·网络空间安全·　　文章编号：1000-3428(2022)04-0143-05　　文献标志码：A　　中图分类号：TP309

PPDM中面向k-匿名的MI Loss评估模型

谷青竹，董红斌

（武汉大学 国家网络安全学院，武汉 430000）

摘　要：隐私保护数据挖掘(PPDM)利用匿名化等方法使数据所有者在不泄露隐私信息的前提下，安全发布在数据挖掘中有效可用的数据集。k-匿名算法作为PPDM研究使用最广泛的算法之一，具有计算开销低、数据形变小、能抵御链接攻击等优点，但是在一些k-匿名算法研究中使用的数据可用性评估模型的权重设置不合理，导致算法选择的最优匿名数据集在后续的分类问题中分类准确率较低。提出一种使用互信息计算权重的互信息损失(MI Loss)评估模型。互信息反映变量间的关联关系，MI Loss评估模型根据准标识符和标签之间的互信息计算权重，并通过Loss公式得到各个准标识符的信息损失，将加权后的准标识符信息损失的和作为数据集的信息损失，以弥补评估模型的缺陷。实验结果证明，运用MI Loss评估模型指导k-匿名算法能够明显降低匿名数据集在后续分类中的可用性丢失，相较于Loss模型和Entropy Loss模型，该模型分类准确率提升了0.73%~3.00%。

关键词：隐私保护数据挖掘；k-匿名算法；数据可用性；分类准确率；MI Loss评估模型

开放科学(资源服务)标志码(OSID)：

中文引用格式：谷青竹，董红斌.PPDM中面向k-匿名的MI Loss评估模型[J].计算机工程，2022，48(4)：143-147.

英文引用格式：GU Q Z，DONG H B.MI Loss evaluation model for k-anonymity in PPDM[J].Computer Engineering，2022，48(4)：143-147.

MI Loss Evaluation Model for k-Anonymity in PPDM

GU Qingzhu，DONG Hongbin

（School of Cyber Science and Engineering，Wuhan University，Wuhan 430000，China）

【Abstract】Privacy Preserving Data Mining（PPDM）uses methods such as anonymization to allow data owners to safely publish data sets that are effectively available in data mining without revealing private information. The k-anonymity algorithm，one of the most widely used algorithms in PPDM research，has the advantages of low computational overhead，small data deformation，and resistance to link attacks.However，in some studies on k-anonymity algorithms，the weight settings of the data utility evaluation model used by the algorithm are unreasonable，which leads to the low classification accuracy of the optimal anonymous data set selected by the algorithm.Mutual Information（MI）reflects the relationship between variables.The MI Loss evaluation model uses the mutual information between the quasi-identifier and the label to calculate the weight.The information loss of each quasi-identifier is obtained through the Loss formula，and the sum of all weighted quasi-identifier information losses is taken as the information loss of the data set，which makes up for the shortcomings of the existing evaluation model. Experiments show that using the MI Loss evaluation model to guide the k-anonymity algorithm can significantly reduce the utility loss of anonymous data sets in subsequent classification problems.The classification accuracy of the proposed model shows an improvement of 0.73%~3.00% compared with the accuracies of the Loss and Entropy Loss models.

【Key words】Privacy Preserving Data Mining（PPDM）；k-anonymity algorithm；data utility；classification accuracy；MI Loss evaluation model

DOI：10. 19678/j. issn. 1000-3428. 0061707

0　概述

k-匿名作为隐私保护数据挖掘(Privacy Preserving Data Mining，PPDM)研究中的一项关键技术，具有计算开销低、数据形变小、能抵御链接攻击等优点[1]。在一些追求高数据可用性的k-匿名算法中，通常会提出度量数据集信息损失的数据可用性评估模型，但这些评估模型在属性的可用性权重设置上存在缺陷，导致

基金项目：国家自然科学基金“计算机免疫智能的连续免疫应答机制及其应用研究”(61877045)。

作者简介：谷青竹(1997—)，女，硕士研究生，主研方向为隐私保护数据挖掘；董红斌，教授。

收稿日期：2021-05-20　　**修回日期：**2021-07-10　**E-mail：**qqqzzz0519@163.com

算法选择出的最优匿名数据集在后续的分类问题中表现较差。

目前数据集的发布共享越来越普遍，为避免在数据使用过程中造成的隐私泄露，研究人员希望通过修改原始数据集来保护隐私信息。但是，转化原始数据可能会导致信息损失和数据挖掘结果不准确。为此，隐私保护数据挖掘技术旨在保护隐私信息的前提下，最大化数据可用性，使得转化后的数据仍然可以被有效用于数据挖掘[2-3]。隐私保护数据挖掘是一项在数据挖掘整个过程中实现隐私保护的研究，从数据的产生、发布，以及到挖掘的各个阶段，都需要依靠PPDM技术防止隐私泄露。数据发布阶段的PPDM技术主要包括匿名化、扰动和加密，这些技术适用的场景各不相同。匿名化主要用在数据集被公开发布的场景下，扰动多用于防止统计泄露的数据库查询中，加密则是不同地点的数据所有者在协同计算时应用[1]。

本文构建一种面向k-匿名的互信息损失（Mutual Information Loss，MI Loss）评估模型，该模型利用互信息计算属性的可用性权重，使与标签相关性高的属性可被较低程度泛化，从而减少关键属性的信息损失，以提升匿名数据集在后续分类问题中的准确率。

1 匿名化

匿名化是指在匿名化算法的指导下对原始数据集执行一系列数据转化操作，使其满足相应的匿名化模型。其中，只有数据转化操作是对数据集的具体转化，而匿名化算法和匿名化模型只提供理论的指导和约束。

1.1 k-匿名模型

k-匿名模型由SWEENEY等[4]于1998年提出，由于能抵御链接攻击，已经成为使用最广泛的匿名化模型之一。k-匿名模型要求在发布的数据集中，每条记录至少要与$k-1$条记录拥有相同的准标识符值，这样攻击者推测一条记录所关联个人的可能性被降低，达到了匿名化的效果。准标识符是指可以联合起来标识一个个人的属性，例如当用<性别，职业，邮编>这样一个属性组合就能识别出一个个体时，“性别”“职业”“邮编”就被称为是准标识符。参数k越大，数据失真程度越高，隐私保护的效果越好，但相应的信息损失也越大。经过二十多年的发展，在k-匿名模型的基础上，l-diversity[5]、t-closeness[6]、(alpha，k)-anonymity等[7]众多匿名化模型被相继提出。这些模型通常使用一个或多个隐私参数来降低发布数据中一个个体被重标识的风险。

1.2 k-匿名算法

现有的k-匿名算法大致可以分为两类：一类是基于泛化的k-匿名算法[8]；另一类是基于微聚合的k-匿名算法[9]。泛化是指使用一个更一般化的值替换原始的精确值[10]。在泛化操作中，数值型准标识符往往被泛化成一个区间，例如年龄17可以被泛化成(15，20]；而类别型准标识符则需要定义泛化层级关系树，树的叶子节点是原始值，越往上层值的泛化程度越大。职业的泛化层级树如图1所示。抑制（suppression）是程度最高的泛化，指使用特定符号（例如*，&等）来代替原始值，使原始值失效。

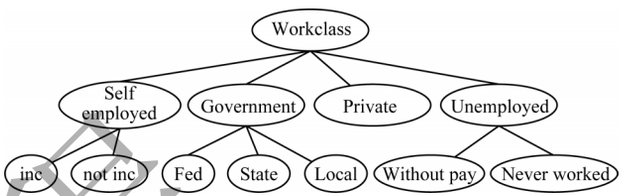

图1 职业的泛化层级关系树

Fig.1 Generalized hierarchical tree of occupation

由于基于泛化的k-匿名算法更适合运用于类别型准标识符，对于数值型准标识符会丢失更多的语义[11]，一些研究便将SDC（Statistical Disclosure Control）中的微聚合技术运用到匿名化算法中[12]。微聚合的主要思想是给定一个参数k，然后将数据集划分成多个组，每个组至少包含k条记录，组内记录最大程度的相似，而组间记录最大程度的不同。最后每个组内原始的准标识符值会被替换成所在组的准标识符质心[13]。因此，每条记录会与$k-1$条记录相似，满足了k-匿名模型要求。

2 评估模型研究现状

k-匿名算法在保证数据集安全发布的同时，会不可避免地导致数据可用性降低。根据可用性指导匿名化过程，得出可用性最高的匿名化数据集（即最优匿名数据集），研究人员提出了评估数据可用性模型[14-15]。在可用性模型中，数据集的信息损失等于所有加权后的准标识符信息损失之和。这里的权重是指准标识符的可用性权重，权重越大，该准标识符在匿名化时被泛化程度越低。

目前，可用性权重的设定方式主要分为三类：一类是假设所有准标识符的可用性权重相等，如文献[16]提出的评估模型Loss，显然这种方式并不符合使用数据时的实际情况；另一类是由使用数据的人指定，如文献[17]提出评估模型UC，这类方式很大程度上依赖于操作者的个人能力，且不具有普遍适用性；最后一类是依靠特定指标自动计算出权重因子，如文献[18]提出的Entropy Loss模型，首先计算出准标识符的信息熵，然后将信息熵标准化之后作为可用性权重。但是信息熵仅描述了属性所包含的信息量，并不能反映属性在后续挖掘中的重要程度，有些信息熵低的属性反而更应该被较低程度泛化。例如性别，虽然信息熵低，但是在分析一个人的兴趣爱好时却十分重要。

上述评估模型中可用性权重的设置方式均存在缺陷，直接影响了k-匿名算法选取最优解。为此，本文提出使用互信息计算可用性权重的互信息损失评估模型。该模型可适用于PPDM研究，指导k-匿名算法在后续分类问题中选择出可用性更高的匿名数据集。

3　MI Loss模型

MI Loss评估模型通过准标识符和标签之间的互信息计算可用性权重，利用Loss公式计算各个准标识符的信息损失，最后将所有加权后的准标识符信息损失之和作为数据集的信息损失。

3.1　MI Loss评估模型

互信息通常被用来描述两个变量之间的相关性，若两个变量之间不存在相关性，则它们的互信息为零，否则它们的互信息大于零。通过计算数据集中每个准标识符与标签的互信息，可以判断出它们与标签的相关性大小[19]。若相关性大，则说明该准标识符应该被较低程度泛化，以保留其后续在数据挖掘中的可用性。因此，相较于信息熵，互信息反映了准标识符对标签信息量的影响，更适合作为k-匿名过程中计算准标识符的可用性权重指标。

利用互信息计算准标识符的可用性权重方法如下：首先计算出每个准标识符与标签之间的互信息，然后将所有互信息值标准化得到权重。如式(1)所示，假设在数据集D中，准标识符的个数为q，依次表示为$X_1, X_2, \cdots, X_q$，标签表示为Y，将式(2)、式(3)代入式(1)计算出准标识符X_i与标签Y之间的互信息$\mathrm{MI}(Y, X_i)$，最后通过标准化式(4)得到准标识符X_i的可用性权重w_i。

$$M_{\mathrm{MI}}(Y, X_i) = H(Y) - H(Y|X_i), i \in \{1, 2, \cdots, q\} \tag{1}$$

$$H(Y) = -\sum_{y \in Y} p(y)\log_a(p(y)) \tag{2}$$

$$H(Y|X_i) = -\sum_{x \in X_i} \sum_{y \in Y} p(x, y)\log_a(p(y|x)) \tag{3}$$

$$w_i = \frac{M_{\mathrm{MI}}(Y, X_i)}{\sum\limits_{i \in \{1, 2, \cdots, q\}} M_{\mathrm{MI}}(Y, X_i)} \tag{4}$$

在匿名化的过程中，准标识符的原始值会被替换为经过泛化的值，MI Loss评估模型使用文献[16]提出的信息损失公式Loss来估计准标识符的信息损失$\mathrm{Loss}(X_i)$。Loss公式用一个值被泛化后的区间度量除以其泛化前的区间度量作为该值的信息损失。对于数值型属性，区间度量为泛化区间最大值和最小值之差，对于类别型属性，区间度量为泛化层级关系树的同层节点数目。

根据所有准标识符X_i的信息损失$\mathrm{Loss}(X_i)$和其相应的可用性权重w_i，MI Loss模型计算数据集D的信息损失公式如下：

$$M_{\mathrm{MI}}\mathrm{Loss}(D) = \sum_{1 \leqslant i \leqslant q} w_i \cdot \mathrm{Loss}(X_i) \tag{5}$$

式(5)可用于指导k-匿名算法寻找最优解。

3.2　基于MI Loss模型的k-匿名算法

基于MI Loss模型的k-匿名算法流程如图2所示。在搜索阶段，依据一定规则转化原始数据，并将所有满足k-匿名模型的数据集加入解空间。建立解空间后，根据MI Loss评估模型计算每个解的信息损失，最后从中选出信息损失最低的一个匿名数据集作为最优解输出。若原始数据集体积过大，维数过多，则搜索阶段可能会遭遇维数灾难[20]，这时要依靠MI Loss模型计算当前数据集的信息损失来剪枝，达到降低计算量及优化算法性能的目的。

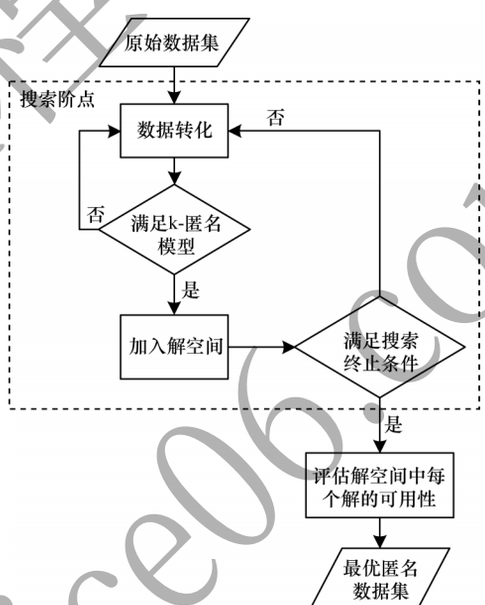

图2　基于MI Loss模型的k-匿名算法流程

Fig.2　Procedure of k-anonymity algorithm based on MI Loss model

4　实验评估

4.1　实验目的及步骤

为证明在PPDM的研究中利用MI Loss评估模型指导k-匿名算法得到的最优匿名数据集能有更好的分类表现，本文将其与Loss模型[16]和Entropy Loss模型[18]进行了对比。比较3种模型选择出的最优匿名数据集在分类问题中的准确率，分类准确率越高则说明对应的匿名数据集在分类方面的可用性降低越少[21]。

本文实验中使用的数据集是UCI Machine Learning Repository中的Adult数据集，Adult数据集经常被用来研究二分类问题和隐私保护机制，通过一个人的年龄、受教育年限、性别等属性预测其薪资是否超过50K，该数据集中包含了14个特征和1个标签。为最大程度地模拟真实环境，保证数据安全，实验从14个特征中选出11个作为准标识符，准标识符的具体描述如表1所示。

表1 Adult数据集中的准标识符

Table 1 Quasi-identifiers in Adult dataset

属性	类型
age	数值型
education-num	数值型
capital-gain	数值型
hours-per-week	数值型
race	类别型
relationship	类别型
workclass	类别型
country	类别型
marital Status	类别型
occupation	类别型
sex	类别型

本文在安装了JDK 11.0.7的Windows10环境中进行了实验，使用开源的匿名化软件ARX3.8.0[22]对数据匿名化，该工具可以由用户自定义匿名化模型、数据转化规则以及评估数据可用性的模型。

本文实验步骤如下：

步骤1 预定义匿名化模型及规则。在ARX中设置匿名数据集须满足的匿名模型为k-匿名，k分别取2、3、5、10，并且针对不同类型的数据设置了相应的数据转化规则和泛化层级关系。数值型准标识符的原始值会被转化为泛化区间端点的算术平均值；类别型准标识符则会根据预定义的泛化层级关系树来转化。

步骤2 搜索所有解，建立解空间。ARX会对所有准标识符按照泛化层级关系自底向上逐级泛化，每遇到一个满足k-匿名的匿名数据集就将其加入解空间，直至满足搜索终止条件。

步骤3 使用3种评估模型分别对解空间中的匿名数据集进行评估。Loss模型、Entropy Loss模型和MI Loss模型中准标识符的可用性权重如表2所示。

表2 3种评估模型中属性的可用性权重

Table 2 Utility weights of attributes in three evaluation models

属性	Loss模型	Entropy Loss模型	MI Loss模型
age	1/11	0.229 4	0.035 5
education-num	1/11	0.118 3	0.068 3
capital-gain	1/11	0.035 0	0.297 6
hours-per-week	1/11	0.140 5	0.037 5
race	1/11	0.032 2	0.022 5
relationship	1/11	0.086 9	0.164 3
workclass	1/11	0.066 5	0.028 1
country	1/11	0.038 1	0.019 7
marital Status	1/11	0.074 0	0.182 8
occupation	1/11	0.141 9	0.056 5
sex	1/11	0.037 0	0.087 0

3种模型利用不同的权重通过式(5)计算数据集的信息损失，最后从所有满足k-匿名模型的数据集中搜索出信息损失最低的一个作为该模型得到的最优匿名数据集输出。

步骤4 当k取不同值时，用3种模型得到的最优解分别训练分类器，并比较分类器的分类准确率。分类准确率越高，说明对应的匿名数据集在分类问题中的信息损失越低，相应的可用性评估模型也越能有效地指导k-匿名算法选择出最优解。

4.2 实验结果与分析

使用线性回归在3种模型选择出的最优匿名数据集上分别训练分类器，并采用三折交叉验证计算分类器的准确率，结果如图3所示。当k=0时，即为原始数据集，数据没有形变，因此训练出的分类器的分类准确率最高为89.23%，这也是所有分类器准确率的上界。随着k值的增加，满足k-匿名模型的难度也随之增大，数据的形变程度加大，因此分类器的表现都随着数据可用性的降低有所下降，但不论k取何值，用MI Loss评估模型选择得到的最优解训练出的分类器分类准确率始终最高，比使用其他两种模型高0.73%~3.00%，而且随着k值的增加，MI Loss模型的优势愈加明显。这说明在k-匿名算法中，用MI Loss模型指导匿名化过程并选择最优解，能显著降低最优匿名数据集在分类问题中的可用性丢失，从而取得更高的分类准确率。

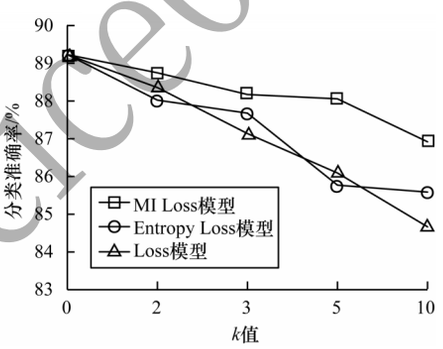

图3 3种模型的分类器准确率

Fig.3 Classification accuracy of three models

5 结束语

本文分析了隐私保护数据挖掘研究中k-匿名算法所采用的数据可用性评估模型在可用性权重设置上的缺陷，提出一种使用互信息计算权重的MI Loss评估模型。该模型将准标识符和标签之间的互信息标准化后作为权重，并与Loss模型、Entropy Loss模型进行对比。实验结果表明，MI Loss模型选择出的最优解在分类问题中有更高的可用性，分类准确率优于另外两种模型。本文仅对MI Loss模型选择出的最优解在分类问题中的表现进行了实验分析，下一步将研究该模型在回归问题中的可用性。

参考文献

[1] SHAH A，GULATI R. Privacy preserving data mining：techniques，classification and implications—a survey[J]. International Journal of Computer Applications，2016，137(12)：40-46.

[2] MENDES R，VILELA J P. Privacy-preserving data mining：methods，metrics，and applications[J]. IEEE Access，2017，5：10562-10582.

[3] SHILPA R，RAJKOT V E C. Survey on privacy preserving data mining techniques[J]. International Journal of Engineering Research and Technical Research，2020，9(6)：265-273.

[4] SWEENEY L. Datafly：a system for providing anonymity in medical data[C]//Proceedings of the 11th International Conference on Database Security. [S. l.]：Chapman & Hall，Ltd.，1997：356-381.

[5] MACHANAVAJJHALA A，GEHRKE J，KIFER D，et al. L-diversity：privacy beyond k-anonymity[C]//Proceedings of the 22nd ACM International Conference on Data Engineering. New York，USA：ACM Press，2006：24.

[6] LI N H，LI T C，VENKATASUBRAMANIAN S. T-closeness：privacy beyond k-anonymity and l-diversity[C]//Proceedings of the 23rd IEEE International Conference on Data Engineering. Washington D. C.，USA：IEEE Press，2007：106-115.

[7] CHI-WING R，LI J Y，FU A W C，et al.(α，k)-anonymity：an enhanced k-anonymity model for privacy preserving data publishing[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York，USA：ACM Press，2006：754-759.

[8] SWEENEY L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty，Fuzziness and Knowledge-Based Systems，2002，10(5)：571-588.

[9] MONEDERO D R，MEZHER A M，COLOMÉ X C，et al. Efficient k-anonymous microaggregation of multivariate numerical data via principal component analysis[J]. Information Sciences，2019，503：417-443.

[10] XU Y，MA T H，TANG M L，et al. A survey of privacy preserving data publishing using generalization and suppression[J]. Applied Mathematics & Information Sciences，2014，8(3)：1103-1116.

[11] DOMINGO-FERRER J，TORRA V. Ordinal，continuous and heterogeneous k-anonymity through microaggregation[J]. Data Mining and Knowledge Discovery，2005，11(2)：195-212.

[12] PALLARÈS E，REBOLLO-MONEDERO D，RODRÍGUEZ-HOYOS A，et al. Mathematically optimized，recursive prepartitioning strategies for k-anonymous microaggregation of large-scale datasets[J]. Expert Systems with Applications，2020，144(8)：113-126.

[13] ABIDI B，BEN YAHIA S，PERERA C. Hybrid microaggregation for privacy preserving data mining[J]. Journal of Ambient Intelligence and Humanized Computing，2020，11(1)：23-38.

[14] KOHLMAYER F，PRASSER F，ECKERT C，et al. Flash：efficient，stable and optimal K-anonymity[C]//Proceedings of 2012 International Conference on Privacy，Security，Risk and Trust and 2012 International Confernece on Social Computing. Washington D. C.，USA：IEEE Press，2012：708-717.

[15] KOHLMAYER F，PRASSER F，KUHN K A. The cost of quality：implementing generalization and suppression for anonymizing biomedical data with minimal information loss[J]. Journal of Biomedical Informatics，2015，58：37-48.

[16] IYENGAR V S. Transforming data to satisfy privacy constraints[C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York，USA：ACM Press，2002：279-288.

[17] LOUKIDES G，GKOULALAS-DIVANIS A. Utility-preserving transaction data anonymization with low information loss[J]. Expert Systems with Applications，2012，39(10)：9764-9777.

[18] BHATI B S，IVANCHEV J，BOJIC I，et al. Utility-driven k-anonymization of public transport user data[J]. IEEE Access，2021，9：23608-23623.

[19] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transactions on Neural Networks，1994，5(4)：537-550.

[20] AGGARWAL C C. On k-anonymity and the curse of dimensionality[C]//Proceedings of International Conference on Very Large Data Bases. New York，USA：ACM Press，2005：901-909.

[21] RODRÍGUEZ-HOYOS A，ESTRADA-JIMÉNEZ J，REBOLLO-MONEDERO D，et al. Does k-anonymous microaggregation affect machine-learned macrotrends?[J]. IEEE Access，2018，6：28258-28277.

[22] PRASSER F，EICHER J，SPENGLER H，et al. Flexible data anonymization using ARX—current status and challenges ahead[J]. Software：Practice and Experience，2020，50(7)：1277-1304.

编辑　索书志