Parzen窗核密度估计的大规模数据模式分类隐私保护方法

原永滨[1,2],杨静[1],张健沛[1],于旭[3]

1. 哈尔滨工程大学计算机科学与技术学院,哈尔滨 150001
2. 福州大学电气工程与自动化学院,福州 350108
3. 青岛科技大学信息科学与技术学院,青岛 266001

摘要 针对大规模数据集上的模式分类任务,提出基于Parzen窗核密度估计的模式分类隐私保护算法。利用Parzen窗算法对原始大规模训练集服从的概率密度进行估计,根据估计的概率密度函数构造*la*个替换训练样本,其中*l*为原始样本的数目,*a*通过10折交叉验证方式确定。最后发布替换训练样本进行模式分类,以实现原始数据上的隐私保护。在Adult数据集上的仿真实验充分验证了算法的有效性。

关键词 Parzen窗;核密度估计;数据发布;隐私保护

中图分类号 TP309.2 **文献标志码** A **doi** 10.3981/j.issn.1000-7857.2014.36.017

A Pattern Classification Privacy Preservation Algorithm Based on Parzen Window Kernel Density Estimation for Large Data Set

YUAN Yongbin[1,2], YANG Jing[1], ZHANG Jianpei[1], YU Xu[3]

1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
2. College of Electrical Engineering & Automation, Fuzhou University, Fuzhou 350108, China
3. College of Information Science and Technology, Qingdao University of Science & Technology, Qingdao 266001, China

Abstract In this paper, a pattern classification privacy preservation algorithm is proposed based on the Parzen window kernel density estimation on large scale dataset. Firstly, the probability density is estimated through the original large scale training set. Then the replacement training samples are constructed by the estimated probability. Finally, the replacement training samples are published for the pattern classification training. Thus the privacy on the original training set can be protected effectively. The simulation experiments on Adult datasets fully verify the effectiveness of the proposed algorithm.

Keywords Parzen window; kernel density estimation; data publish; privacy preserving

数据挖掘技术的发展极大地促进了人们对海量数据的利用,同时也引起了数据隐私的泄露[1]。为了进行隐私保护,同时又能对数据中隐藏的有用信息进行挖掘,面向隐私保护的数据挖掘应运而生[2]。本文针对大规模数据集上的模式分类任务,提出了一种基于Parzen窗核密度估计的模式分类隐私保护算法,避免了原始数据上的隐私泄露[3]。

模式分类是对表征事物或现象的各种形式的信息进行处理和分析,以对事物或现象进行描述、辨认、分类和解释的过程,是人类以及动物的最基本的智能表现。随着人类收集和存储数据能力的不断增长,以及计算机运算能力的飞速发

收稿日期:2014-01-09;修回日期:2014-09-03

基金项目:国家自然科学基金项目(61073041,61073043,61370083,61402126);黑龙江省自然科学基金项目(F200901);福建省自然科学基金项目(2011J1296);高等学校博士学科点基金项目(20112304110011,20112304110012)

作者简介:原永滨,副教授,研究方向为隐私保护、机器学习,电子信箱:yyb1688@163.com;于旭(通信作者),博士,研究方向为隐私保护支持向量机,电子信箱:yuxu0523@163.com

引用格式:原永滨, 杨静, 张健沛, 等. Parzen窗核密度估计的大规模数据模式分类隐私保护方法[J]. 科技导报, 2014, 32(36): 104-109.

展,利用计算机来分析数据进行模式分类的要求越来越广泛和迫切。近些年随着研究得深入,出现了许多优秀的分类算法,如人工神经网络(artificial neural network,ANN)[4],支持向量机(support vector machines,SVMs)[5]和决策树(decision tree,DT)[6]等。这些算法的出现极大地促进了模式分类技术在各领域中的应用。

训练样本数据的获取是模式分类工作基础,所以模式分类任务很容易造成一些敏感数据的泄露。为了保护用来分类的训练数据,同时又尽可能不影响模式分类算法的性能,本文提出了一种基于Parzen窗核密度估计的模式分类隐私保护算法。该算法的主要思想是通过核密度估计方法估计原始数据的概率密度分布,然后根据其密度函数生成一定数目的新样本,最后用这些新样本替换原始样本进行训练,实现原始数据的隐藏。因为本文算法针对的是大规模数据集,所以通过Parzen窗核密度估计算法可以较为准确的对原始数据集服从的密度函数进行估计,从而保障了分类器在替换数据集上的学习性能。

1　模式分类及隐私保护

1.1　隐私保护

1.1.1　数据实例

设数据表$D=\{E_x,A_1,A_2,\cdots,A_d,S\}$为一个待发布的数据表,其中$E_x$为显式标识符,$A_i(1\leqslant i\leqslant d)$为准标识符,$S$为敏感属性。$D$中包含$N$个元组,每个元组记作$t_k(1\leqslant k\leqslant n)$。若存在具有相同准标识符属性的元组的集合,则称该集合为数据表D的一个等价类,记作QI。

例如,表1为待发布的原始数据[7],其中"姓名"为显式标识符,"年龄、性别、邮编"为准标识符,"疾病"为敏感属性。

表1　原始数据

Table 1　Table of raw data

姓名	年龄	性别	邮编	疾病
Andy	4	M	12000	胃溃疡
Bill	5	M	14000	消化不良
Ken	6	M	18000	肺炎
Nash	9	M	19000	支气管炎
Alice	12	F	22000	流感
Betty	19	F	24000	肺炎

1.1.2　k-匿名隐私保护算法

为了达到敏感属性数据保护的目的,Sweeney提出了k-匿名原则[8]。对于数据表D,删除显式标识符后,所有等价类中包含的元组个数$\geqslant k$,则称D是k-匿名的。例如,表2中的元组t_1、t_2构成1个等价类,表中的3个等价类均满足2-匿名。一般而言,k值越大,隐私保护效果越好,但信息损失越大。

表2　2-匿名化数据

Table 2　2-anonymity data

年龄	性别	邮编	疾病
[1,5]	M	[10k,15k]	胃溃疡
[1,5]	M	[10k,15k]	消化不良
[6,10]	M	[15k,20k]	肺炎
[6,10]	M	[15k,20k]	支气管炎
[11,20]	F	[20k,25k]	流感
[11,20]	F	[20k,25k]	肺炎

1.1.3　l-多样性隐私保护算法

设数据表$D=\{E_x,A_1,A_2,\cdots,A_d,S\}$为待发布的数据表,其中$E_x$为显式标识符,$A_i(1\leqslant i\leqslant d)$为准标识符,$S$为敏感属性。若存在具有相同准标识符属性的元组的集合,则称该集合为数据表D的一个等价类记作QI。若对$\forall s\in S$,设(QI,s)为等价类QI中包含敏感值s的元组的集合,若对任意的QI,都有$|QI,\ s|\geqslant l(l\geqslant 2)$,则称$D$满足$l$-多样性[9]。$l$-多样性模型使得攻击者最多以$1/l$的概率确认某个体的敏感信息。同样,表2发布的数据也是满足2-多样性,即每一个等价类中至少有2个不同的敏感属性值。

1.1.4　基于噪声添加的隐私保护分类算法

在模式分类过程中,需要对原始训练样本进行学习和对新样本进行预测,这都会造成数据隐私的泄露。目前提出的k-匿名隐私保护算法和l-多样性隐私保护算法等的主要思想是通过数据泛化实现个体敏感信息的隐藏,但数据的泛化破坏了原始数据信息,使得模式分类算法无法良好地运行。

为了实现具有分类性能的隐私保护,Agrawal等[10]提出了满足独立同分布的噪音添加方法,为了描述方便,称之为ASN算法。ASN算法的主要思想是对于每一个原始样本的每一维属性值添加服从正态分布的噪声,以实现原始样本隐私保护的目的。然而该算法存在不足:如果噪声范围过大,隐私保护效果较为明显,但原始数据值的有效性受到破坏,导致分类算法效果不理想;如果噪声范围太小,则隐私保护效果就不够理想,很容易通过发布后的数据推算原始数据信息。

1.2　核密度估计

核密度估计在概率论中用来估计未知的密度函数,属于非参数检验方法,由Rosenblatt和Parzen提出。该方法又称为Parzen窗方法。核密度估计的主要思想是通过某范围内各点密度的均值对总体密度函数进行估计,该方法能够较好的描述多维数据的分布状态。

一个向量$\boldsymbol{x}$落在区域R中的概率P为

$$P=\int_R p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{1}$$

因此,可以通过统计概率估计概率密度函数$p(\boldsymbol{x})$。假设N个样本的集合$X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$是根据概率密度函数为$p(\boldsymbol{x})$的分

布独立抽取得到的。那么,有k个样本落在区域R中的概率服从二项式定理:

$$P_k=\binom{N}{k}P^k(1-P)^{N-k} \tag{2}$$

对P的估计为$\hat{P}=k/N$。

假设$p(\boldsymbol{x})$是连续的,且R足够小使得$p(\boldsymbol{x})$在R内几乎没有变化。令R是包含样本点$\boldsymbol{x}$的一个区域,其体积为V,设有N个训练样本,其中有k落在区域R中,则可对概率密度做出估计:

$$P=\int_R p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=p(\boldsymbol{x})V \tag{3}$$

$$\hat{P}=\frac{k}{N} \tag{4}$$

$$\hat{p}(\boldsymbol{x})=\frac{k/N}{V} \tag{5}$$

当N固定时,V的大小对估计的效果影响很大。过大则平滑过多,不够精确;过小则可能导致在此区域内无样本点,k=0。

假设$\mathbf{R}_n$是一个d维超立方体。令h_n为超立方体一条边的长度,则体积$V_n=h_n^d$。

立方体窗函数为

$$\varphi(\boldsymbol{u})=\begin{cases}1 & |\boldsymbol{u}_j|\leqslant\frac{1}{2}\ (j=1,2,\cdots,d)\\ 0 & \text{其他}\end{cases} \tag{6}$$

落入以$\boldsymbol{x}$为中心的立方体区域的样本数为

$$k_n=\sum_{i=1}^{n}\varphi\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h_n}\right) \tag{7}$$

$\boldsymbol{x}$处的密度估计为

$$\hat{p}_n(\boldsymbol{x})=\frac{k_n/n}{V_n}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h_n}\right) \tag{8}$$

可以验证

$$\hat{p}_n(\boldsymbol{x})\geqslant 0 \tag{9}$$

$$\int\hat{p}_n(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1 \tag{10}$$

Parzen窗估计过程是一个内插过程,样本$\boldsymbol{x}_i$距离$\boldsymbol{x}$越近,对概率密度估计的贡献越大,越远贡献越小。

只要满足如下条件,就可以作为窗函数

$$\phi(\boldsymbol{u})\geqslant 0 \tag{11}$$

$$\int\phi(\boldsymbol{u})\mathrm{d}\boldsymbol{u}=1 \tag{12}$$

常见的窗函数为

方窗函数

$$\phi(\boldsymbol{u})=\begin{cases}1 & |\boldsymbol{u}|\leqslant\frac{1}{2}\\ 0 & \text{其他}\end{cases} \tag{13}$$

正态窗函数

$$\phi(\boldsymbol{u})=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\boldsymbol{u}^2\right) \tag{14}$$

指数窗函数

$$\phi(\boldsymbol{u})=\exp(-|\boldsymbol{u}|) \tag{15}$$

2 算法设计

模式分类中的训练数据通常包括很多属性,其中有很多涉及到个人的隐私信息,如收入和信用级别等,所以原始数据的公开很容易造成个人隐私的泄露。如何在不泄露原始训练数据的情况下得到满意的分类决策标准,就成了亟需解决的问题,具有很高的研究价值。

本文提出一种基于Parzen窗核密度估计的模式分类隐私保护算法(a pattern classification privacy preserve algorithm based on Parzen window kernel density estimation, CPPPW)。该算法首先利用Parzen窗核密度估计算法对原始训练样本所服从的数据分布进行密度估计,然后根据该密度函数生成一定数目的替换样本。算法设定生成la个替换样本,其中l为原始训练样本的个数,a代表构造样本的数目和原始样本数据的比例。综合考虑在替换样本集上分类算法的分类性能和运行效率,设定$a\in[1,2]$,即生成替换样本的个数不少于原始样本的个数,同时不多于原始样本数目的两倍。根据10折交叉验证[11]方式确定最合理的a值。最后用这些新样本替换原始样本进行分类学习。

以二分类模式分类为例,基于核密度估计原始数据替换的数据分类隐私保护算法:CPPPW算法的伪码实现如下。

输入:原始样本集合。

$T=\{(\boldsymbol{x}_i,0),i=1,2,\cdots,n_0\}\cup\{(\boldsymbol{x}_i,1),i=1,2,\cdots,n_1\}$,基分类器M,Parzen窗函数$\varphi(\boldsymbol{x})$;式中,$\boldsymbol{x}_i$为训练样本;$n_0$为第一类样本个数;$n_1$为第二类样本个数。

输出:分类决策函数F。

方法:

1)l=n_0+n_1。

2)针对D维向量进行相应的正态窗函数赋值:

$$\varphi(\boldsymbol{x})=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{u})\right]=\frac{1}{(2\pi)^{D/2}}\exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right);$$

//选择正态窗函数对窗函数进行赋值,其中均值向量为零向量,协方差矩阵取单位矩阵。式中,$\varphi(\boldsymbol{x})$为正态窗函数;$\boldsymbol{u}$为均值向量;$\boldsymbol{x}$为D维向量;Σ为协方差矩阵。

3)利用如下核密度估计算子进行密度估计:

$$p(\boldsymbol{x})=\frac{1}{(2\pi)^{D/2}h^D N}\sum_{n=1}^{N}\exp\left(-\frac{\boldsymbol{x}-\boldsymbol{x}_n}{2h^2}^2\right)$$

//对训练集T利用Parzen窗核密度估计方法估计密度函数$p(\boldsymbol{x})$。式中,h为核函数的宽度参数;$\boldsymbol{x}_n$为训练样本。

4)RS=Sample_Generation($p(\boldsymbol{x})$,la);//根据密度函数$p(\boldsymbol{x})$生成la个替换训练样本,得到替换样本集RS,其中根据交叉验证方式确定最合理的a数值。

5)F=M(RS);//利用分类器M对替换样本集RS进行学习,获得分类决策函数F。

由于本算法针对的是大规模数据集,概率密度函数可以得到较为准确的估计,从而使得分类器在替换数据集的分类

性能得到有效地保障。同时该算法利用替换样本集RS进行分类学习,有效地避免了原始样本数据信息的泄露。

3　实验验证

3.1　实验1

3.1.1　数据来源及处理

选用UCI标准机器学习数据库中的Adult数据集进行实验。该数据集的目的是根据统计数据预测收入是否超过50k美元,共包含48842个样本,其中3620个样本包含缺失数据。数据集有14个属性,其中6个为连续属性,8个为标称属性。Adult数据集见表3。

表3　Adult数据集介绍

Table 3　Description of adult data set

属性	属性类型	标称属性取值
age	连续	N/A
workclass	标称	Private, Self-emp-not-inc, Self-emp-inc等8个
fnlwgt	连续	N/A
education	标称	Bachelors, Some-college, 11th, HS-grad等16个
education-num	连续	N/A
marital-status	标称	Married-civ-spouse, Divorced, Never-married等7个
occupation	标称	Tech-support, Craft-repair, Other-service等14个
relationship	标称	Wife, Own-child, Husband等6个
race	标称	White, Asian-Pac-Islander, Amer-Indian-Eskimo等5个
sex	标称	Female, Male
capital-gain	连续	N/A
capital-loss	连续	N/A
hours-per-week	连续	N/A
native-country	标称	United-States, Cambodia, England等41个

首先对数据集进行预处理,将具有缺失属性的数据记录删除,然后从处理后的数据中选取9000个元组进行实验,其中6000个作为训练样本,3000个作为测试样本。

为了避免分类过程中造成的数据泄露,采用Parzen窗核密度估计算法进行原始样本处理。为了使用Parzen窗核密度估计算法,实验首先对所选数据进行标称属性的数值化处理。本文将标称属性的属性值离散化为从1到N的自然数,其中N对应该标称属性的取值个数。这样将原始数据集转变为$\mathbf{R}^{14}$向量空间内的向量。

3.1.2　分类性能评价指标

为了更精确地对算法的性能进行评价,不采用传统的分类准确率作为评价指标,而是选择正确率(precision,P)和召回率(recall,R)作为评价指标。计算公式为

$$P=\frac{n_1}{n_2} \tag{16}$$

$$R=\frac{n_1}{n_3} \tag{17}$$

式中,n_1为事实属于此类且被分类正确的样本数目;n_2为被判为此类的样本数;n_3为属于此类的总样本数。可以看出,只有算法的正确率和召回率都较高时,算法的性能才更优越。

3.1.3　实验方法

实验平台为Intel Core2 Duo CPU T6500,2.10 GHz,2.00 GB RAM,Windows 7操作系统,选择Matlab7.0软件进行实验。分别在原始训练集合上和替换数据集合上进行分类学习,其中替换数据利用本文算法生成。生成的替换样本个数为la,具体地,生成n_0a个第一类样本,n_1a个第二类样本,$l=n_0+n_1$。当a=1时表示生成与原始样本数目一致的替换样本,当a=2时表示生成的替换样本数目是原始样本数目的两倍。采用10折交叉验证方式确定最合理的a数值,求得a=1.6。

为了说明,本文提出的CPPPW算法是一种通用的模式分类隐私保护算法(即对各种不同的分类器均有效),本文采取当前最为经典的3种分类器进行作为基分类器,即人工神经网络分类器、决策树分类器和支持向量机分类器。其中人工神经网络采用BP算法,并设定神经网络结构为3层,其中隐含层设定节点个数为5个,设定人工神经网络步长为0.1,迭代次数设为1000。

决策树使用C4.5决策树算法,对于连续属性C4.5算法将属性进行排序,选择相邻值得中点进行分裂,并最终确定具有最大信息增益率的分界点,另外本文C4.5算法采用了一种后剪枝方法。

支持向量机采用C-SVM分类算法,并使用如下高斯核函数作为分类核函数

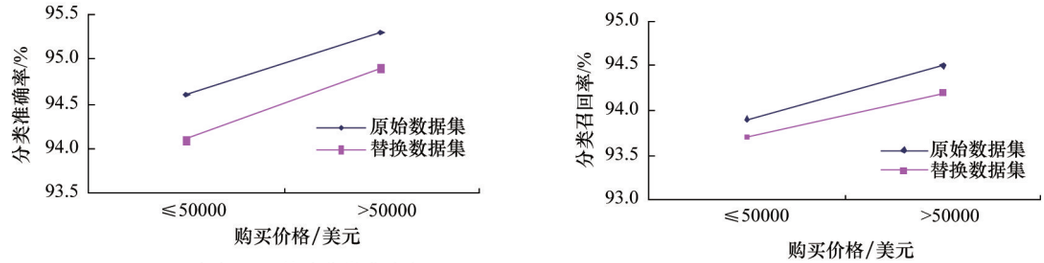
$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-g\ \|\boldsymbol{x}-\boldsymbol{y}\|^2\right) \tag{18}$$

式中,g与C(惩罚因子)为可调参数。同样通过10折交叉验证求得最合适的g和C值,求得g=1/256,C=1000。

3.1.4　实验结果与分析

由于本文算法使用新生成的样本替换原始样本进行学习,所以算法隐私保护的效果是显然的,图1~图3仅给出在替换数据集和原始数据集上,各种分类算法的分类性能。

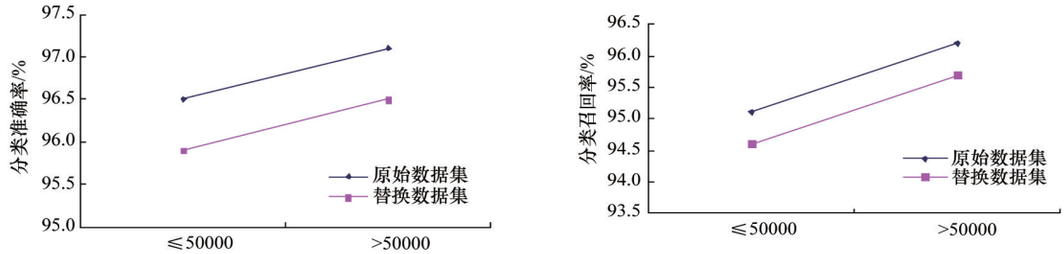
由图1~图3可以看出,3种经典的分类算法在替换数据集上同样可以取得较好的分类性能。因为大规模数据集使得Parzen窗算法能够较好地对样本的分布函数进行估计,从而保障了替换数据集的质量。又考虑到本文算法使用替换数据集代替原始数据集,避免了用户隐私数据的泄露,所以是一种有效的面向隐私保护的数据分类算法。本实验也充分说明本文算法是一种独立于分类器的模式分类隐私保护算法,可以与经典分类器结合,构建不同分类器算法下的隐私保护模型。



(a) C4.5算法分类准确率　　(b) C4.5算法分类召回率

图1　两种数据集上C4.5算法分类准确率及召回率

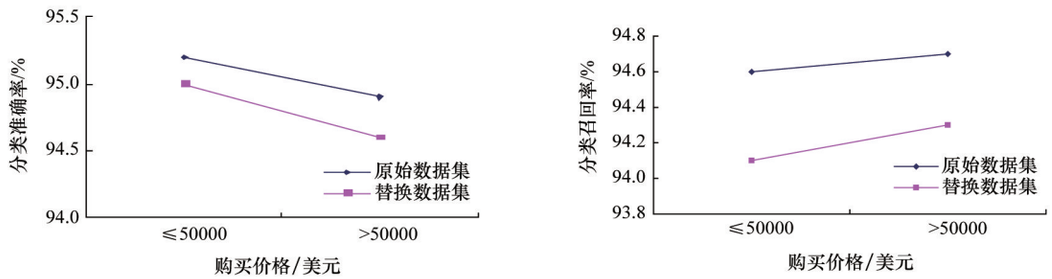
Fig. 1　Classification recall rate comparison of C4.5 algorithm in two datasets



(a) SVM算法分类准确率　　(b) SVM算法分类召回率

图2　两种数据集的SVM算法分类准确率及召回率

Fig. 2　Classification recall rate comparison of SVM algorithm in two datasets



(a) BP算法分类准确率　　(b) BP算法分类召回率

图3　两种数据集的BP算法分类准确率及召回率

Fig. 3　Classification recall rate comparison of BP algorithm in two datasets

3.2　实验2

3.2.1　数据来源与数据描述

利用Car评估数据集进行实验。Car Evaluation数据集来自于UCI标准数据库，是模式识别领域应用最为普遍的数据集之一，包含6个条件属性，分别是buying，maint，doors，persons，lug_boot，safety。buying属性指购买价格，maint指维护费用，doors指汽车的门数，persons指汽车容纳人数，lug_boot指汽车行李箱的大小，safety指汽车的安全性。数据集共包含4类，分别是unacc，acc，good，vgood，其中unacc指结果不可以接受，acc为可以接受，good为比较满意，vgood为非常满意。数据集共有1728条数据，其中unacc和acc类别样本数目较多，分别是1210条和384条。由于本文算法针对的是大规模数据，为了获得较好的实验效果，仅选取unacc和acc类别样本进行实验。本实验选取1000条unacc数据和300条acc数据做训练数据，将剩余的数据作为测试数据。

3.2.2　实验方法

在3.1节，通过实验对比了本文设计的隐私保护分类算法(CPPPW算法)和传统分类算法的分类性能。为了更充分地测试CPPPW算法的隐私保护性能和分类性能，将CPPPW算法与具有隐私保护性能的分类算法进行对比分析。尽管在1.1小节中综述了几种隐私保护方法，如k-匿名和l-多样性，但不具有分类效果，所以选取具有分类性能的隐私保护算法ASN算法进行对比。下面将从理论上对两种方法的隐私保护性能进行分析对比，并通过分类实验结果对它们的分类性能进行比较。

实验中CPPPW算法仍然选取正态窗函数作为窗函数，

为简单起见,其中均值向量为零向量,协方差矩阵取单位矩阵。选取正态窗函数如下:

$$\varphi(\boldsymbol{x})=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{u})\right]$$

$$=\frac{1}{(2\pi)^{D/2}}\exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right)$$

对于ASN算法,选择 $N(0,\sigma^2)$ 正态分布进行噪声生成,并将噪声添加到每一维属性值上。

3.2.3　两种算法分类性能分析

首先给出两种算法的分类结果,如表4所示,其中 P_1 和 P_2 分别表示本文算法和ASN算法的分类准确率,R_1 和 R_2 分别表示本文算法和ASN算法的分类召回率。

表4　两种算法的分类结果对比

Table 4　Comparison of results processed by two algorithms

数据类别	分类结果			
	P_1	P_2	R_1	R_2
unacc	95.3	94.3	94.6	93.1
acc	94.1	93.3	93.8	92.2

σ 太大导致噪声过分地影响原始的训练样本数据,造成分类精度不高。σ 太小原始数据信息几乎不发生改变,隐私保护效果不明显。综合考虑分类算法的隐私保护效果和分类精度,设定 $\sigma=5$。另外与实验1类似,本文CPPPW算法基于10折交叉验证方式确定最合理的 a 数值,a=1.8。

从表4可以看出,本文算法在分类性能上要优于ASN算法,这主要是因为ASN算法通过加入噪声,破坏了原始数据的有效性,对分类性能产生了不好的影响。而本文算法虽然也没有利用精确的原始数据进行学习,但是由于数据规模较大,使得估计的概率密度较为精确,为新生成的数据的有效性提供了保障,从而保证了分类算法的分类性能。

3.2.4　两种算法隐私保护性能的理论分析

为了评估隐私保护方法的隐私保护性能,引入Agrawal和Srikant提出的隐私保护性能度量指标[10],其思想是根据被保护属性的原始值推算出的可能性评价隐私保护算法的隐私保护性能,更严格的定义如下。

对于某种隐私保护算法,如果原始值能够以 c% 的置信度位于置信区间 $[x_1,x_2]$ 中,则定义该算法的隐私保护性能为 $\{c\%,x_2-x_1\}$。

根据定义,显然 c 越小,置信区间 $[x_1,x_2]$ 的区间长度越大,则隐私保护算法的隐私保护性能越好。同时可知ASN算法在加入服从正态分布 $N(0,\sigma^2)$ 的噪声时,隐私保护性能为 $\{50\%,1.34\sigma\}$,也就是说ASN隐私保护算法以50%的置信度确定扰动后的数值 x,其真值的区间为 $[x-0.67\sigma,x+0.67\sigma]$。当标准差 σ 较小时,原始数据和隐私保护后的数据差别不大,隐私保护效果不够理想。

本文CPPPW算法由于从概率密度函数重新构造生成了一批新样本,与原始数据没有直接的数据联系,难以由新样本数据反推得到原始数据,所以隐私保护性能优于ASN算法。

4　结论

针对大规模数据集,本文提出了一种基于Parzen窗核密度估计的模式分类隐私保护算法。充足的训练样本使得Parzen窗核密度估计算法可以较准确的估计密度函数,保障了替换数据集的质量。在替换数据集进行分类学习,有效地避免了原始数据上的隐私泄露,并且较好地保持了分类算法的分类性能。本文算法有效的前提是数据集包含大量地样本,研究在小样本数据集上有效的模式分类隐私保护算法将是进一步的研究内容。

参考文献(References)

[1] Han J W, Kamber M. Data mining: Concepts and techniques[M]. San Francisco, CA: Morgan Kaufmann, 2001: 257-259.

[2] 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 Shuigeng, Li Feng, Tao Yufei, et al. Privacy preservation in database applications: A survey[J]. Chinese Journal of Computers, 2009, 32(5): 847-861.

[3] 周恩策, 刘纯平, 张玲燕, 等. 基于时间窗的自适应核密度估计运动检测方法[J]. 通信学报, 2011, 3(2): 106-114.
Zhou Ence, Liu Chunping, Zhang Lingyan, et al. Foreground object detection based on time information window adaptive kernel density estimation[J]. Journal on Communications, 2011, 3(2): 106-114.

[4] Yang J, Yu X, Xie Z Q. A novel virtual sample generation method based on Gaussian distribution[J]. Knowledge-Based Systems, 2011, 24 (6): 740-748.

[5] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20(8): 273-297.

[6] Quinlan J R. C4.5: Programs for Machine Learning[M]. San Mateo, CA: Morgan Kaufmann, 1993, 17-69.

[7] Xiao X, Tao Y. Personalized privacy preservation[C]//Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. Illinois, Chicago: ACM, 2006: 229-240.

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

[9] Machanavajjhala A, Kifer D, Gehrke J, et al. L-diversity: Privacy beyond *K*-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007(1): 3-15.

[10] Agrawal R, Srikant R. Privacy-preserving data mining[J]. ACM Sigmod Record, 2000, 29(2): 439-450.

[11] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 1995, 14(2): 1137-1145.

(编辑　季超)