属性分布模拟了整个数据集中的属性分布，所以只要数据集中没有语义相似，那么组中就不会存在语义相似。

缺点：破坏了数据集的信息完整性，1.限制了有用信息的发布数量，2.破坏了QI和敏感信息值的相关性，不利于数据挖掘的进行，可以通过放宽t值来减小危害[15]。

总结：不同的匿名原则，都是发布静态的隐私保护数据集，无法有效地应对新产生的特定攻击模型。

### 常见的脱敏方法

数据匿名化一般采用两种基本操作：

(1)抑制：抑制某数据项,亦即不发布该数据项；

(2)泛化算法：它通过将数据表中的QI属性值进行一定程度的泛化，即将具体的数据值转化为概括和抽象的形式，从而防止攻击者使用目标用户的QI值获取用户在数据表中的个体标识。对于数字类型的属性，数值将被转化为值域的类型，如将数值24转化为[10-30]。而对于分类类型的属性，数值将根据用户自定义的泛化层次树进行泛化。

常用的噪音机制分别为拉普拉斯机制(Laplace mechanism)与指数机制，其他噪音机制包括高斯机制[16]、几何机制、矩阵机制、函数机制等。

### 常用脱敏效果评测方法

1. 敏感数据保护性评估

将脱敏后数据集合输入到相应的攻击模式中，计算攻击模型推断敏感数据的准确性。

1. 信息可利用性评估

目前，应用比较广泛的信息损失量评估方法包括：

1. 最小的失真（Minimal Distortion）：最小的失真[17]是被用于计算单个属性的信息损失量。对于一个属性，每当某个用户在该属性中的一个具体值被泛化时，惩罚值就会加1。文献[19]等使用最小的失真惩罚值作为算法中的启发式控制算法的分组策略。
2. 损失度量（Loss Metric）：损失度量[18]被用于计算将一个具体的数据值进行泛化时产生的信息损失量。它的表达公式为：

其中，表示数据值在泛化树中包含子孙的数量，表示属性A中值域包含值的数量。

# 主要研究内容



图 1 基于人工智能的隐私保护模型

本文主要研究内容为基于人工智能的隐私保护模型，针对攻击者通过非敏感信息，构造推断攻击模型去推断敏感信息的攻击方式，在给定一个黑盒推断攻击模型时，拟首先抓取该攻击模型的特征作为深度学习的输入数据，比如时间复杂度，需要非敏感数据的数量和种类等，然后用大量的过往数据测试来记录推断攻击的准确性，最后通过深度学习模型来生成噪音，以最小化推断成功率和最大化数据可用性。

该模型大致工作流程如图所示，首先，将非敏感数据输入到攻击者的推断攻击模型，提取出攻击模型的特征，计算出攻击模型推断敏感数据的准确性；然后将这两部分输出和非敏感数据作为深度学习模型的输入，作为模型为非敏感数据添加噪声的依据；最后模型的两个优化目标是：1.最小化推断攻击的准确性2.最小化非敏感数据的噪音，并以此来实现针对相应攻击，隐私数据保护和数据完整性之间的平衡。其中1的计算拟通过攻击模型再次处理深度学习模型输出的加噪非敏感数据计算推断准确性，2拟通过相应的信息可利用性标准来计算。

针对上述流程，需要明确具体的推断攻击模型，及其所作用的非敏感数据集和相对应的敏感数据，来进一步明确深度学习模型的输入；同时需要针对不同类型的非敏感数据研究适用的噪声添加方法，来构建模型。

其中研究的重点是深度学习模型的实现，包括考虑模型为不同类型数据添加噪声所应使用何种算法，是否可以满足相应的脱敏原则，以及选用适合当前问题领域的机器学习方法来实现模型。如文献[20]为保护训练数据集中用户的敏感信息，提出一个基于深度卷积生成式对抗网络(deep convolutional generative adversarial networks，DCGAN)反馈的深度差分隐私保护方法。该方法在深度网络参数优化计算时结合差分隐私理论添加噪声数据，基于差分隐私与高斯分布可组合特点，计算深度网络每一层的隐私预算，在随机梯度下降(stochastic gradient descent，SGD)计算中添加高斯噪声使之总体隐私预算最小；用DCGAN生成数据选取可能得到的最优结果。

# 已完成的研究工作

调研了基本的脱敏算法如：匿名算法和噪音机制，以及有关的脱敏原则如：匿名原则及差分隐私；此外，提出了基于人工智能隐私保护模型的大致流程，分析了美国人口普查等基本数据集。

# 研究方案及进度安排，预期达到的目标和取得的研究成果

## 研究方案

### 脱敏算法和原则的研究

针对不同的数据类型，寻找适合的脱敏算法和原则，尝试采用适当的噪音添加机制，力求与差分隐私等相应的脱敏原则相结合。为深度学习模型添加噪声提供相应的思路和算法支持。

### 基于人工智能的隐私保护模型的研究

首先，收集和选取数据集以及针对相应非敏感数据的推断攻击模型，实现测试攻击模型准确性和提取攻击模型特征的方法；其次，去了解不同深度学习模型的擅长的领域，针对不同类型的非敏感数据，尝试采用合适的深度学习模型，探求为不同类型数据添加噪声的方法，通过深度学习模型来为非敏感数据添加噪声。

## 预期达到的目标和取得的研究成果

通过本文提出的基于人工智能的隐私数据保护模型，拟实现针对不同数据类型来加入不同程度噪音的效果，针对具体攻击推断模型，通过深度学习模型对非敏感数据添加噪音，最小化推断攻击准确性和非敏感数据的噪音。

## 进度安排

|  |  |
| --- | --- |
| 时间 | 进度 |
| 2018.07~2018.08 | 调研隐私数据保护算法及脱敏原则 |
| 2018.09~2018.10 | 调研推断攻击模型以及其针对的非敏感数据和敏感数据集，实现提取攻击模型特征的方法和测量推断敏感数据准确性的方法，以明确深度学习模型的输入 |
| 2018.10~2018.12 | 了解适用不同领域的机器学习方法，学习深度学习模型构建和训练的方法，了解添加噪声的方法，针对非敏感数据集，确定模型如何对非敏感数据添加噪声 |
| 2019.01~2019.03 | 针对本文研究问题的输入，以及噪声添加方式，尝试构建合适的深度学习模型，同时还需要引入攻击模型，来计算设置的相应优化目标 |
| 2019.04~2019.05 | 训练深度学习模型，评价模型最小化噪声和推断准确性的效果 |
| 2019.06~2019.07 | 论文撰写，准备答辩 |

# 为完成课题已具备和所需的条件和经费

已具备条件：基于人工智能的隐私保护模型大致框架

# 预计研究过程中可能遇到的困难和问题，以及解决的措施

难点1：如何明确深度学习模型的输入，为构建合适的神经网络等后续任务打好基础。解决的措施：与导师交流，明确攻击模型的类型及其作用范围，调研相应领域的推断攻击模型以及其针对的非敏感数据和敏感数据集，并实现提取攻击模型特征的方法和测量推断敏感数据准确性的方法。

难点2：在明确非敏感数据集的前提下，如何选择深度学习模型添加噪声的方法和机制，如何结合相应的脱敏原则。解决的措施：首先理解不同噪声机制的原理和使用范围，针对不同类型的数据，寻求通过深度学习模型添加噪声的有效方法。

难点3：如何构建包含深度学习模型和攻击模型的隐私保护模型，以及训练深度学习模型的方法、调参方法等。解决的措施：了解适用不同领域的机器学习方法，学习深度学习模型构建和训练的方法。

# 主要参考文献

1. 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(06): 647-657.
2. 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述[C]. 中国计算机学会人工智能会议, 2013.
3. 陈天莹, 陈剑锋. 大数据环境下的智能数据脱敏系统[J]. 通信技术, 2016, 49(7): 915-922.
4. 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.
5. Walker S J, Viktor M-S, Kenneth C. Big Data: A Revolution That Will Transform How We Live, Work, and Think[J]. International Journal of Advertising, 2014.
6. 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014(4): 927-949.
7. Barbaro M, Zeller T. A face is exposed for AOL Searcher no.4417749[J]. New York Times, 2006.
8. Cox L. Suppression Methodology and Statistical Disclosure Control[J]. Publications of the American Statistical Association, 1980, 75(370): 377-385.
9. 周水庚, 李丰, 陶宇飞,等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5):847-861.
10. JUSTIN B， VITALY S． The cost of privacy: destruction of data-mining utility in anonymized data publishing［C］ ∥ ACM SIGKDD International Conference on Knowledge Discovery and Data M ining ．New York: ACM， 2008: 70- 78．
11. Latanya S. k-Anonymity: a model for protecting privacy[J]. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 2002, 10(05):557-570.
12. Dalenius T. Finding a needle in a haystack-or identifying anonymous census record[J]. Journal of Official Statistics, 1986, 2(3):935-936.
13. Dwork C. Differential privacy[J]. Lecture Notes in Computer Science, 2006, 26(2):1-12.
14. Han C, Wang K. Sensitive Disclosures under Differential Privacy Guarantees[C] . IEEE International Congress on Big Data. IEEE, 2015:110-117.
15. Domingo-Ferrer J, Torra V. A Critique of k-Anonymity and Some of Its Enhancements[C]// International Conference on Availability, Reliability and Security. IEEE, 2008:990-993.
16. Dwork C, Naor M, Vadhan S. The Privacy of the Analyst and the Power of the State[C]// Foundations of Computer Science. IEEE Computer Society, 2012:400-409.
17. Sweeney L. Datafly: A system for providing anonymity in medical data[M]. Database Security XI. Springer US, 1998: 356-381.
18. Xiao X, Tao Y. Personalized privacy preservation[C]. Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006: 229-240.
19. Latanya S. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05):571-588.
20. 毛典辉, 李子沁, 蔡强,等. 基于DCGAN反馈的深度差分隐私保护方法[J]. 北京工业大学学报, 2018(6).