基于 Shannon 信息熵与 BP 神经网络的隐私数据度量与分级模型

俞艺涵，付钰，吴晓平
（海军工程大学信息安全系，湖北 武汉 430033）

摘 要：针对当前网络环境下由隐私数据识别困难问题所引出的隐私度量与分级需求，提出了一种基于 Shannon 信息熵与 BP 神经网络的隐私数据度量与分级模型。该模型从 3 个维度建立了两层隐私度量要素，基于数据集本身，利用 Shannon 信息熵为二级隐私要素定权，并由此计算数据集中各条记录在一级隐私度量要素下的隐私量；利用 BP 神经网络在不预设度量权值的情况下，输出隐私数据分级结果。实验表明，该模型能够在极低的误判率和较小的误判偏差下实现对隐私数据的度量与分级。

关键词：隐私安全；信息熵；BP 神经网络；隐私度量
中图分类号：TP301
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018286

Metric and classification model for privacy data based on Shannon information entropy and BP neural network

YU Yihan, FU Yu, WU Xiaoping
Naval University of Engineering, Wuhan 430033, China

Abstract: Aiming at the requirements of privacy metric and classification for the difficulty of private data identification in current network environment, a privacy data metric and classification model based on Shannon information entropy and BP neural network was proposed. The model establishes two layers of privacy metrics from three dimensions. Based on the dataset itself, Shannon information entropy was used to weight the secondary privacy elements, and the privacy of each record in the dataset under the first-level privacy metrics was calculated. The trained BP neural network was used to output the classification result of privacy data without pre-determining the metric weight. Experiments show that the model can measure and classify private data with low false rate and small misjudged deviation.

Key words: privacy security, information entropy, BP neural network, privacy metrics

1 引言

当前，移动互联网、大数据计算平台等信息产业的飞速发展给人们生活带来了极大便利，众多服务型互联网产业应运而生。这些产业在为用户提供服务的同时，海量的数据信息在期间流转。以网络约车平台为例，用户的个人信息、行程信息、司机信息、车辆信息等数据在用户、平台和司机之间不断交互，交互的过程中数据往往以满足服务为首要目的进行呈现，而数据安全常常被忽略。而在海量数据信息流转间又往往蕴含着巨大的信息价值，其中不缺乏涉及与隐私相关的数据信息[1]，即使仅仅是网络约车平台上的一则简单评论都可能造成用户个人隐私的泄露，如何保证这类数据中隐私信息

收稿日期：2018–05–25；**修回日期**：2018–08–01
基金项目：国家自然科学基金资助项目（No.61100042）；国家社会科学基金资助项目（No.15GJ003-201）
Foundation Items: The National Natural Science Foundation of China (No.61100042), The National Social Science Foundation of China (No.15GJ003-201)

的安全是一个亟需解决的现实问题。

多年来，国内外众多学者已经就如何保护隐私数据做了大量的研究，在基于数据扰乱、数据匿名等策略下，提出了一些卓有成效的隐私保护模型和方法，例如 k-anonymity 模型[2]、l-diversity 模型[3]以及差分隐私保护技术[4-5]。这些隐私保护技术的提出与发展为隐私数据安全打下了坚实的基础，但在实际应用中仍受到隐私数据类型多、隐私应用场景复杂等问题的制约，其中，隐私数据的识别困难问题尤为突出。由于隐私是一个十分抽象的概念，在不同隐私场景与不同隐私主体的情况下，隐私的范畴存在极大差异，很难形成一套通用的隐私界定标准，这给隐私数据的识别造成了巨大的困难[6]。而当前，隐私信息的载体往往是海量流转在网络间的数据流，如若不能成功地在大数据环境中遴选出需要实施隐私保护的数据，而将隐私保护技术无差别地实施在整个网络数据流中将造成时间和空间 2 个维度上的巨大开销。对数据进行科学高效的隐私度量与分级是解决隐私识别困难问题的必要前提。

当前，国内外针对数据隐私度量问题已经有了许多卓有成效的研究成果。Li 等[7]利用 k-匿名模型提出一种基于计算敏感属性分布值的隐私度量方法，通过计算数据中敏感属性值的全局分布以及同一敏感属性在各个等价类中分布的差异程度来度量隐私泄露风险；Gkountouna 等[8]同样基于匿名理论，构建攻击者背景知识与匿名数据的二叉树图，通过贝叶斯理论推理构建出预测二叉树图，将其与隐私信息比较来度量隐私泄露的风险；Clauß 等[9]利用信息熵描述数据集中隐私信息的不确定度，在此基础上，Peng 等[10]用通信模型描述隐私保护的过程，用信息熵度量通信信宿中带有的信息量以此度量隐私泄露的风险，并利用条件熵对拥有背景知识攻击者的攻击进行隐私度量，构建了对应的隐私保护信息熵模型；在差分隐私保护中，则通常以差分隐私预算 ε 直观地度量隐私保护效果[11-12]。可以发现，当前国内外学者针对隐私度量问题的研究主要集中在对经隐私保护后的数据进行隐私泄露风险的度量上，而针对原始数据集自身原有隐私信息量的度量方法研究成果较少。

由于缺乏通用的隐私界定标准[13]，要对某条数据在隐私层面进行“是”与“否”的判定十分困难，一种可行的方法是通过制定某种度量与分级规则来代替隐私界定标准，将评估理论应用到隐私度量与分级中，即将对单条数据的隐私度量与分级问题转化为对数据集隐私状况的评估问题，通过选取需度量的隐私要素作为评估的指标，基于相关评估手段，以数据集总体隐私状况为标准对数据集中的单条数据进行隐私度量与分级。这样做在绕过了隐私界定标准不明确这一“壁垒”的同时，基于数据集对单条记录进行隐私度量与分级更能反映出某条记录在即时情况下的隐私重要程度，更能为隐私保护技术与策略的实施提供依据。但仍将面临以下 2 个关键问题：1) 由于隐私概念的宽泛性所带来的隐私度量要素种类多且复杂而引起的效率性问题；2) 由于隐私应用场景多样性以及隐私拥有者主观因素不确定性所造成的隐私度量要素定权困难问题。

基于此，本文在 3 个隐私维度下建立 2 层隐私度量要素的基础上，提出了一种无需事先设定隐私度量要素权重的隐私数据度量与分级模型。该模型通过 3 个隐私维度下 2 层隐私要素的设置，借助 Shannon 信息熵对二级隐私要素进行合理定权，并由此计算出一级隐私要素下的数据隐私量，实现对隐私要素的降维，随后借助 BP（back propagation）神经网络实现隐私数据的分级。

2　基础知识

2.1　Shannon 信息熵

信息熵[14]（information entropy）这个词是信息论之父香农（C.E.Shannon）从热力学中借用过来的，热力学中的热熵是表示分子状态混乱程度的物理量，香农则用信息熵的概念来描述信源的不确定度。

假设某系统 X 存在 n 种状态，记为 $X\{x_1, x_2, \cdots x_n\}$，$p(x_i)\,(i=1,2,\cdots,n)$ 表示状态 x_i 在系统 X 中出现的概率，则系统 X 的 Shannon 信息熵 $H(x)$ 定义为[15]

$$H(x)=-\sum_{i=1}^{n}p(x_i)\log\left(p(x_i)\right) \tag{1}$$

其中，$0\leqslant p(x_i)\leqslant 1$ 且 $\sum_{i=1}^{n}p(x_i)=1$，规定当 $p(x_i)=0$ 时，$0\log(0)=0$。

Shannon 信息熵理论认为，通过信息熵对信息的无序程度进行度量，信息的信息熵越大，表示信息的无序程度越高，其带有的信息量就越少；信息熵越小，信息的无序程度越低，其带有的信息量就越大。

2.2 BP 神经网络

BP 神经网络是一种按误差反向传播训练的多层前馈网络，其算法称为 BP 算法，该算法的基本思想是梯度下降法，利用梯度搜索技术，以期使网络的实际输出值和期望输出值的误差均方差为最小[16]。

BP 神经网络是一种多层网络，分为输入层、隐含层和输出层 3 个层次。各神经元与下一层的神经元采取全连接，同层神经元之间相互无连接。一个包含 4 个输入层神经元、5 个隐含层神经元及 3 个输出层神经元的 BP 神经网络具体结构如图 1 所示。

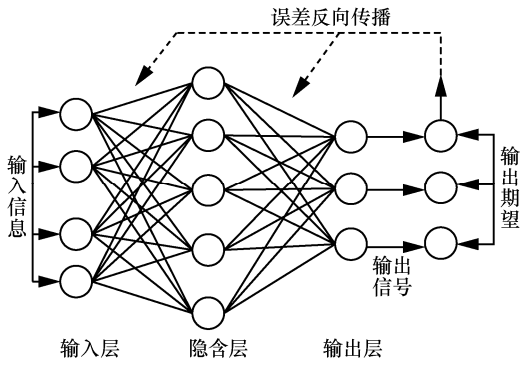

图 1 BP 神经网络结构

BP 神经网络最大的优势在于其能够学习和存储大量的输入输出关系，并且不用在事先揭示这种数学关系，包括信号的前向传播和误差的反向传播 2 个过程。正向传播时，输入信号通过隐含层作用于输出节点，经过非线性变换，产生输出信号，若实际输出与期望输出不相符，则转入误差的反向传播过程。误差反传是将输出误差通过隐含层向输入层逐层反传，并将误差分摊给各层所有单元，以从各层获得的误差信号作为调整各单元权值的依据。

3 基于 IE-BPDN 的隐私度量与分级模型

网络流量中的每一条记录都或多或少涉及隐私信息。而这些信息在隐私层面并没有明确的界定。一方面，相同一条记录中所蕴含的信息对于不同的用户来说，其带有的隐私量存在巨大差异；另一方面，以不同目的进行数据挖掘的隐私信息收集者来说，一条数据是否存在隐私价值也不尽相同。同时，各条记录中蕴含信息之间的关联性、信息的时效性、应用场景的多样性以及隐私拥有者对于隐私概念的主观性都是影响量化数据隐私的关键因素。若将以上所提因素全都考虑到数据隐私量中来，隐私度量要素的定权将十分困难，同时度量过程也是多维且复杂的，而针对特定需求进行度量的结果也不具有通用性。而本文的目的就是提出一种通用的数据隐私度量与分级模型，旨在事先不进行预设隐私度量要素权值的前提下，以较低计算开销实现对数据隐私量的合理度量与分级。

由此，本文提出了一种基于 IE-BPNN(information entropy-BP neural network)的隐私度量与分级模型，其基本框架分为隐私数据规则化、隐私要素度量与定权以及隐私分级 3 大模块。其基本思路是将网络流量中的数据以单位时间窗分割成各条记录，解析记录中所蕴含的隐私要素，并将其规则化表示；随后，通过计算不同记录之间相同二级隐私要素的信息熵确定其权重并依此计算各条记录在一级隐私要素下的隐私量，在对数据进行初步度量的同时，实现了对隐私要素的降维；最后由训练好的隐私分级 BP 神经网络得出各记录的最终隐私级别。

3.1 隐私数据规则化模块

隐私的概念十分宽泛，要对数据隐私实现准确度量与分级需涉及众多隐私要素，而过多的隐私要素将给隐私度量与分级的效率性提出挑战。根据相关文献[13,17]对于不同隐私个体对数据隐私不同方面敏感度的分析，本文从隐私内容（P_1）、隐私状态（P_2）、隐私详情（P_3）等 3 个维度选取以下具有代表性的要素作为隐私度量与分级的指标。具体如表 1 所示。

其中，在隐私详情（P_3）维度下，本文假设隐私度量与分级是针对用户位置轨迹进行的，由此选取了二级要素时刻和时间、坐标与地区分别对应一级要素精确信息与模糊信息。二级要素的选择可根据实际所需进行更替。记进行度量的数据集为 D，以单位时间窗将 D 分割成 n 条隐私记录，记为 $D\{d_1,\cdots,d_n\}$；以隐私内容（P_1）、隐私状态（P_2）、隐私详情（P_3）3 个维度解析各条记录，得到以 $L\{L_1,\cdots,L_8\}$ 为一级要素的二级要素 $L_a\{l_{a1},\cdots,l_{ab}\}$ 值，记为 $d_{ia}=\left\{L_a\left\{l_{a1}^*,\cdots,l_{ab}^*\right\}\right\}$，表示记录 d_i 在一级要素 L_a 下规则化后的记录值，$a=1,2,\cdots,8$，b 为 L_a 所对应的二级要素的维度，其中

$$l_{ab}^*=\begin{cases}1, & \text{记录中含有}l_{ab}\text{信息}\\ 0, & \text{记录中无}l_{ab}\text{信息}\end{cases} \tag{2}$$

举例来说，假设某条记录 d 在隐私内容（P_1）中，存在一级要素人口统计学信息（L_1）中除包含姓名

（l_{12}）、年龄（l_{12}）外没有其他二级要素的信息，则得到规则化表示后的记录值 $d=\{0,1,1,0,0,0,0,0\}$。

表 1　　隐私度量要素

维度（P）	一级要素（L）	二级要素（l）
隐私内容（P_1）	人口统计学信息（L_1）	证件（l_{11}）
		姓名（l_{12}）
		年龄（l_{13}）
		身高（l_{14}）
		体重（l_{15}）
		民族（l_{16}）
		籍贯（l_{17}）
	偏好信息（L_2）	兴趣特长（l_{21}）
		宗教信仰（l_{22}）
		政治倾向（l_{23}）
		购物偏好（l_{24}）
	财务信息（L_3）	信贷资料（l_{31}）
		收入（l_{32}）
		银行卡号（l_{33}）
		网银等虚拟账号（l_{34}）
	通信信息（L_4）	社交信息（l_{41}）
		家庭、工作地址（l_{42}）
		电子邮箱（l_{43}）
		手机、电话号码（l_{44}）
隐私状态（P_2）	静态信息（L_5）	家庭（l_{51}）
		工作（l_{52}）
		学历（l_{53}）
		社会关系（l_{54}）
	动态信息（L_6）	Cookies（l_{61}）
		位置移动信息（l_{62}）
		网络行为信息（l_{63}）
隐私详情（P_3）	精确信息（L_7）	时刻（l_{71}）
		坐标（l_{72}）
	模糊信息（L_8）	时间（l_{81}）
		地区（l_{82}）

3.2　基于信息熵的隐私要素度量模块

本文对 n 条记录在 3 个度量维度上的各个一级要素分别建立信息熵度量矩阵，假设某个一级要素 L_a 包含 b 个二级要素，其对于 n 条记录的度量结果通过建立 $n\times b$ 大小的二级要素信息熵度量矩阵来计算实现，具体步骤如下。

步骤 1　由 n 条记录经规则化后对应该一级要素 L_a 的 b 个二级要素记录值建立二级要素信息熵度量矩阵 $\boldsymbol{B}_{L_a}$

$$\boldsymbol{B}_{L_a}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1b} \\ \vdots & & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nb} \end{bmatrix}$$

矩阵中的 b_{ij} 是经规范化后第 i 个记录中对应第 j 个二级要素的记录值，取值为 0 或 1。

步骤 2　对矩阵 $\boldsymbol{B}_{L_a}$ 中的元素进行变换，如式(3)所示。

$$b_{ij}^{\ *}=\frac{b_{ij}}{\sum_{o=1}^{n} b_{oj}} \tag{3}$$

得到矩阵

$$\boldsymbol{B}_{L_a}^{*}=\begin{bmatrix} b_{11}^{*} & b_{12}^{*} & \cdots & b_{1b}^{*} \\ \vdots & & \ddots & \vdots \\ b_{n1}^{*} & b_{n2}^{*} & \cdots & b_{nb}^{*} \end{bmatrix}$$

步骤 3　根据信息熵式(1)对各个二级要素 j 计算其信息熵

$$E_j=-k\sum_{0=1}^{n} b_{oj}^{\ *}\ln(b_{oj}^{\ *}) \tag{4}$$

其中，$k=\dfrac{1}{\ln(n)}$，j 代表二级要素的个数，j=1, 2,$\cdots$,b，当 $b_{oj}^{\ *}=0$ 时，$E_j=0$。

步骤 4　二级要素 l_j 权重计算

$$W_{l_j}=\frac{1-E_j}{b-\sum_{j=1}^{b} E_j} \tag{5}$$

步骤 5　得到一级要素 L_a 对单条记录 d_i 的度量值

$$L_{d_{ia}}=\sum_{j=1}^{b} W_{l_j} b_{ij}^{\ *}, i=1,\cdots,n \tag{6}$$

$L_{d_{ia}}$ 的值越大，代表记录 d_i 中以一级要素 L_a 为度量标准的隐私量越大。

步骤 6　重复步骤 1~步骤 5，在该度量维度下求解单条记录 d_i 在各个一级要素下的隐私度量值，生成 d_i 在该维度的隐私度量值向量：$\boldsymbol{F}_o(d_i)=\{L_{d_{i1}},\cdots,L_{d_{ia}}\}$，$o=1,2,3$。

步骤 7　重复以上步骤，得到单条记录 d_i 在 3 个度量维度下的隐私度量值向量

$$\boldsymbol{F}_{d_i}=\{F_1(d_i),F_2(d_i),F_3(d_i)\}$$

3.3　基于 BP 神经网络的隐私数据分级模块

本文建立在 BP 神经网络下得到隐私数据的最终分级结果。设计神经网络层数为 3 层，输入层节点数设置为 b，使之与一级要素数量对应，分别对应 b 个一级要素的隐私度量值；输出层节点数为 3，将度量为最高隐私等级的输出定为（1,1,1），度量为最低隐私等级的输出定为（0,0,0），以此类推将输出向量分别对应 8 个隐私等级；以 singmoid 型正切函数 tansig 和 singmoid 型对数函数 logsig 分别作为隐含层和输出层的传递函数。

在每轮 BP 神经网络训练中，随机抽取训练数据集中 10%的记录，以 3.2 节中基于信息熵的隐私要素度量方法得到训练数据的隐私要素度量向量值，将其归一化后组成该轮训练样本。具体训练过程如图 2 所示。

3.4　基于 IE-BPDN 的隐私度量与分级模型的实现

基于 IE-BPDN 的隐私度量与分级模型的具体实现过程如图 3 所示。

模型的隐私度量结果输出由 3 个方面组成：1) 由隐私要素度量模块所得到的隐私度量值向量集合经筛选计算和直接存储生成的度量值集合，本文采用

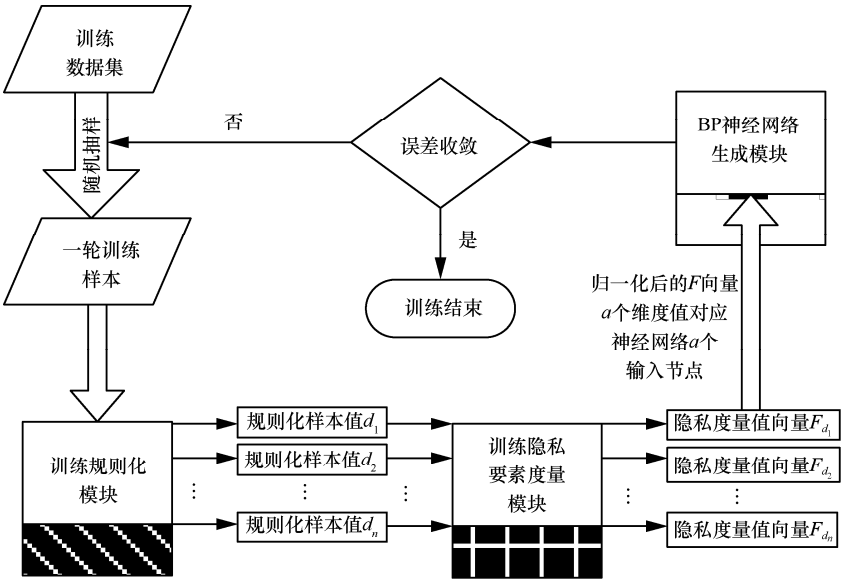

图 2　BP 神经网络分级模块训练

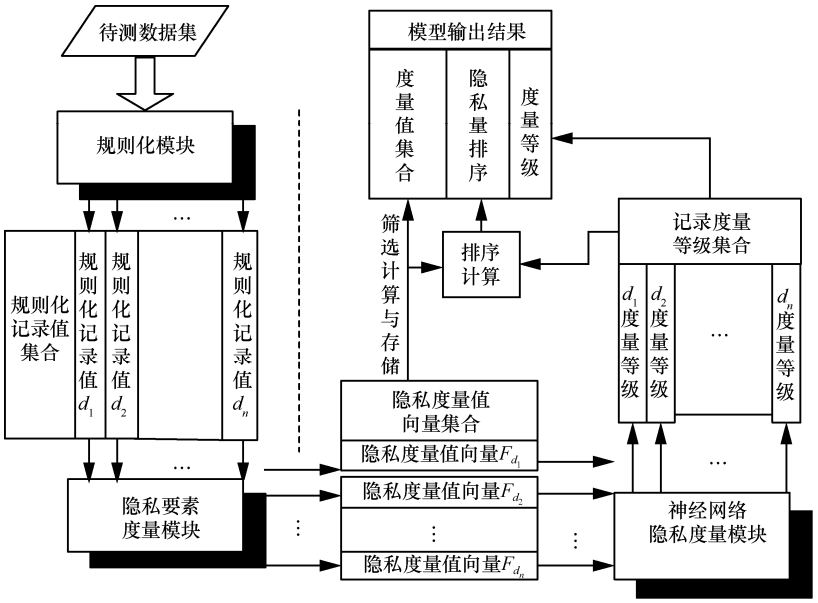

图 3　模型实现

筛选计算的方法是通过主成分分析法筛选出在 3 个隐私度量维度下主要反映记录隐私状况的一级要素，将其度量值相加得到记录在3个维度下的度量值，将其与隐私度量值向量一并存储作为输出；2) 由 BP 神经网络分级模块所得出的记录隐私度量等级；3) 记录在数据集中的隐私度排序。具体排序规则以记录度量等级为先，当度量等级相同时，依次比较 3 个维度下的隐私度量值大小，在 2 个或以上维度下度量值大的记录排序靠前。

模型的先进性主要表现为以下几个方面。

1) 隐私度量与分级要素的多维度。以本文所提隐私度量与分级要素为例，其包含 3 个层面的 8 个一级要素和 30 个二级要素。度量与分级要素的多维度使隐私数据在规则化的过程中，能够尽可能多地从不同方面将数据所蕴含信息的隐私量呈现出来，使隐私度量与分级的依据更为全面与合理。

2) 层次化的隐私量计算与呈现。以二级隐私要素下的隐私量计算一级隐私要素下的隐私量，即以二级隐私要素权重 W_{l_j} 计算一级隐私要素下的隐私量 $L_{d_{ia}}$，以此作为神经网络的输入向量并呈现在模型最终的输出结果中，令隐私量向量的维度由二级隐私度量要素维度数 b，降低至一级隐私度量要素维度 a（以本文所提隐私度量与分级要素为例，维度数 30→8），为高效可行的神经网络隐私分级打下基础的同时使输出结果更能反映出数据的隐私属性。

3) 度量与分级权值参数的零输入。不同的隐私要素在隐私度量与分级中的重要程度本就不相同，加之隐私应用场景的多样性与主观因素的差异性，使得很难预设隐私度量与分级中不同隐私要素的权重值。模型利用信息熵对二级隐私要素在数据级隐私度量与分级中重要程度的刻画，以及 BP 神经网络能在不揭示输入与输出之间数学关系的情况下进行学习和存储的能力，在权值参数零输入的情况下，实现了对数据高效合理的隐私度量与分级。

4) 度量并描述了单条记录在数据集中的隐私重要程度。在一些情况下，抛开数据集对某条记录进行单一的隐私度量并不能达到进行隐私度量的目的。模型在利用信息熵对二级隐私要素定权的过程中，已经在隐私层面刻画了单条记录蕴含的隐私信息相对于整个数据集蕴含的隐私信息的重要程度，并最终在模型输出中结合分级结果进行了呈现。记录的隐私等级最为直观地反映了模型对记录的隐私度量结果；记录在 3 个层面的度量值则体现了记录的隐私属性，即该条记录所带有隐私信息的具体形态趋势；记录的隐私度排名则直观地体现了该条记录在数据集中隐私量的贡献程度，即记录相对于数据集在隐私层面的重要程度。即通过基于 IE-BPDN 的隐私度量与分级模型，隐私保护人员可以直观地得到隐私数据的以下信息：

① 数据集中的某一条记录的隐私度量等级；

② 数据集的总体隐私度量信息；

③ 数据集中的某条记录相对整个数据集在隐私层面的重要程度；

④ 某 2 条记录的隐私重要程度比较；

⑤ 数据集中的某条记录在各维度、各隐私要素下的具体度量值。

以上信息，将使隐私保护方获得数据集隐私状况的定量评估，同时可以区分数据集中各条记录在隐私层面上的不同，即相对隐私保护对象区分记录在隐私层面上的重要程度，这些信息均为高效、准确、有针对性的隐私保护措施提供了实施依据。

4　测试与分析

本文所提模型中，隐私要素度量模块中的隐私要素定权与隐私量计算为简单的对数与乘法运算，为此本文将测试的重点放在基于 BP 神经网络的隐私数据分级模块上，主要进行神经网络训练测试与分级准确性测试。以 3.1 节中各项二级度量要素的有无为标准模拟生成 8 个隐私等级的数据集。其中，每个隐私等级包含 1 000 条隐私记录，共 8 000 条数据记录作为训练数据集。

4.1　神经网络训练

将训练轮次定为 1 200，学习效率为 0.1，目标误差为 0.000 1。根据公式 $N=\sqrt{I+O}+a$ 初选隐层节点数，I 和 O 分别为输入层节点数和输出层节点数，a 为调节参数，最终选定隐层节点数为 7。训练过程的误差曲线如图 4 所示，神经网络在 282 轮次训练后达到训练目标，可以满足模型需求。

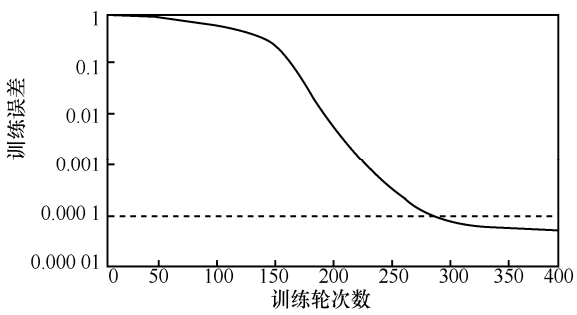

图 4　BP 神经网络训练误差曲线

4.2 模型准确性测试

由于本文所提模型输出结果的 3 个部分中，度量值集合由数据相应隐私要素的信息熵计算所得，隐私量排序是经度量值集合与度量等级综合对比产生，所以本文将模型的准确性测试重点放在了隐私分级的准确性上。首先提出误判偏差率的概念，用来刻画度量结果与实际隐私状况的偏差程度，计算式如式(7)所示。

$$\varepsilon = \sum_{i=1}^{k} \frac{\mathrm{e}^{|D_i - D_i'|}}{k} \tag{7}$$

其中，k 为进行误判偏差率计算的记录数，e 为自然对数，D_i 和 D_i' 分别代表序号为 i 的记录样本的实际隐私等级和预测隐私等级，$|D_i - D_i'| = 0,1,2,\cdots,7$。

在此基础上，本文从训练数据集中随机抽取 1 000 条记录作为测试样本，对所提隐私数据度量与分级模型进行如下分级准确性测试。测试结果如表 2 所示。

表 2 模型隐私分级准确性测试

样本隐私等级（P）	样本数量/个	误判数/个	误判率	误判偏差率
1	97	1	1.03%	2.80%
2	79	1	1.26%	3.44%
3	120	1	0.83%	2.27%
4	183	7	3.82%	12.95%
5	165	6	3.64%	9.88%
6	114	4	3.51	9.53%
7	98	1	1.02%	2.77%
8	144	1	0.69%	1.89%
1~8	1 000	22	2.20%	6.45%

由测试结果可以看出，本文所提的隐私数据度量与分级模型的分级总体准确率可达 97.8%，对于单个隐私等级的样本也能提供 96%以上的分级准确率。另一方面，从各个隐私等级的误判率可以发现，模型对处于隐私相对边界等级（隐私等级 0,1,2,6,7）的数据分级准确率高于处于隐私中间等级（隐私等级 4,6,7）的数据，这符合当数据提供的隐私信息区分度越大时，对其的隐私度量越容易这一现实情况。这也反映在各隐私等级的误判篇差率上，各等级误判率（E）、误判偏差率（ε）和 $\frac{E}{\varepsilon}$ 的变化趋势如图 5 和图 6 所示。

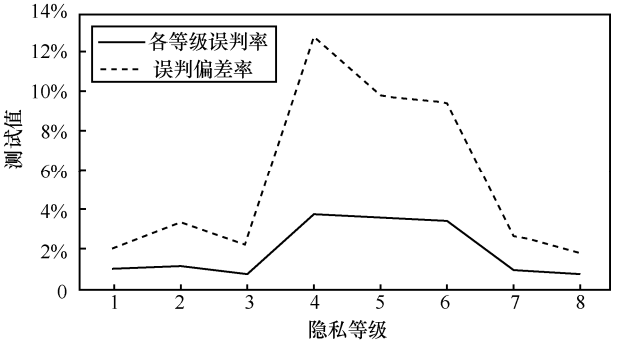

图 5 误判率与误判偏差率变化趋势

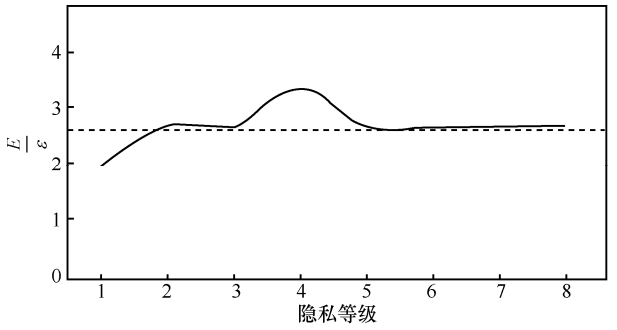

图 6 $\frac{E}{\varepsilon}$ 变化趋势

由图 5 可知，模型对于数据集整体的误判偏差程度，即误判偏差率的变化趋势与误判率相似。分析式(7)可知，ε 的值与 $|D_i - D_i'|$ 呈幂次关系，则 $\frac{E}{\varepsilon}$ 也应该与 $|D_i - D_i'|$ 呈幂次关系，而由图 6 可以看出 $\frac{E}{\varepsilon}$ 的变化趋于一条靠近自然对数 e 的直线。因此可以推断 $|D_i - D_i'|$ 的取值多为 0 和 1，这就说明模型在发生误判时，极少发生跨级误判，即发生误判的结果基本都在相邻隐私等级中。在误判率不高的情况下，模型这样的误判偏差程度是能够被接受的。

5 结束语

本文提出了一种基于信息熵和 BP 神经网络的隐私数据度量与分级模型。该模型借助信息熵对数据集隐私量进行分层计算，随后利用 BP 神经网络无需事先揭示输入与输出之间数学关系这一优势，能够在不预先设定隐私度量要素权重的情况下，实现对隐私数据的准确度量与分级。下一步可进行以下两方面工作：1) 研究海量网络流量环境下的数据自动化解析技术，为数据隐私要素值的自动化获取提出相应解决方案；2) 在本文所提隐私度量与分级模型基础上，研究 BP 神经网络的内在原理与优化

技术，进一步优化模型的效率性与准确性。

参考文献：

[1] CHANG V, SUN G, LI J. Guest editorial: security and privacy for multimedia in the internet of things (IoT)[J]. Multimedia Tools & Applications, 2018:1-2.

[2] SWEENEY L. K-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5):557-570.

[3] MACHANAVAJJHALA A, KIFER D, GEHRKE J. L-diversity: privacy beyond k -anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1):3.

[4] DWORK C, MCSHERRY F, NISSIM K. Calibrating noise to sensitivity in private data analysis[J]. Proceedings of the Vldb Endowment, 2006, 7(8):637-648.

[5] DWORK C, ROTH A. The algorithmic foundations of differential privacy[M].Boston: Now Publishers Inc. 2014

[6] GUZMAN J A D, THILAKARATHNA K, SENEVIRATNE A. Security and privacy approaches in mixed reality: a literature survey[J]. arXiv: 1802.05797v2[cs.CR], 2018.

[7] LI N, LI T, VENKATASUBRAMANIAN S. Closeness: a new privacy measure for data publishing[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(7):943-956.

[8] GKOUNTOUNA O, TERROVITIS M. Anonymizing collections of tree-structured data[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(8):2034-2048.

[9] CLAUß S, SCHIFFNER S. Structuring anonymity metrics[C]//The Workshop on Digital Identity Management. ACM, 2006: 55-62.

[10] PENG C G, DING H F, ZHU Y J, et al. Information entropy models and privacy metrics methods for privacy protection[J]. Journal of Software, 2016, 27(8): 1891-1903.

[11] ZHANG W J, HUI L I. A differentially-private mechanism for multi-level data publishing[J]. Chinese Journal of Network & Information Security, 2015, 1(1): 58-65.

[12] JORGENSEN Z, YU T, CORMODE G. Conservative or liberal? Personalized differential privacy[C]//International Conference on Data Engineering. IEEE, 2015:1023-1034.

[13] WAGNER I, ECKHOFF D. Technical privacy metrics: a systematic survey[J]. Computer Science, 2018, 51(3).

[14] SHANNON C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(4):379-423.

[15] BARNUM H, BARRETT J, CLARK L O, et al. Entropy and information causality in general probabilistic theories[J]. New Journal of Physics, 2012, 14(12): 129401.

[16] ZHANG G, ZHANG Z, HE Y. Review on BP neural network applied in the textile and clothing Field[J]. Shandong Textile Science & Technology, 2013.

[17] PHELPS J, NOWAK G, FERRELL E. Privacy concerns and consumer willingness to provide personal information[J]. Journal of Public Policy & Marketing, 2013, 19(1):27-41.

[作者简介]

俞艺涵（1992–），男，浙江金华人，海军工程大学博士生，主要研究方向为信息安全、隐私保护。

付钰（1982–），女，湖北武汉人，海军工程大学副教授，主要研究方向为信息安全、风险评估。

吴晓平（1961–），男，山西新绛人，海军工程大学教授、博士生导师，主要研究方向为信息安全、密码学。