文章编号:1001-506X(2011)09-2139-05

基于灰关联分析的 K 匿名方法及其在聚类中的应用

郭 昆[1],张岐山[2]

(1. 福州大学数学与计算机科学学院,福建 福州 350108;2. 福州大学管理学院,福建 福州 350108)

摘 要:采用泛化和抑制技术对数据进行 K 匿名化处理,需要在数据的有用性和隐私保护度之间保持平衡。提出一种新的利用基于差异信息理论的灰关联分析实现 K 匿名的方法,利用数据序列之间的均衡接近度描述数据点之间的相似程度,据此进行相应的泛化和抑制操作,并将 K 匿名后的数据应用于聚类分析。在真实数据集上的测试验证了该方法的有效性。

关键词:聚类;K 匿名;灰关联分析

中图分类号:TP 301 **文献标志码**:A **DOI**:10.3969/j.issn.1001-506X.2011.09.41

K-anonymity method based on grey relational analysis and its application in clustering

GUO Kun[1], ZHANG Qi-shan[2]

(1. *College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China*;
2. *School of Management, Fuzhou University, Fuzhou 350108, China*)

Abstract: It is important to keep balance between the usefulness and the degree of privacy protection when the techniques of generalization and suppression are applied to the K-anonymization of data. A novel K-anonymity method implemented by grey relational analysis based on difference information theory is proposed. The similarity between any two data sequences is described by their balanced closeness degrees. The generalization and suppression operations are carried out accordingly. The K-anonymized data are then applied in cluster analysis. The experimental results on the real data sets prove the effectiveness of the proposed method.

Keywords: clustering; K-anonymity; grey relational analysis

0 引 言

随着数据挖掘技术的快速发展以及公众对个人隐私的日益关注,如何在从数据中挖掘信息的同时尽可能地保护用户的隐私不被泄露已经成为数据挖掘领域的一个重要研究方向[1]。保护用户隐私的一种常用的方法是对数据进行改动,删除可能用于唯一标识个人的敏感属性,如:姓名、身份证号码等。但是,简单地删除敏感信息并不能保证用户隐私不被泄露。

文献[2-3]将这种通过多表关联获取个人隐私的攻击方法称为链接攻击,并提出一种能够防止链接攻击的模型—K 匿名模型。K 匿名要求数据表中的每条记录与至少其他 K-1 条记录在准标识符上的投影值相同。

实现 K 匿名的常用技术包括泛化和抑制。文献[4]将泛化和抑制技术具体分为属性泛化(attribute generalization, AG)、元组泛化(tuple generalization, TG)、单元泛化(cell generalization, CG)、属性抑制(attribute suppression, AS)、元组抑制(tuple suppression, TS)和单元抑制(cell suppression, CS)等 6 种。这些技术的单独或联合使用形成了不同的 K 匿名方法。由于求最优 K 匿名方案为 NP 难问题[5-6],K 匿名方法又可以分为精确求解方法和近似求解方法。前者能保证找到最优 K 匿名方案,但其时间复杂度为指数级,只适用于小规模数据,后者只能找到近似最优 K 匿名方案,但其时间复杂度为线性或近似线性,可应用于大规模数据。典型的精确求解方法包括文献[7]提出的方法、Datafly 算法[8]、K-Optimize 算法[9]、Incognito 算法[10]和 Mondrain 算法[11]等。它们均采用 AG 与 TS 相结合的技术,并通过各种优化策略以缩小搜索范围,提高算法效率,但最坏情况下的时间复杂度仍为指数级。典型的近似求解方法包括基于遗传算法的方法[12]、基于模拟退火的方

收稿日期:2010-09-28; **修回日期**:2011-04-03。

基金项目:国家自然科学基金(70871024);福建省自然科学基金(2010J01358);福建省教育厅科技项目(JB09006)资助课题

作者简介:郭昆(1979-),男,博士研究生,主要研究方向为灰色系统、数据挖掘、决策支持系统和计算机软件理论。

E-mail:gukn@fzu.edu.cn

法[12-13]、自顶向下泛化法[14]、自底向上泛化法[15]、以及一些近似算法[5-6,16]等。通过采用多种启发策略，这些方法可以在多项式时间内找到满足特定目标函数的局部最优方案，但不能保证找到全局最优方案。Classfly 算法引入特征类的概念，克服了 Datafly 算法在属性泛化时可能导致信息过量损失的问题，提高了数据的可用性[17]。

K 匿名方法与数据挖掘方法的结合主要有两种模式：匿名-挖掘模式和挖掘-匿名模式，前者是先进行 K 匿名化再运行挖掘算法，后者则相反[18]。匿名-挖掘模式将 K 匿名操作与挖掘操作分离，便于分别应用成熟的算法，在隐私保护和挖掘精度间寻求最佳平衡，但较高的 K 匿名要求可能造成挖掘算法无法弥补的较大的精度损失。目前基于 K 匿名的隐私保护数据多采用这种模式。挖掘-匿名模式由于是在挖掘结果(关联规则、决策树等)上进行 K 匿名操作，挖掘精度不受影响，但需要针对挖掘结果的不同表示方式设计不同的 K 匿名方法[19]，且挖掘操作只能由数据持有者进行。

灰色系统理论是处理现实中存在的部分信息明确、部分信息不明确的“小样本”、“贫信息”不确定性问题的有力工具，已经在油气勘探、工业控制、图像处理、经济预测等多个学科领域得到广泛应用[20-25]。灰关联分析是灰色系统理论的重要组成部分。本文提出将基于差异信息理论的灰关联分析应用于数据表的 K 匿名化，利用均衡接近度描述记录之间的相似程度，根据相似度的高低对记录进行泛化和抑制处理，以实现数据表的 K 匿名化，并通过在真实数据集上的聚类实验证明了新方法的有效性。

1 数据的 K 匿名化

1.1 K 匿名的定义

定义 1 属性集。设 $T(A_1,\cdots,A_m)$表示有限个元组构成的二维表，则 T 的属性集为$\{A_1,\cdots,A_m\}$。

给定表 $T(A_1,\cdots,A_m)$，设$\{A_{i1}\cdots,A_{id}\}\in\{A_1,\cdots,A_m\}$为属性子集$(i1,\cdots,id\in\{1,\cdots,m\})$，则 $T(A_{i1},\cdots,A_{id})$表示 T 在$\{A_{i1},\cdots,A_{id}\}$上的投影，$t(A_{i1},\cdots,A_{id})$表示元组 t 在$\{A_{i1},\cdots,A_{id}\}$上的值序列。

定义 2 准标识符。给定一个实体集 U，描述实体的表 $T(A_1,\cdots,A_m)$，映射 $f_c:U\to T$，$f_g:T\to U'$，$U'\in U$。若对属性子集 $QI=\{A_{i1},\cdots,A_{id}\}\in\{A_1,\cdots,A_m\}(i1,\cdots,id\in\{1,\cdots,m\})$，$\exists\, p_i\in U$ 满足 $f_g(f_c(p_i)[QI])=p_i$，则 QI 称为 T 上的准标识符。

定义 3 K 匿名。给定表 $T(A_1,\cdots,A_m)$和准标识符 $QI=\{A_{i1},\cdots,A_{id}\}\in\{A_1,\cdots,A_m\}(i1,\cdots,id\in\{1,\cdots,m\})$，若 T 在 QI 上的投影 $T(QI)$中，每个元组重复出现至少 K 次，则称表 T 满足关于准标识符 QI 的 K 匿名。

1.2 K 匿名的实现

在实现 K 匿名时，对表中的每个属性，可以建立一个域泛化层次(domain generalization hierarchy, DGH)[8]，图 1 给出 DGH 的示例。

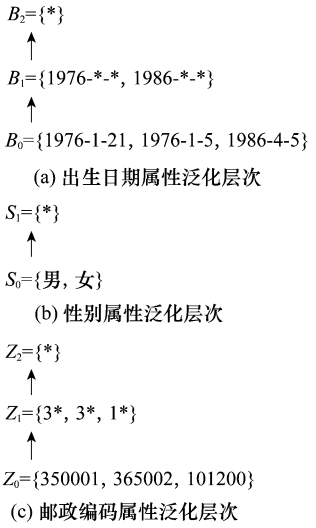

图 1 出生日期、性别和邮政编码的 DGH

DGH 的最底层由元组在属性上的真实值组成，每上升一层就将相似的属性值进行归并，使该层属性值代表对下一层属性值的泛化。随着层次的上升，泛化程度逐渐提高，属性值逐渐减少，到最高层时只剩一个代表最高泛化层次的属性值。此时，所有元组在该属性上的取值相同，这意味着该属性的信息完全被隐藏。因此，当属性被泛化至最高层次时也称为属性的抑制。

K 匿名化程度越高，隐私保护效果越好，但信息丢失也越多。因此 K 匿名算法的研究重点在于如何在满足 K 匿名的条件下使信息损失最小，即寻找最优 K 匿名方案。

2 灰关联分析

灰色系统理论是一种处理不确定性和非线性问题的系统理论。灰关联分析是灰色系统理论中的重要组成部分，它对样本的数量和分布规律不做要求，量化结果与定性分析的结果保持一致，因此特别适用于对小样本、无明显规律的数据进行研究[20]。文献[21]提出灰朦胧集的差异信息理论，通过在灰关联分析中引入灰关联熵，克服了传统灰关联分析可能造成局部点关联倾向的问题。

定义 4 设 $\boldsymbol{X}$ 为灰关联因子集，$\boldsymbol{X}=\{\boldsymbol{x}_i\mid \boldsymbol{x}_i=(x_i(1),x_i(2),\cdots,x_i(n)),i\in N\}$，$N=\{0,1,2,\cdots,m\}$，$m\geqslant 2$，$K=\{1,2,\cdots,n\}$，$n\geqslant 3$，$\boldsymbol{x}_0\in\boldsymbol{X}$ 为参考序列，$\boldsymbol{x}_i\in\boldsymbol{X}$ 为比较序列。若存在一个非负实数 $\gamma(\boldsymbol{x}_0(k),\boldsymbol{x}_i(k))$为 $\boldsymbol{X}$ 上一定环境下的比较测度，且满足

$$\gamma(\boldsymbol{x}_0,\boldsymbol{x}_i)=\frac{1}{n}\sum_{k=1}^{n}\gamma(x_0(k),x_i(k)) \tag{1}$$

当其满足规范性、偶对称性、整体性、接近性时，称 $\gamma(x_0(k),x_i(k))$ 为 $\boldsymbol{x}_i$ 对 $\boldsymbol{x}_0$ 在第 k 点的灰关联系数，称 $\gamma(\boldsymbol{x}_0,\boldsymbol{x}_i)$为 $\boldsymbol{x}_i$ 对 $\boldsymbol{x}_0$ 的灰关联度[20]。

定义 5 设 $\boldsymbol{r}_i=(\gamma(x_0(k),x_i(k)))$为第 i 个比较序列的灰关联系数序列，$\boldsymbol{C}=\{\boldsymbol{r}_i\mid i\in N\}$为灰关联系数序列集，则称映射

$$\phi:\boldsymbol{r}_i\to\boldsymbol{v}_i$$
$$\gamma(x_0(k),x_i(k))\mapsto v(x_0(k),x_i(k))$$

$$v(x_0(k),x_i(k)) \triangleq \frac{\gamma(x_0(k),x_i(k))}{\sum_{k=1}^{n}\gamma(x_0(k),x_i(k))} \quad (2)$$

为灰关联系数分布映射,映射值 $v(x_0(k),x_i(k))$ 为第 i 个比较序列在第 k 点的灰关联密度值。此比较序列的所有关联密度值的全体构成灰关联密度序列,记为 $\boldsymbol{v}_i$。

定义 6　设 $\boldsymbol{V}=\{\boldsymbol{v}_i \mid i\in N\}$ 为灰关联密度序列集,则称函数

$$I(\boldsymbol{v}_i) \triangleq -\sum_{k=1}^{n} v(x_0(k),x_i(k))\ln v(x_0(k),x_i(k)) \quad (3)$$

为第 i 个比较序列的灰关联熵。

定义 7　设 $I(\boldsymbol{v}_i)$ 为第 i 个比较序列的灰关联熵,I_m 为灰关联系数序列的最大关联熵,则称

$$E(\boldsymbol{x}_0,\boldsymbol{x}_i) \triangleq I(\boldsymbol{v}_i)/I_m \quad (4)$$

为第 i 个比较序列的熵关联度。

定义 8　设 $\gamma(\boldsymbol{x}_0,\boldsymbol{x}_i)$ 和 $E(\boldsymbol{x}_0,\boldsymbol{x}_i)$ 分别为第 i 个比较序列的灰关联度和熵关联度,则称

$$B(\boldsymbol{x}_0,\boldsymbol{x}_i) = E(\boldsymbol{x}_0,\boldsymbol{x}_i)\times\gamma(\boldsymbol{x}_0,\boldsymbol{x}_i) \quad (5)$$

为第 i 个比较序列的均衡接近度。

由于均衡接近度既考虑了灰关联因子序列间点的距离接近性,又考虑了整体的无差异性接近,因此可以克服点关联强倾向问题。

3　K 匿名的灰关联分析方法

通过计算两个数据序列之间的均衡接近度,可以得到它们之间的相似度信息。如果将原始数据表在准标识符 QI 上的投影表示为矩阵 $\boldsymbol{D}_{n\times m}$($n$ 为元组数;$m=|QI|$),则可以将矩阵的每个行向量看作一个数据序列,利用基于差异信息熵的灰关联分析计算所有数据序列之间的均衡接近度,形成一个均衡接近度矩阵,为

$$\boldsymbol{B} = \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{bmatrix} \quad (6)$$

定义 K 近邻序列集如下:

定义 9　设 n 个数据序列 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 的均衡接近度矩阵如式(6)所示,B_{ik} 表示矩阵 $\boldsymbol{B}$ 的第 i 个行向量的第 k 个最大元素,则序列 $\boldsymbol{x}_i$($i=1,\cdots,n$)的 K 近邻序列集 $AP_{ik}=\{\boldsymbol{x}_j \mid B_{ij}\geqslant B_{ik}, j=1,\cdots,n\}$。

所有 AP_{ik}($i=1,\cdots,n$)构成数据表的一个覆盖。当两个 AP_{ik} 的交集非空时,将交集元素分配给均衡接近度高的 AP_{ik},由此形成的 AP'_{ik} 构成数据表的一个划分,每个 AP'_{ik} 称为一个 K 近邻等价类($i=1,\cdots,s$,s 为近邻等价类数)。显然,$s\leqslant$floor (n/k),floor $(\cdot)$ 为向下取整函数。当 n/k 为整数时,所有元组都被划分到等价类,此时将同一个等价类中的元组用元组 i 代替;当 n/k 不为整数时,存在至多 $k-1$ 个元组无法划分等价类,此时采用抑制技术将这些元组的各属性值都泛化至最高层次。由此可以设计一种基于灰关联分析的 TG+TS 技术的 K 匿名方法,称为 KAoGRA 算法。相对于采用 AG 技术的 K 匿名方法,采用 TG 技术的 K 匿名方法不要求所有元组的同一个属性都取相同的泛化值,允许根据元组间的相似程度进行不同程度的泛化,因此泛化后的信息损失通常比较小。

3.1　KAoGRA 算法流程

输入　原始数据矩阵 $\boldsymbol{D}_{n\times m}$

输出　K 匿名数据矩阵 $\boldsymbol{X}_{n\times m}$

步骤 1　将 $\boldsymbol{D}_{n\times m}$ 的 n 个行向量看作 n 个数据序列,计算它们之间的均衡接近度,生成均衡接近度矩阵 $\boldsymbol{B}$。

步骤 2　根据矩阵 $\boldsymbol{B}$ 将数据表划分成 s 个 K 近邻等价类,$s\leqslant$floor (n/k)。

步骤 3　将第 i 个等价类中的元组替换成元组 i。

步骤 4　若 n/k 不为整数,抑制剩余的不能被划分到任何一个等价类的元组。

步骤 5　返回 K 匿名化后元组形成的矩阵 $\boldsymbol{X}_{n\times m}$。

3.2　算法复杂度分析

由于计算两个长度为 m 的序列的均衡接近度的时间复杂度为 $O(m)$,生成整个均衡接近度矩阵的时间复杂度为 $O(n\times m)$。搜索一个序列的 K 近邻序列集在最坏情况下需要扫描整个序列集,其时间复杂度为 $O(n)$,因此找出所有的 K 近邻序列集需要的时间复杂度为 $O(n^2)$。算法步骤 3 和步骤 4 的时间复杂度不超过 $O(n/k)$ 和 $O(k)$。这样,算法总的时间复杂度为 $O(n^2)$。

算法中的原始数据矩阵和 K 匿名矩阵占用的空间均为 $O(n\times m)$,计算过程生成的均衡接近度矩阵需要 $O(n^2)$ 的空间。因此,算法总的空间复杂度为 $O(n^2)$。

4　实验与分析

为了验证 KAoGRA 算法的有效性,在一台硬件配置为 1.66GHz CPU、2GB 内存,软件配置为 Windows XP(SP2)的台式机上进行实验。实验采用两组数据集:由文献[26]提供的 CENSUS 数据集和由 UCI 提供的 Adult 数据集[27]。CENSUS 数据集包含 1 080 条个人信息记录,13 个属性均为数值型数据。由于该数据集不包含分类信息,在实验时先将记录按 ERNVAL 属性值递增排序,再根据 ERNVAL 属性值的中位数将整个数据集划分为 4 个类别。Adult 数据集只使用其中的 age、workclass、education、marital-status、occupation、race、sex 和 native-country 等 8 个属性,除 age 为整数数值型外,其他属性均为分类型。根据 income 属性的取值可以将数据分为 2 个类别。在忽略属性值有缺失的记录后,实际参与实验的记录共 30 162 条。CENSUS 数据集的所有属性的 DGH 均为 3 层,Adult 数据集的属性的泛化层次如表 1 所示。

表 1　Adult 数据集属性泛化层次

属性	不同值数	泛化层次
age	72	3
workclass	7	2
education	16	3
marital-status	7	2
occupation	14	2
race	5	1
sex	2	1
native-country	41	2

实验中选择 Classfly 算法作为对比算法，将 K 匿名化后的数据集应用于聚类分析，聚类算法采用 K-Means 算法。KAoGRA 算法在 Matlab 7.0 环境下实现，Classfly 算法基于 MySQL 5.0 数据库实现。实验比较了在不同数据集大小和不同 K 值条件下，聚类精确度和信息损失度的变化。其中，信息损失度采用文献[9]中提出的分辨度损失指标(discernibility metric)来衡量。精确度实验结果如图 2 和图 3 所示。

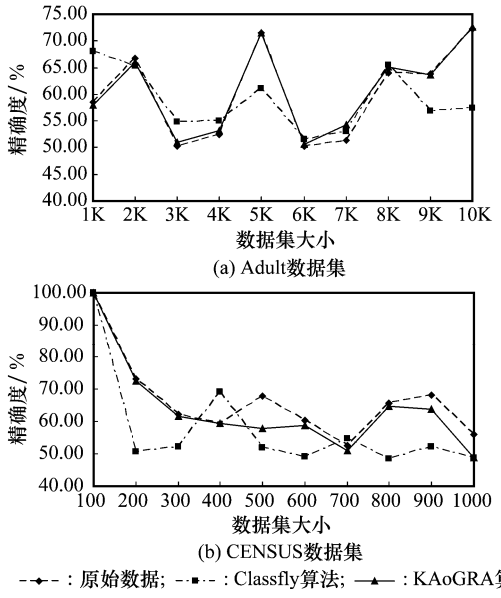

图 2 精确度随数据集大小的变化

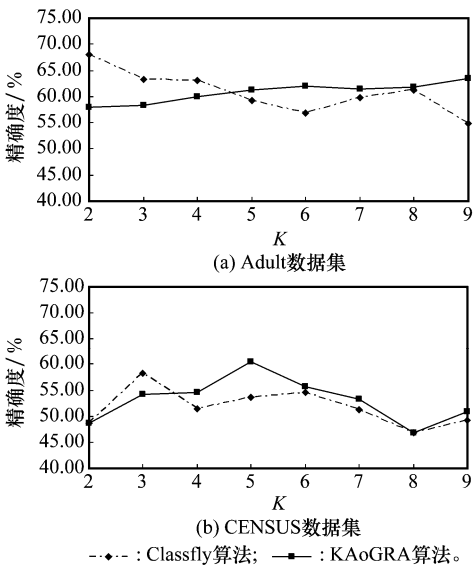

图 3 精确度随 K 值的变化

从图 2 可以看出，KAoGRA 算法的精确度和原始数据的精确度非常接近，在 Adult 数据集上两条曲线几乎重合，在 CENSUS 数据集上 KAoGRA 算法的精确度比原始数据略有下降，但两者都高于 Classfly 算法的精确度。这说明 KAoGRA 算法在保证 K 匿名的条件下，造成的信息损失比较小，因而在 KAoGRA 算法处理后的数据上进行聚类的结果和在原始数据上的聚类结果相比变化较小。图 3 显示了当数据集大小为 1K 时，KAoGRA 算法和 Classfly 算法处理后数据聚类精度的变化。由图 2 可以发现，当 K 值较小时，Classfly 算法有一定优势，而当 K 值大于 3 后，KAoGRA 算法开始优于 Classfly 算法，在平均意义上 KAoGRA 算法的精确度高于 Classfly 算法。此外，Classfly 算法的精确度随着 K 值的增大而逐渐下降，这和第 3 节中的分析相符：K 值越大，泛化的程度就越高，信息损失也就越大，导致在 K 匿名处理的数据集上的聚类精确度降低。而 KAoGRA 算法的表现在两个数据集上有比较大的差异。在 Adult 数据集上，KAoGRA 算法的精确度随着 K 值的增大缓慢上升，这可能是由于 Adult 数据集中的数据只有2 个类别，较易于划分，且 KAoGRA 算法中采用的均衡接近度指标能够较好反映数据点间的相似性，使生成的等价类具有较好的紧凑性，在一定程度上抵消泛化带来的信息损失，使聚类精度略有提高。在 CENSUS 数据集上，KAoGRA 算法精确度先升后降，反映当数据集的类别数较多时，利用均衡接近度描述数据点间相似性带来的优势随着 K 值的增加逐渐被泛化造成信息损失的劣势抵消。当 K 等于 5 时，两者达到平衡，此时聚类精确度最高，之后信息损失的劣势开始超过均衡接近度的优势，精确度逐渐下降。信息损失度实验结果如图 4 和图 5 所示。

从图 4 和图 5 显示的实验结果可以看出，随着数据量的增大以及 K 值的增大，KAoGRA 算法的信息损失程度总是小于 Classfly 算法，反映采用基于灰关联分析的 TG+TS 技术的 KaoGRA 算法相对于采用基于特征类的 AG+TS 技术的 Classfly 算法能够更好地保存原始数据的距离信息，使聚类精确度损失较小，这也与第 3 节中的分析相符。

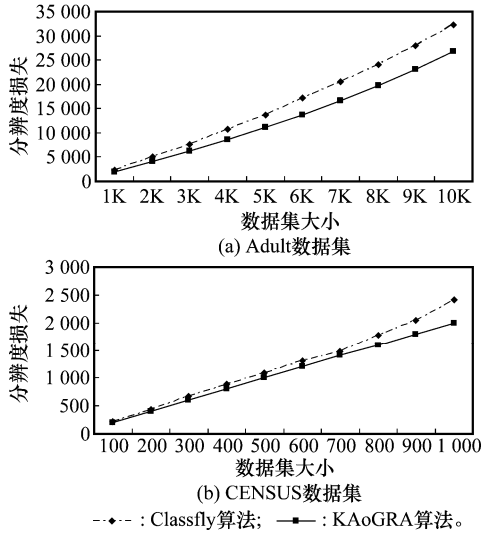

图 4 分辨度损失随数据集大小的变化

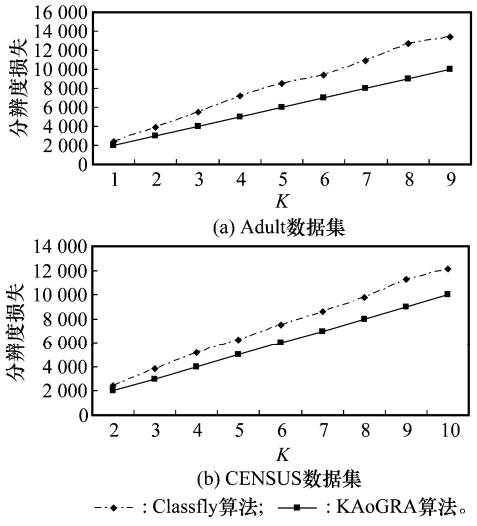

图 5　分辨度损失随 K 值的变化

5　结束语

本文应用灰关联分析方法并结合差异信息理论,给出了一种新的基于灰关联分析的 K 匿名方法,实验结果分析证明了新方法的有效性。未来可以在如下两个方面展开研究:一是将基于灰关联分析的 K 匿名方法拓展到聚类分析以外的分类、关联规则分析等其他数据挖掘应用中;二是将灰色系统理论的其他组成部分(如灰色模型等)应用于 K 匿名的研究。

参考文献:

[1] 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. (Zhou S G, Li F, Tao Y F, et al. Privacy preservation in database applications: a survey[J]. *Chinese Journal of Computers*, 2009, 32(5):847 - 861.)

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

[3] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information[C]// *Proc. of the 7th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 1998.

[4] Ciriani V, De Capitani di Vimercati S, Foresti S, et al. K-anonymity[M]// *Advances in Information Security*. Berlin: Springer, 2007.

[5] Meyerson A, Williams R. On the complexity of optimal K-anonymity[C]// *Proc. of the 23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 2004: 223 - 228.

[6] Aggarwal G, Feder T, Kenthapadi K, et al. Anonymizing tables[C]// *Proc. of the 10th International Conference on Database Theory*, 2005: 246 - 258.

[7] Samarati P. Protecting respondents' identities in microdata release[J]. *IEEE Trans. on Knowledge and Data Engineering*, 2001, 13(6): 1010 - 1027.

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

[9] Bayardo R J, Agrawal R. Data privacy through optimal K-anonymization[C]// *Proc. of the 21st International Conference on Data Engineering*, 2005: 217 - 228.

[10] Lefevre K, Dewitt D J, Ramakrishnan R. Incognito: efficient full-domain K-anonymity[C]// *Proc. of the ACM SIGMOD International Conference on Management of Data*, 2005: 49 - 60.

[11] Lefevre K, Dewitt D J, Ramakrishnan R. Mondrian multidimensional K-anonymity[C]// *Proc. of the 22nd International Conference on Data Engineering*, 2006.

[12] Iyengar V S. Transforming data to satisfy privacy constraints[C]// *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002: 279 - 288.

[13] Winkler W E. Using simulated annealing for K-anonymity[R]. Washington: Statistical Research Division, U. S. Census Bureau, 2002.

[14] Fung B C M, Wang K, Yu P S. Top-down specialization for information and privacy preservationtop-down specialization for information and privacy preservation[C]// *Proc. of the 21st International Conference on Data Engineering*, 2005: 205 - 216.

[15] Wang K, Yu P S, Chakraborty S. Bottom-up generalization: a data mining solution to privacy protection[C]// *Proc. of the 4th IEEE International Conference on Data Mining*, 2004: 249 - 256.

[16] Park H, Shim K. Approximate algorithms for K-anonymity[C]// *Proc. of the ACM SIGMOD International Conference on Management of Data*, 2007.

[17] Liu X Y, Yang X C, Yu G. A representative classes based privacy preserving data publishing approach with high precision[J]. *Computer Science*, 2005, 32(9A): 368 - 373.

[18] Aggarwal C C, Yu P S. *Privacy preserving data mining: models and algorithms*[M]. New York: Springer, 2008.

[19] Friedman A, Wolff R, Schuster A. Providing K-anonymity in data mining[J]. *The International Journal on Very Large Data Bases*, 2008, 17(4): 789 - 804.

[20] 邓聚龙. 灰理论基础[M]. 武汉: 华中科技大学出版社, 2002. (Deng J L. *Elements on grey theory*[M]. Wuhan: Huazhong Univerisity of Science and Technology Press, 2002.)

[21] 张岐山. 灰朦胧集的差异信息理论[M]. 北京: 石油工业出版社, 2002. (Zhang Q S. *Difference information theory of grey hazy set*[M]. Beijing: Petroleum Industry Press, 2002.)

[22] Liu H, Zhang Q S. Life prediction of mechanical products of GM(1,1) based on particle swarm optimization[C]// *Proc. of the IEEE International Conference on Grey Systems and Intelligent Services*, 2007: 409 - 413.

[23] 刘虹, 张岐山. 基于微粒群算法的 GM(2,1,λ,ρ)优化模型[J]. 系统工程理论与实践, 2008, 28(10): 96 - 101. (Liu H, Zhang Q S. GM(2, 1, λ, ρ) based on particle swarm optimization[J]. *Systems Engineering-Theory & Practice*, 2008, 28(10): 96 - 101.)

[24] Zhang Q S, Wang H Y. Measuring the greyness of grey cluster knowledge[J]. *Journal of Grey System*, 2009, 21(3): 259 - 268.

[25] Zhang Q S, Chen K J. Grey sets and their greyness measure[C]// *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*, 2008: 2045 - 2048.

[26] Brand R, Domingo-Ferrer J, Mateo-Sanz J M. Reference data sets to test and compare sdc methods for protection of numerical microdata[EB/OL]. [2010 - 9 - 1]. http://neon.vb.cbs.nl/casc.

[27] Frank A, Asuncion A. UCI machine learning repository[EB/OL]. [2010 - 9 - 1]. http://archive.ics.uci.edu/ml.