一种基于权重属性熵的分类匿名算法

廖 军[1] 蒋朝惠[1] 郭 春[1] 平 源[2]

(贵州大学计算机科学与技术学院 贵阳 550000)[1] (许昌学院信息工程学院 许昌 461000)[2]

摘 要 为了在高效地保护数据隐私不被泄露的同时保证数据效用,提出了一种基于权重属性熵的分类匿名方法(Weight-properties Entropy for Classification Anonymous,WECA)。该方法在数据分类挖掘的特定应用背景下,通过信息熵的概念来计算数据集中不同准标识符属性对敏感属性的分类重要程度,选取分类权重属性熵比率最高的准标识符属性对分类树进行有利的划分,同时构建了分类匿名信息损失度量,在更好地保护隐私数据的前提下确保了数据分类效用。最后,在标准数据集上的实验结果表明,该算法在保证较少的匿名损失的同时具有较高的分类精度,提高了数据可用性。

关键词 隐私保护,分类匿名,权重属性熵,分类精度

中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.07.008

Classification Anonymity Algorithm Based on Weight Attributes Entropy

LIAO Jun[1] JIANG Chao-hui[1] GUO Chun[1] PING Yuan[2]

(School of Computer Science and Technology,Guizhou University,Guiyang 550000,China)[1]

(School of Information and Engineering,Xuchang University,Xuchang 461000,China)[2]

Abstract In order to efficiently protect data privacy being not leaked,which have high availability,a classification anonymous method based on weight attributes entropy(WECA) was proposed. The method builds on application-specific background of data classification mining,and calculates the classification importance of different standard identifier to sensitive attribute by the concept of information entropy in the data set,which selects the highest ratio of weight attributes entropy in classification quasi-identifier attributes to favorably divide the classification tree. The method also constructs the anonymous information loss measures of classification,which ensures the utility of classification on the premise of protecting privacy data. Finally,the experimental results on the standard data set show that the algorithm has fewer anonymous losses and higher classification accuracy,improving data availability.

Keywords Privacy protection,Classification anonymous,Weight attributes entropy,Classification accuracy

1 引言

伴随着数据库信息技术的广泛应用,人们可以通过数据挖掘等技术来分析数据库中的大量数据以获得实际有用的信息,但在对数据进行挖掘以产生知识和给人们现实生活带来便利的同时,随之而来的就是数据共享和用户隐私信息泄露问题。例如,通过数据挖掘对用户网上浏览记录进行分析,能够发现用户浏览记录之间的联系,但是在数据挖掘应用过程中难免会使得用户的个人隐私数据泄露。已有很多保护敏感信息的方法被提出,但是不论采用哪种方式进行隐私保护,都会对数据质量产生不同程度的破坏。因此,如何在高效地保护用户数据隐私不被泄露的同时还能够最大限度地保证分类匿名数据效用的问题成为了现在最热门的研究内容之一[1-2]。

目前大多数匿名模型和算法在保证数据隐私和可用性二者之间的平衡问题上仅单独地考虑了处理准标识符的方法或只考虑了敏感属性的敏感度量对原始数据进行匿名的方法,并没有考虑到不同的准标识符属性对敏感属性的重要程度。因此,本文主要针对数据分类挖掘的特定应用背景下的隐私信息保护需求,提出了一种基于权重属性熵的分类匿名算法,引入了权重熵的概念来衡量不同准标识符属性对敏感属性的重要程度,对数据集进行有利的分类,并构建数据匿名损失度量标准,以确保将数据隐私泄露降到最低。最后实验结果分析及验证表明,该算法能够在较好地保护数据隐私性的前提下获得一个较高的分类数据可用性。

2 相关工作

近年来,在数据分类应用场景上隐私保护与数据可用性之间的平衡已成为众多研究者的方向。匿名模型便是既满足

到稿日期:2016-08-20 返修日期:2016-11-01 本文受国家自然科学基金项目(61303232,61540049),贵州省基础研究重大项目(黔科合JZ字[2014]2001-21),贵州大学研究生创新基金(院项目),河南省高等学校重点科研项目(16A520025),许昌学院优秀青年骨干教师资助项目资助。

廖 军(1990-),女,硕士生,主要研究方向为网络与信息安全、数据挖掘,E-mail:amy_lj1220@163.com;蒋朝惠(1965-),男,教授,硕士生导师,主要研究方向为数据库与软件工程、网络与信息安全,E-mail:jiangchaohui@126.com;郭 春(1986-),男,博士,主要研究方向为网络安全、数据挖掘、风险评估;平 源(1981-),男,博士,副教授,硕士生导师,主要研究方向为信息安全、机器学习、数据挖掘。

保护数据隐私不被披露的需求,同时也能使分类的数据具有较高的数据效用的最佳方法。早在 2002 年,L. Sweeney 等人[3]提出了经典 k-anonymity 模型,该模型要求数据发布后数据表中的每一条元组对应至少 k 个不可区分的元组,从而使得攻击者无法推断出匿名后元组中的具体信息,以此来抵抗链接攻击。通过发布一定数量的不可区分的个体,使攻击者不能判别隐私信息所属的个体,从而防止链接攻击。很多学者在经典 K-匿名模型的基础上对解决数据隐私和可用性问题从不同方面进行了以下相关研究。

文献[4]进一步对 k-匿名模型进行了阐述,提出了一种基于聚类的 K-匿名算法;Benjamin C M Fung 等人[5]定义了信息熵来度量隐私与信息间的权衡,提出了自顶向下的 TDS 泛化算法,TDS 算法既能对离散属性泛化,也能对连续属性进行泛化,在泛化操作的每一次迭代中处理冗余的数据,保留数据信息从而引导分类信息,同时也要符合隐私要求。Xu Jian 等人[6]对准标识符属性进行了局部泛化,提出了两种匿名算法,以确保数据的可用性。Li 等人[7]提出了数据隐私和可用性的权衡方法。在考虑数据可用性的基础上,申艳光等人[8]利用安全多方计算方法,提出了保护隐私的 PPC4.5 的分类决策树算法。文献[9]和文献[10]在分类决策树模型的基础上,分别提出了改进的奇异值分解和多维隐匿的思想来指导匿名处理。赵爽[11]和杨静[12]都对敏感属性进行条件约束,分别提出了敏感度个性化(a,l)-匿名模型和敏感属性熵的微聚集方法。文献[13]提出了一种基于分类效用的匿名模型,通过对每个准标识符属性进行互信息计算来确定属性最大化分集能力,以属性分类性而不是隐私性为目标,该方法最终得到的匿名数据分类准确性较高。文献[14]提出了一种考虑属性权重的隐私保护方法,该方法对准标识符属性引入了权重泛化路径,制约了对不同特定数据应用分析场合的效用差异。

上述研究为隐私保护的发展奠定了坚实的基础,但大多数研究并没有考虑到准标识符属性对敏感属性之间的分类效用影响,因此本文提出了一种基于权重属性熵的分类匿名算法(Weight properties Entropy for Classification Anonymous,WECA),该算法根据不同准标识符属性对不同敏感属性的权重熵大小来对标准数据集进行有利的分类,即对分类有利的准标识符属性进行弱泛化,对分类作用不大的准标识符属性进行强泛化,并构建了一种泛化层次树的数据匿名损失度量标准,提高了数据的安全性和可用性。

3　基于权重属性熵的分类匿名模型

3.1　匿名属性的基本概念

表 1 列出了一所医院病人的诊断原始数据,其中将表中病人姓名、身份证号等唯一标识个体的信息舍去。这里设给定共享数据集为一个包含多属性的二维表 D,其中 d 为 D 中与某一个体相关的元组,D 中有 n 个元组,则 $D=\{d_1,d_2,\cdots,d_n\}$,且数据表 D 中包含了 m 个准标识符属性 $QI=\{Q_1,Q_2,\cdots,Q_m\}$ 和 k 个敏感属性 $S=\{S_1,S_2,\cdots,S_k\}$。数据表 D 包含两类属性:1)准标识符属性(quasi-identifiers)[3],它指数据表 D 中的一组属性 QI,$QI=\{Q_1,Q_2,\cdots,Q_m\}\subseteq D=\{d_1,d_2,\cdots,d_n\}$,通常与外部数据信息进行连接来标识个体记录,例如,表 1 中的 age,sex,zip 属性,它用于链接攻击,攻击者据此可推演出标识个体敏感信息的若干个属性的组合。2)敏感属性[4],指数据表 D 中一组属性 S,$S\subseteq D$,用于描述隐私信息不愿被披露的属性,如表 1 中的疾病属性。

表 1　病人原始数据表

序号	年龄	性别	邮编	疾病
1	43	男	3237	癌症
2	31	男	3142	支气管炎
3	19	女	3135	肺炎
4	37	男	3141	癌症
5	41	女	3229	肺炎
6	22	女	3132	支气管炎
7	32	男	3142	肺炎

定义 1(等价类[2])　准标识符属性 QI 在数据表 D 上具有全体相同的元组构成的子集称为等价类,其形式化表示:准标识符 QI 元组集合 $QI\subseteq D$,对于 $\forall x,y\in\{Q_1,Q_2,\cdots,Q_m\}$,有 $x[d_i,\cdots,d_j]=y[d_i,\cdots,d_j]=(a_i,\cdots,a_j)$,并且 $\forall r\in D-\{Q_1,Q_2,\cdots,Q_m\}$ 有 $r[d_i,\cdots,d_j]\neq(a_i,\cdots,a_j)$。

定义 2(K-匿名[3])　数据表 D 中,要求任何一个等价类至少包含 k 个元组,同时每一个元组至少与同一等价类中的 $k-1$ 个元组的取值无法辨别。这种情况称为数据表 D 满足 K-匿名,其中匿名一般要求 K 值大于 1。2-匿名的数据表如表 2 所列。

表 2　2-匿名的数据表

序号	年龄	性别	邮编	疾病
1	19~22	女	313*	肺炎
2	19~22	女	313*	支气管炎
3	31~37	男	314*	支气管炎
4	31~37	男	314*	癌症
5	31~37	男	314*	肺炎
6	41~43	*	32**	癌症
7	41~43	*	32**	肺炎

定义 3(泛化层次树)　给定准标识符属性 QI 中的属性 Q_i,其值域为 Z(Z 为有限集),设属性 Q_i 的层次树为 TQ_i,其节点集合为 $\{R,l_1,l_2,\cdots,l_n\}$(其中 R 为根节点,l 为叶子节点),称属性 Q_i 上的泛化树为映射函数 $f:TQ_i\to Z$,TQ 中存在父子关系的节点 x 和 y 满足 $f(y)\subseteq f(x)$,则对于根节点和叶节点有 $|f(l_i)|=1(1\leqslant i\leqslant n)$,$f(l_1)\cup f(l_2)\cup\cdots\cup f(l_n)=f(R)$ 且 $f(R)\in Z$。图 1－图 3 分别示出了年龄的层次泛化树、邮编的泛化树和性别的泛化树。

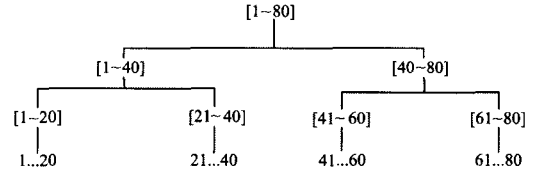

图 1　年龄的层次泛化树

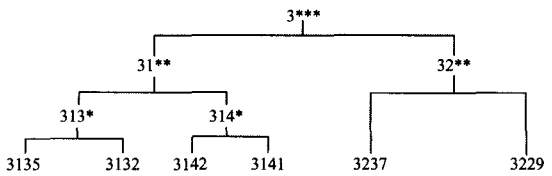

图 2　邮编的泛化树

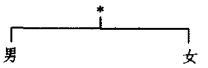

图 3 性别的泛化树

3.2 权重属性熵度量

在数据挖掘隐私保护中，熵[11]是一种自顶向下的分裂方法，用于表示属性的不确定性，属性的不确定性越小，纯度越高，分类的效果越好。

定义 4(熵) 给定一个随机变量 x，其熵的定义为：

$$E(x)=E(p_1,p_2,\cdots,p_k)=\sum_{i=1}^{k}p_i\log_2 p_i \tag{1}$$

其中，p_i 为 x 的各个值出现的概率($0\leqslant p_i\leqslant 1$)，$\sum_{i=1}^{k}p_i=1$。

设给定样本数据表 D 中含有 k 个类别样本 $C_1,C_2,\cdots,C_k$，数据表中类 C_i 的元组个数为 S_i，$i=1,2,\cdots,k$。故给定样本数据分类的信息熵为：

$$E(S)=E(S_1,S_2,\cdots,S_k)=\sum_{i=1}^{k}\frac{R_i}{|D|}\log_2\frac{R_i}{|D|} \tag{2}$$

其中，$\frac{R_i}{|D|}$ 为类别样本的概率，$|D|$ 为数据表 D 的元组数。

定义 5(权重属性熵) 假如属性 Q 将数据集 D 划分为 v 个划分区 $\{D_1,D_2,\cdots,D_v\}$，其中 Q 有 v 个不同的值 $\{q_1,q_2,\cdots,q_v\}$，D_j 包含了 D 中属性值为 q_j 的所有元组，则属性 Q 划分 D 的元组分类所需要的权重属性熵为：

$$E(Q)=\sum_{j=1}^{v}w_jE(S_1,S_2,\cdots,S_k)=\sum_{j=1}^{v}\frac{D_j}{|D|}\sum_{i=1}^{k}\frac{R_i}{|D|}\log_2\frac{R_i}{|D|} \tag{3}$$

其中，$w_j=\frac{D_j}{|D|}$ 表示 j 个划分大小的权重，需要的权重属性熵越小，划分区的准确性越高，越有利于分类；反之，分类效果越差。

定义 6(权重属性熵增量) 属性 Q 对给定数据集 D 类类别产生的权重属性熵增量为：

$$\Delta E=E(S)-E(Q)=E(S_1,S_2,\cdots,S_k)-E(Q) \tag{4}$$

从式(4)可以看出，如果权重属性熵 $E(Q)$ 值越大，则 ΔE 越小，分类效果越差；如果权重属性熵 $E(Q)$ 值越小，则 ΔE 越大，分类效果越好。

定义 7(权重属性熵增量比率) 由于权重属性熵增量偏重于选择属性值出现比较多的情况，因此选用了权重属性熵增量比率来消除偏重问题，权重属性熵增量比率定义为：

$$\Delta E'=\Delta E/E(Q)' \tag{5}$$

其中，$E(Q)'=\sum_{j=1}^{v}\frac{D_j}{|D|}\log_2\frac{D_j}{|D|}$ 表示准标识符属性 Q 中元组分类熵信息，从式(4)和式(5)可知，$\Delta E'$ 旨在更好地选择属性值，所以权重属性熵增量比率越大，越有利于分类。

如表 1 所列，分类类别号已经确定，有 3 个类，分别是类 1：癌症，类 2：支气管炎，类 3：肺炎。

$$E(S)=E(2,2,3)=-\frac{2}{7}\log\frac{2}{7}-\frac{2}{7}\log\frac{2}{7}-\frac{3}{7}\log\frac{3}{7}$$
$$=1.557$$

$$E(\text{性别})=\frac{4}{7}E(2,1,1)+\frac{3}{7}E(0,1,2)=1.092$$

$$E(\text{性别})'=-\frac{4}{7}\log\frac{4}{7}-\frac{3}{7}\log\frac{3}{7}=0.985$$

$$\Delta E(\text{性别})'=\Delta E/E(Q)'=E(S(\text{疾病}))-E(\text{性别})/E(\text{性别})'=0.465/0.985=0.472$$

由以上计算同理可求出：

$\Delta E(\text{年龄})'=1.557/2.807=0.554$

$\Delta E(\text{邮编})'=1.2713/2.521=0.504$

因为性别、年龄和邮编 3 个属性中年龄的权重属性熵增量比最大，所以选择年龄属性作为属性分裂点，即分类树的根节点。

3.3 隐私数据损失度量

准标识符属性在进行泛化时难免会出现隐私数据泄露，因此本文采用了加权确定性代价[6]来分别计算准标识符属性中分类型属性和数值型属性的匿名损失。

定义 8(分类型属性的匿名损失) 假定准标识符属性 $QI=\{QC_1,QC_2,\cdots,QC_{m1},QN_1,QN_2,\cdots,QN_{m2}\}$，其中 QC_j ($j=1,2,\cdots,m1$) 为分类型属性，相应的分类树为 T_j ($j=1,2,\cdots,m1$)。元组 t 中任意分类型属性 QC_j ($j=1,2,\cdots,m1$) 的值 v_j 泛化为祖先节点值 p_j 后，定义产生的匿名损失为：

$$PL_{QC_j}(t)=\sum_{j=1}^{k}w_j\frac{|p_j|}{|T_j|} \tag{6}$$

其中，k 为敏感属性值的个数，w_j 表示准标识符属性的分类权重，$|p_j|$ 表示子树 p_j 的叶子节点数，$|T_j|$ 表示分类树 T_j 的叶子节点个数。

设元组 t 中所有分类型准标识符属性 QC_j ($j=1,2,\cdots,m1$) 匿名化后，产生的匿名损失为：

$$PL_{QC}(t)=\sum_{j=1}^{m1}PL_{QC_j}(t)=\sum_{j=1}^{m1}\sum_{j=1}^{k}w_j\frac{|p_j|}{|T_j|} \tag{7}$$

定义 9(数值型属性的匿名损失) 假定准标识符属性 $QI=\{QC_1,QC_2,\cdots,QC_{m1},QN_1,QN_2,\cdots,QN_{m2}\}$，其中 QN_j ($j=1,2,\cdots,m2$) 为数值型属性，相应的取值域为 QD_j ($j=1,2,\cdots,m2$)。元组 t 中任意数值型属性 QN_j ($j=1,2\cdots,m2$) 的值由 b_j 泛化到区间 $[a_j,c_j]$ ($a_j\leqslant b_j\leqslant c_j$) 后，产生的匿名损失定义为：

$$PL_{QN_j}(t)=\sum_{j=1}^{k}w_j\frac{a_j-c_j}{|QD_j|} \tag{8}$$

其中，k 为敏感属性值的个数，w_j 表示准标识符属性分类权重，$|QD_j|=\max(QN_j)-\min(QN_j)$。

设元组 t 中所有数值型准标识符属性 QN_j ($j=1,2,\cdots,m2$) 匿名化后，产生的匿名损失为：

$$PL_{QN}(t)=\sum_{j=1}^{m2}PL_{QN_j}(t)=\sum_{j=1}^{m2}\sum_{j=1}^{k}w_j\frac{a_j-c_j}{|QD_j|} \tag{9}$$

定义 10(元组匿名损失) 假定准标识符属性 $QI=\{QC_1,QC_2,\cdots,QC_{m1},QN_1,QN_2,\cdots,QN_{m2}\}$，其中 QC_j ($j=1,2,\cdots,m1$) 为分类型属性，相应的分类树为 T_j ($j=1,2,\cdots,m1$)。QN_j ($j=1,2\cdots,m2$) 为数值型属性，相应的取值域为 QD_j ($j=1,2,\cdots,m2$)，元组 t 匿名化后，产生的匿名损失为：

$$PL(t)=PL_{QC_j}(t)+PL_{QN_j}(t) \tag{10}$$

故数据表的 $PL(D)$ 为表中所有 PL 之和，可表示为：

$$PL(D)=\sum_{t\in D}PL(t) \tag{11}$$

定义 11(分类匿名保护度)

$$cap=\frac{\Delta E'}{PL(D)} \tag{12}$$

由式(12)可知，$\Delta E'$ 越大，$PL(D)$ 越小，则分类匿名保护度 cap 越大，即分类效果就越好。

4　基于权重属性熵的分类匿名算法

为了在保护隐私不被泄露的同时保证数据的高可用性，本文提出了一种基于权重属性熵的分类匿名算法(WECA)，该算法采用分类决策树模型在不同准标识符属性对分类敏感属性的分类效用影响程度下将数据集转换成 K-匿名形式的等价类划分，对划分进行强弱泛化处理。该算法的基本思想：1)在给定的数据表 D 中，对每个敏感属性 S 的信息熵 $E(S_1, S_2, \cdots, S_k)$ 进行计算，得到敏感属性的数值，便于准标识符属性从敏感属性熵值中得到分类信息；2)根据不同准标识符属性对敏感属性分类权重熵增量比进行计算，并将计算结果按从大到小的顺序排序，以便选择分裂节点；3)选择分类权重熵增量最大的作为根节点的分裂属性，根据分裂属性分类能力将其划分到相对应的分支区域，并将其存到匿名数据表中；4)对分类树中剩下的的叶子节点递归地进行上述匿名操作，直到输出满足 K-匿名要求的数据表为止。

WECA 算法的伪代码如下所示。

输入：标准数据表 D，准标识符 QI，敏感属性 S，K-匿名参数 K，其中有 n 个数据表元组，m 个准标识符属性，k 个敏感属性

输出：分类匿名数据表 D*

1. D 为非 ϕ，if n<k then return
2. 计算 D 中 S 的熵值 $E(S_1, S_2, \cdots, S_k)$
3. for each QI 计算出每个准标识符属性 Qi 对敏感属性 Si 的分类权重属性熵增量比 $\Delta E'$
4. 对 $\Delta E'$按从大到小的顺序排列，假设排列为 list={$list_1, list_2, \cdots, list_m$}
5. 对于排列 list={$list_1, list_2, \cdots, list_m$}($1 \leqslant i \leqslant m$)，按分类排列顺序，对排列中记录以隐私损失最小和高可用性即以较高的分类匿名保护度量 cap 为目标进行分类匿名处理
   - 5.1 初始化 T=0
   - 5.2 当 i<=m and $list_i$ 为最大的分类权重属性熵增量比 $\Delta E'$，循环执行
     - 5.2.1 将 $list_i$ 作为分类树的根节点，并将 $list_i$ 划分为 r_i 个不同值
     - 5.2.2 If $list_i$ 划分的 r_i 个不同值满足 K-匿名要求
     - 5.2.3 then 生成划分匿名的 r_i 个叶节点
     - 5.2.4 list=list−{$list_i$}
     - 5.2.5 T=T∪{r_i}
     - 5.2.6 else delete $list_i$ 的最优划分
   - 5.3 return 分类树
6. 当分类树中仍有叶子节点，循环处理//所有的数据集合所包含的记录都大于或等于 k 个
   - 6.1 遍历分类树
   - 6.2 if 左右兄弟叶子节点所包含的元组<K
   - 6.3 then 合并左右兄弟节点的元组
   - 6.4 else 左右兄弟匿名叶子节点的元组存入 T and flag=1//该叶子节点标识为已匿名处理
7. D* ←T，输出分类数据 D*

5　实验数据及结果分析

实验数据集为 http://kdd.ics.uci.edu 网站提供的美国人口统计 adult 数据集，该数据集常被用作数据挖掘和隐私保护的实验数据。Adult 数据集中包含有 32561 条记录，其中包含 15 个属性，分别为 age，workclass，fnlwgt，education，education-num，martial-satus，occupation，relationship，race，sex，capital-gain，capital-loss，hours-per-week，native-country，salary。本实验通过 weka 软件选取了 9 个属性作为实验对象(表 3 列出了各个属性的信息)，并设其中 8 个为准标识符属性，1 个为敏感属性。实验环境：AMD FX(tm)-6100 six-core processor 3.3GHz CPU，4GB 内存，Microsoft Windows 7 的操作系统，本算法所涉及的代码采用 java 软件和 weka 软件编写。

表 3　属性的信息

序号	属性名	属性类型	属性数目
1	native-country	Nominal	42
2	relationship	Nominal	6
3	marital-status	Nominal	7
4	occupation	Nominal	15
5	education	Nominal	16
6	workclass	Nominal	9
7	sex	Nominal	2
8	age	numeric	73
9	race	sensitive	5

本文将所提的 WECA 算法与 Top-down 算法[5] 和 IACK 算法[13] 进行对比实验，分为数据可用性分析、隐私信息损失分析和执行时间分析 3 个部分来完成。

5.1　数据可用性分析

为了验证算法的分类数据可用性，本文采用分类的精度来度量匿名化后的数据可用性。分别在决策树 C4.5 分类模型和贝叶斯分类模型上对 WECA 算法、Top-down 算法、IACK 算法和原始数据进行分类精度比较，取准标识符属性个数$|QI|=8$，设参数 K 的值为 2,4,6,8,10。

图 4 和图 5 分别描述了在 C4.5 分类模型和贝叶斯分类模型上对 WECA 算法、Top-down 算法、IACK 算法和原始数据分类精度的比较。由图可知，随着 K 值的不断增大，3 种算法的分类精度都有所下降。这是因为在进行数据匿名分类时，需要对准标识符属性进行泛化，或多或少地导致数据失真，使得数据可用性降低。但 WECA 算法的分类精度比 Top-down 算法和 IACK 算法的都高，更接近于原始数据的精度，其分类精度始终保持在 75%以上，最差情况下分类精度也有77.2%，说明 WECA 算法具有较高的分类精度。

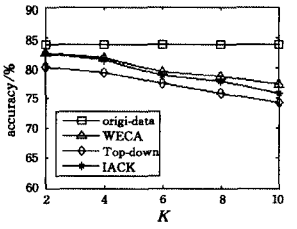

图 4　在 C4.5 分类模型上的分类精度

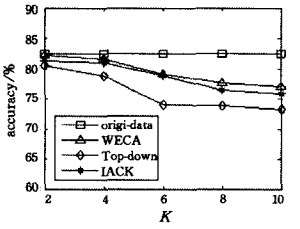

图 5　在贝叶斯分类模型上的分类精度

5.2　匿名信息损失分析

图 6 示出了在 K 值不断变化下 WECA 算法、Top-down 算法和 IACK 算法的匿名信息损失比较情况，其中准标识符属性$|QI|=8$。由图可知，在 K 值不断增大时，3 种算法的匿名信息损失随之增加，这是因为当 K 值不断增加时，等价类中的元组数相应地增加，使得准标识符属性泛化程度也随之加强。但由图 6 可看出，WECA 算法的匿名信息损失比 IACK 算法的小，其原因是 WECA 算法在对属性进行划分时以分类匿名保护度最高为目标，考虑到了不同准标识符属性

对敏感属性的分类效用性，其平均匿名损失也只有 18.16%，所以其匿名损失比 IACK 算法更小。

图 7 展示了当 $K=6$ 时不同准标识符属性 QI 在 WECA 算法、Top-down 算法和 IACK 算法下的匿名信息损失比较。从图 7 可以看出，随着准标识符属性 QI 的增加，3 种算法的匿名信息损失也相应增大，这是由于随着 QI 个数的增加，等价类中属性泛化个数也增加，匿名损失自然也就增加了。

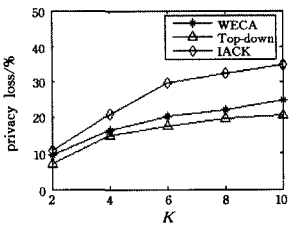

图 6 不同 K 值下的匿名信息损失

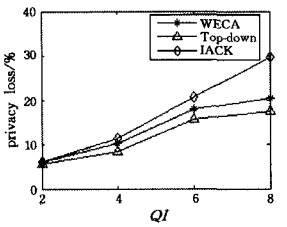

图 7 不同 QI 值下的匿名信息损失

5.3 执行时间分析

在给定准标识符属性 $|QI|=8$ 的情况下，在参数 K 增加时，3 种算法的执行时间比较结果如图 8 所示。由图 8 可知，WECA 算法、Top-down 算法和 IACK 算法在准标识符属性一定的情况下执行时间都不断增加，WECA 算法的执行时间比 Top-down 算法和 IACK 算法略高，其原因在于本文算法在进行属性划分时不仅需要考虑敏感属性的信息熵，还要计算不同准标识符属性对敏感属性的分类重要程度，在更好地保护隐私的前提下提高了数据的可用性，该时间花销是可以接受的。

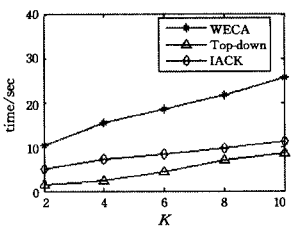

图 8 不同 K 值下的执行时间

上述实验表明，在较好地保护数据隐私性的前提下，WECA 算法的分类精度仅比原始数据得到的分类精度低 4.1%，从而大大地提高了数据可用性。因此，与其他算法相比，WECA 算法能够牺牲较小的时间开销和较少的分类匿名信息损失来获得较高的分类精度，从而提高了数据可用性。

结束语 为了在保证隐私信息不被泄露的同时保证数据匿名后有较高可用性，本文提出了一种基于权重属性熵的分类匿名算法，该算法引入信息熵的概念，通过不同准标识符属性对敏感属性的分类重要程度的大小构建分类匿名模型，同时构建了隐私信息损失标准度量，在更好地保护数据隐私的同时使得数据集有较高的分类可用性。

参 考 文 献

[1] FENG D G, ZHANG M, LI H. Big data security and privacy protection [J]. Chinese Journal of Computers, 2014, 37(1): 246-258. (in Chinese)
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.

[2] LIU Y H, ZHANG T Y, JIN X L, et al. Personal privacy protection in the era of big data [J]. Journal of Computer Research and Development, 2015, 52(1): 229-247. (in Chinese)
刘雅辉，张铁赢，靳小龙，等. 大数据时代的个人隐私保护[J]. 计算机研究与发展，2015，52(1)：229-247.

[3] SWEENEY L. K-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, 10(5): 571-578.

[4] AGGARWAL G, PANIGRAHY R, FEDR T, et al. Achieving anonymity via clustering [J]. ACM Transactions on Algorithms, 2010, 6(3): 1-19.

[5] B C, FUNG M, WANG K, et al. Top-Down Specialization for Information and Privacy Preservation [C] // Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE2005). Tokyo Japan, 2005: 205-216.

[6] XU J, WANG WEI, PEI J, et al. Utility-based anonymization using local recoding[C]// Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (SIGKDD). Philadelphia, PA, USA, 2006: 785-790.

[7] LI T C, LI N H. On the tradeoff between privacy and utility in data publishing [C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2009: 517-525.

[8] SHEN Y G, SHAO H, ZHANG Y Q. Research on privacy preserving distributed decision-tree classification algorithm[J]. Application Research of Computers, 2010, 27(8): 3070-3072. (in Chinese)
申艳光，邵慧，张永强. 隐私保护的分布式决策树分类算法的研究[J]. 计算机应用研究，2010，27(8)：3070-3072.

[9] LI G, WANG Y D. An improved privacy-preserving classification mining method based on singular value decomposition [J]. ACTA Electronica Sinica, 2012, 40(4): 739-744. (in Chinese)
李光，王亚东. 一种改进的基于奇异值分解的隐私保持分类挖掘方法[J]. 电子学报，2012，40(4)：739-744.

[10] KISILEVICH S, ROKACH L, ELOVICI Y, et al. Efficient multidimensional suppression for K-anonymity[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(3): 334-347.

[11] ZHAO S, CHEN L. Personalized (a, l)-anonymity method based on sensitivity[J]. Computer Engineering, 2015, 41(1): 115-120. (in Chinese)
赵爽，陈力. 基于敏感度个性化 (a, l)-匿名方法[J]. 计算机工程，2015，41(1)：115-120.

[12] YANG J, WANG C, ZHANG J P, et al. Micro-aggregation algorithm based on sensitive attribute entropy[J]. ACTA Electronica Sinica, 2014, 42(7): 1327-1337. (in Chinese)
杨静，王超，张健沛，等. 基于敏感属性熵的微聚集算法[J]. 电子学报，2014，42(7)：1327-1337.

[13] LI J Y, LIU J X, BAIG M. Information based data anonymization for classification utility[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 70(12): 1030-1045.

[14] XU Y, QIN X L, YANG Y T, et al. A QIweight-aware approach to privacy preserving publishing data set[J]. Journal of Computer Research and Development, 2012, 49(5): 913-924. (in Chinese)
徐勇，秦小麟，杨一涛，等. 一种考虑属性权重的隐私保护数据发布方法[J]. 计算机研究与发展，2012，49(5)：913-924.