一种基于主成分分析的稀疏数据模式分类隐私保护算法

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 模式分类过程涉及到对原始训练样本的学习,容易导致用户隐私的泄露。为了避免模式分类过程中的隐私泄露,同时又不影响模式分类算法的性能,提出一种基于主成分分析(PCA)的模式分类隐私保护算法。该算法利用PCA提取原始训练数据的主成分,并将原始训练样本集合转化为主成分的新样本集合,然后利用新样本集合进行分类学习。选用Adult 数据集和KDD CUP 99数据集进行仿真实验,并采用正确率和召回率进行性能评价,结果表明,该隐私保护算法通过PCA提取原始数据特征属性的主成分,可避免原始属性的泄露,同时PCA在一定程度上可实现去噪,从而使分类器的分类性能优于原始数据集的分类性能。与已有算法比较,该隐私保护算法具有更好的模式分类精度和隐私保护性能。

关键词 主成分分析;模式分类;隐私保护算法

中图分类号 TP391 **文献标志码** A **doi** 10.3981/j.issn.1000-7857.2014.12.010

A Pattern Classification Privacy Preserve Algorithm for Sparse Data Based on Primary Component Analysis

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. School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266001, China

Abstract The pattern classification process involves the learning from the original training samples, which easily leads to privacy disclosure. In order to avoid the leaks of privacy in the pattern classification process and not to affect the performance of the algorithm, this paper proposes a pattern classification privacy preserve algorithm based on the primary component analysis (PCA). This algorithm extracts the principal component of the original training data and converts the original training samples to new samples corresponding to the primary components. Then, a classification model is trained on the new samples. Experiments are carried out on the Adult data set and the KDD CUP 99 data set, and the precision and recall indexes are used to evaluate the proposed algorithm. It is shown that this algorithm can avoid the leakage of the original attributes through extracting the principal components of the feature attributes about the raw data. PCA can achieve de-noising to some extent, so that the classification performance on the classifier is better than that on the original data set. Therefore, compared with the existing algorithms, this algorithm has better pattern classification accuracy and privacy preserve performance.

Keywords primary component analysis; pattern classification; privacy preserve algorithms

收稿日期:2014-01-06;修回日期:2014-03-16

基金项目:国家自然科学基金项目(61370083,61073043,61073041);高等学校博士学科点专项科研基金(20112304110011,20122304110012)

作者简介:原永滨,副教授,研究方向为隐私保护及机器学习,电子信箱:yuanyongbin74@163.com;于旭(通信作者),博士,研究方向为隐私保护及支持向量机,电子信箱:yuxu0532@163.com

引用格式:原永滨, 杨静, 张健沛, 等. 一种基于主成分分析的稀疏数据模式分类隐私保护算法[J]. 科技导报, 2014, 32(12): 68-73.

数据挖掘[1]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。为了进行隐私保护[2~4],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。

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,是最基本的智能表现。随着收集、存储数据技术以及计算机运算技术的飞速发展,利用计算机分析数据进行模式分类的要求越来越广泛。近年来,在模式分类的研究方面,出现了许多优秀的分类算法,如人工神经网络(artificial neural network, ANN)[5~7]、支持向量机(support vector machines,SVMs)[8~10]和决策树(decision tree,DT)[11~13]等。这些算法极大地促进了模式分类在各领域中的应用。

然而在模式分类过程中,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,已经有很多数据隐私保护方法,如Sweeney等[2]提出的k-匿名模型(k-anonymity),Machanavajjhala等[3]提出的l-多样性模型(l-diversity)等,这几种模型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。Agrawal等[14]提出了满足独立同分布的噪声添加方法,而Kargupta等[15]则指出添加的噪声很容易被过滤掉,且加入噪声后模式分类的效果变得不够理想。Bapna1等[16]提出一种基于小波变化的隐私保护分类方法,该方法首先利用小波变换对原始分类数据进行处理,得到原始数据集合的近似压缩数据,然后对压缩数据进行分类学习,由于难以从压缩数据中反推原始数据,该方法具有较好的隐私保护性能,但该方法借助于小波变换对原始分类数据进行处理,受限于小波变换处理稀疏数据的不足[1],该方法的分类性能不够理想。胡文军等[17]提出一种隐私保护的SVM快速分类方法,可以实现有效的模式分类,并且能够避免分类结果中支持向量的泄漏,但该方法却依然要依赖于对原始数据的学习,因此不可避免地会造成原始数据隐私的泄漏。针对上述问题,本文研究一种基于主成分分析的模式分类隐私保护算法。

1　隐私保护与主成分分析

1.1　隐私保护

1.1.1　相关概念

定义1　设数据表D={E_x, A_1, A_2, …, 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。数据表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[18],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　待发布的原始数据[18]

Table 1　Raw data of wait for release

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-匿名原则

为了达到敏感属性数据保护的目的,Sweeney提出了k-匿名原则。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大于或者等于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-多样性模型

定义2　设数据表D={E_x, A_1, …, 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。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。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-diversity的,即每一个等价类中至少有2个不同的敏感属性值。

1.1.4　隐私保护性能度量指标

为了评估隐私保护方法的隐私保护性能,引入Agrawal等[14]提出的隐私保护性能度量指标,其思想是根据被保护属性的原始值被推算出的可能性,评价隐私保护算法的隐私保护性能,则更严格的定义如下。

定义3 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3,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。

1.2 主成分分析方法

1.2.1 主成分分析的基本原理

主成分分析(Primary Component Analysis,PCA)[19]的主要目的是,从k个特征属性中找出最能代表原始特征属性的c个新属性。与特征选择算法不同的是,主成分分析方法得到的属性并不是原始属性的一个子集,该方法得到的c个新属性中的每一个属性都是对原始属性进行组合产生的。原始属性集合,经过主成分分析方法处理后,可以被投影到一个较小的数据集合中,因此主成分分析是一种有效的降维方法。

1.2.2 主成分分析的求解

设原始样本集合组成的矩阵为

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

其中,包含n个样本,每个样本共有p个特征。主成分分析的目的是将原始变量映射到一个新的空间,记$x_1, x_2, \cdots, x_p$为原变量指标,$z_1, z_2, \cdots, z_m(m\leqslant p)$为新变量指标,则

$$\begin{cases} z_1=l_{11}x_1+l_{12}x_2+\cdots+l_{1p}x_p \\ z_2=l_{21}x_1+l_{22}x_2+\cdots+l_{2p}x_p \\ \vdots \\ z_m=l_{m1}x_1+l_{m2}x_2+\cdots+l_{mp}x_p \end{cases} \tag{2}$$

主成分分析要实现两个目标:1) z_i与$z_j(i\neq j; i,j=1, 2, \cdots, m)$不相关;2) z_1是$x_1, x_2, \cdots, x_p$的所有的线性组合中方差最大者,z_2是与z_1不相关的$x_1, x_2, \cdots, x_p$的一些线性组合中方差最大者,依此类推,z_m是与$z_1, z_2, \cdots, z_{m-1}$都不相关的$x_1, x_2, \cdots, x_p$的一些线性组合中方差最大者。根据这两个目标,即可确定系数l_{ij}。

2 隐私保护算法设计

模式分类中的训练数据通常包括诸多属性,其中有很多涉及到个人的隐私信息,如收入和信用级别等。尽管通过去除用户ID,可在一定程度上避免个人身份的泄露,但是准标示符同样可以泄露用户的身份信息,所以原始数据的公开很容易造成个人隐私的泄露。传统的k-匿名原则和l-多样性模型在进行泛化操作时,会破坏原始数据的信息,使得分类算法无法运行。为了解决这个问题,本文提出一种基于主成分分析的稀疏数据模式分类隐私保护算法(a pattern classification privacy preserve algorithm for sparse data based on primary component analysis,CPPPCA),首先利用主成分分析法提取原始训练数据的主成分,并将原始训练数据集合转化为主成分的新数据集合,然后发布新数据集合进行分类学习。CPPPCA隐私保护算法的实现方法如下。

输入原始样本数据集合F、原始样本类别集合L、阈值a和基分类器M。

输出分类决策函数f。

步骤1 对于原始样本数据集合组成的矩阵$\boldsymbol{F}$中的第i列,计算均值$m(i)$和标准差$s(i)$。

步骤2 对于矩阵$\boldsymbol{F}$中的元素$F(i, j)$,利用如下公式进行标准化处理:

$$F(i,j)=\frac{F(i,j)-m(i)}{s(i)} \tag{3}$$

步骤3 计算矩阵$\boldsymbol{F}$的相关系数矩阵

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \tag{4}$$

式中

$$r_{ij}=\frac{1}{n-1}\sum_{t=1}^{n}x_{ti}x_{tj} \qquad (i,j=1, 2,\cdots, p) \tag{5}$$

步骤4 利用雅克比方法求相关系数矩阵$\boldsymbol{R}$的特征值($\lambda_1, \lambda_2,\cdots, \lambda_p$)和相应的特征向量

$$a_i=(a_{i1}, a_{i2},\cdots,a_{ip}) \qquad (i=1, 2,\cdots, p) \tag{6}$$

步骤5 选取累计贡献率b大于阈值a的前k个主成分,其中,前k个主成分的累计贡献率为

$$b_k=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{7}$$

步骤6 根据所选主成分,重新构造训练数据集合F'。

步骤7 利用基分类器M对训练数据集合F'进行学习,得到决策函数f,即

$$f=M(F') \tag{8}$$

通过主成分分析方法提取原始特征属性的主成分,避免了原始属性值的泄露,同时由于主成分分析方法在一定程度上可以实现去噪,因此分类器的分类性能在一定程度上有可能高于原始数据集的分类性能。下面通过充分的实验对CPPPCA算法的性能进行测试。

3 实验测试

3.1 实验I

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人们的统计数据来预测收入是否超过5万美元。该数据集共包含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。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了9000个元组进行实验,其中6000个元组作为训练样本,3000个元组作为测试样本。由于数据集包括age、work class、education、marital-status、occupation等明显涉及到个人隐私的

属性,很容易在分类的同时造成个人隐私的泄露。

3.1.2　分类性能评价指标

为了更精确地评价算法的性能,本实验并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)作为本实验分类的性能评价指标。计算公式为

$$P=\frac{n_1}{n_2} \tag{9}$$

$$R=\frac{n_1}{n_3} \tag{10}$$

式中,n_1为事实属于此类且被分类正确的样本数;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。采用当前最为经典的3种分类器作为实验的基分类器,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,人工神经网络采用后向传播算法(back propagation,BP),并设定神经网络结构为3层,决策树采用C4.5决策树算法,支持向量机采用C-SVM分类算法,并采用高斯核函数作为分类核函数,即

$$K(x,y)=\exp(-g\|x-y\|^2) \tag{11}$$

其中,g、C(惩罚因子)为可调参数,通过10折交叉验证[20]来求得最合适的g、C值,即将数据集分成10份,轮流将其中9份作为训练数据,1份作为测试数据,进行实验。每次实验都会得出相应的正确率,将10次实验结果的正确率的平均值作为对算法精度的估计。

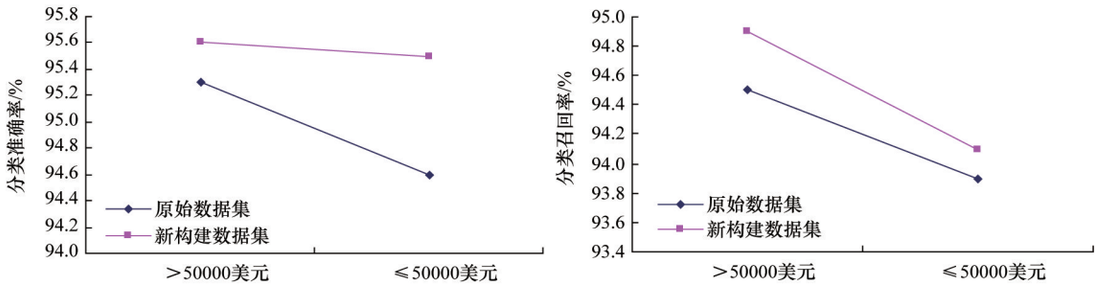
3.1.4　结果与分析

3种分类算法BP、C4.5和C-SVM在原始训练数据集合F'和新构造数据集合F'上的实验结果如图1、图2和图3所示。

本文CPPPCA算法改变了原始数据集合的属性和数值,在构造的新样本集合F'上进行分类学习,其良好的隐私保护效果是显然的。下面主要分析传统的BP、C4.5和C-SVM3种分类算法(对应原始数据集)和本文CPPPCA算法(对应新构造数据集)的分类性能。

图1、图2和图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。

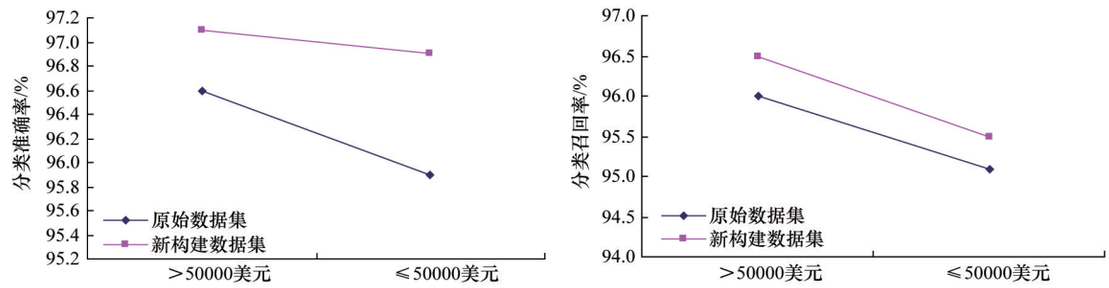
从图1、图2和图3可以看出,3种经典的分类算法在新构建数据集上获得了更好的分类性能,这是因为在新构建数据集上,分类算法不易陷入过拟合。另外,由于本文CPPPCA算法使用新构建数据集代替原始数据集,避免了用户隐私数据的泄露,所以本文算法是一种有效的面向隐私保护的数据分类算法。实验也充分证明本文算法是一种独立于分类器的模式分类隐私保护算法,可与经典分类器结合,构建不同分类器算法下的隐私保护模型。



(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

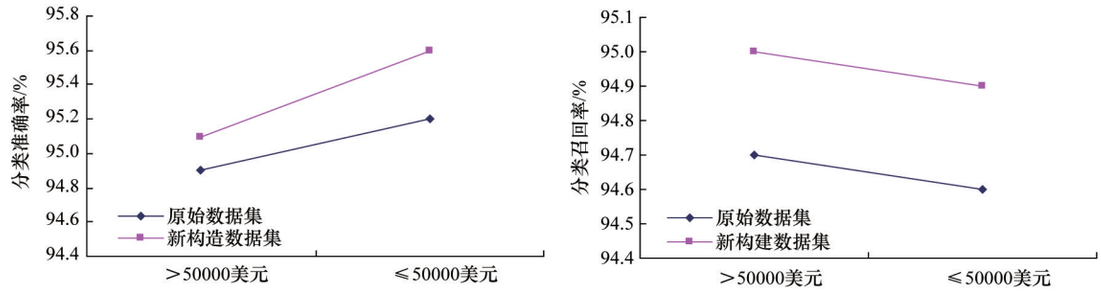
Fig. 1　Classification precision and recall rate comparison of C4.5 algorithm in two datasets



(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 precision and 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 precision and recall rate comparison of BP algorithm in two datasets

3.2 实验Ⅱ

3.2.1 实验数据

为了进一步测试本文CPPPCA算法的性能,利用入侵检测数据集进行分类实验,并将本文算法与文献[14]提出的满足独立同分布的噪声添加方法以及文献[16]中提出的基于小波变换的隐私保护分类方法进行对比。为了方便描述,将文献[14]的方法记为ASN算法,将文献[16]的方法记为WT算法。本实验的入侵检测数据kdd-cup_10_per来源于KDD CUP 99,包含494021条记录。每条记录包含41个属性,其中34个是连续属性,剩余的7个属性是离散的。数据集包含23个类别,其中Normal是正常网络行为,其余22类(如Back、Neptune、Smurf等)是入侵行为。将23类行为映射到5大类,即Normal、Dos、R2L、U2R和Probing。训练数据和测试数据的分布如表3所示。

表3 数据集的数据分布

Table 3 Distribution of data set

类型	训练样本	测试样本
Normal	900	960
Dos	3700	3790
Probing	600	800
R2L	300	398
U2R	30	22

3.2.2 实验方法

由于数据集包含4个字符属性,因此对于字符属性需要进行数值转化。具体方法参照文献[21]中给出的方法。由于原始数据集中每个属性的取值范围是不同的,所以需要对数据进行标准化处理,将连续属性的值映射到区间[0.0,1.0]。标准化公式为

$$V=\frac{v-\min(f_i)}{\max(f_i)-\min(f_i)} \tag{12}$$

式中,V为标准化后的属性值;v为原始数据的属性值;$\min(f_i)$、$\max(f_i)$分别是属性f_i的最小值和最大值。

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。为了简便,本实验仅利用C-SVM分类算法进行测试。本文CPPPCA算法利用PCA对训练数据进行处理,而文献[14]的ASN算法对每一个样本添加1个$N(0,1)$随机噪声。

3.2.3 结果与分析

由于本实验是一个多分类问题,因此选择一对余分类策略。选择RBF作为核函数,利用10折交叉验证确定最合适的参数σ和C。实验得到的平均结果如表4所示。

从表4可以看出,与文献[14]的ASN算法比较,从模式分类效果来看本文CPPPCA算法优势更明显。这主要是因为CPPPCA算法利用PCA保留了原始数据中的重要信息,而ASN算法在原始数据中加入了噪声,影响了其对数据的学习。

表4 KDD CUP 99数据集实验结果

Table 4 Experimental result on KDD CUP 99 data sets

算法	实验结果/%			
	Dos	Probing	U2R	R2L
CPPPCA算法	P=93,R=92	P=91,R=89	P=51,R=46	P=45,R=44
WT算法[16]	P=89,R=89	P=88,R=87	P=46,R=45	P=40,R=40
ASN算法[14]	P=90,R=88	P=89,R=86	P=49,R=44	P=39,R=38

与文献[16]的WT算法比较,发现CPPPCA算法仍然在分类效果上具有优势,这主要是因为WT算法基于小波变换,在稀疏数据上的分类效果不够理想,而CPPPCA算法借助于PCA对数据进行处理,能较好地处理稀疏数据。

另外,本文CPPPCA算法通过对原始属性进行重新组合,合理地避免了原始属性值的泄漏,而且通过新属性值难以反

推得到原始属性值,因此有效地实现了模式分类过程中对隐私的保护。而ASN算法添加的噪声可以很容易被过滤掉,因此其实现隐私保护的效果不够理想。

3.3　隐私保护效果的理论分析

实验Ⅰ和实验Ⅱ展示了本文CPPPCA算法具有较好的模式分类效果,这里从理论上进一步分析CPPPCA算法的隐私保护效果。根据文献[14]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$[x-0.67\sigma,x+0.67\sigma]$。

本文CPPPCA算法对于p维属性$x_1, x_2,\cdots, x_p$,取m个主成分$z_1, z_2,\cdots, z_m$,如下式所示

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases}$$

相应的系数矩阵为

$$\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mp} \end{bmatrix}$$

由矩阵基本理论可知,在主成分$z_1, z_2,\cdots, z_m$和系数矩阵已知的前提下,方程如果有解,则解一定不唯一。由PCA的实现过程可知,方程必有解,因此必然有无穷多满足方程的解存在,即根据$z_1, z_2,\cdots, z_m$的数值,无法反推$x_1, x_2,\cdots, x_p$的数值,也无法确定$x_1, x_2,\cdots, x_p$的有效范围,因此本文CPPPCA算法的隐私保护性能优于ASN算法。

4　结论

CPPPCA算法的主要思想是通过主成分分析将原始的特征属性进行线性组合,从而避免了原始数据的的泄露,实现了模式分类过程中的隐私保护。同时由于主成分分析一定程度上可以实现去噪的目的,避免分类算法陷入过拟合,有效地提高了分类器的分类性能。在Adult数据集和KDD CUP 99数据集上进行的仿真实验,验证了本文算法具有更好的隐私保护和分类性能。

参考文献(References)

[1] Han J, Kamber M, Pei J. Data mining: Concepts and techniques[M]. CA,San Mateo: Morgan kaufmann, 2006.

[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, et al. L-diversity: Privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007(1): 3.

[4] 田秀霞, 王晓玲, 高明, 等. 数据库服务——安全与隐私保护[J]. 软件学报, 2010, 21(5): 991-1006.
Tian Xiuxia, Wang Xiaoling, Gao Ming, et al. Database as a service? security and privacy preserving[J]. Journal of Software, 2010, 21(5): 991-1006.

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

[6] 戴群, 陈松灿, 王喆. 一个基于自组织特征映射网络的混合神经网络结构[J]. 软件学报, 2009, 20(5): 1329-1336.
Dai Qun, Chen Songcan, Wang Zhe. Hybrid neural network architecture based on self-organizing feature maps[J]. Journal of Software, 2009, 20(5): 1329-1336.

[7] 杨静, 辛宇, 谢志强. 面向物联网传感器事件监测的双向反馈系统[J]. 计算机学报, 2013, 36(3): 506-520.
Yang Jing, Xin Yu, Xie Zhiqiang. A bi-feedback system of wireless sensor network event detection in the internet of things[J]. Chinese Journal of Computers, 2013, 36(3): 506-520.

[8] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[9] 曾志强, 高济. 基于向量集约简的精简支持向量机[J]. 软件学报, 2007, 18(11): 2719-2727.
Zeng Zhiqiang, Gao Ji. Simplified support vector machine based on reduced vector set method[J]. Journal of Software, 2007, 18(11): 2719-2727.

[10] 顾彬, 郑关胜, 王建东. 增量和减量式标准支持向量机的分析[J]. 软件学报, 2013, 24(7): 1601-1613.
Gu Bin, Zheng Guansheng, Wang Jiandong. Analysis for incremental and decremental standard support vector machine[J]. Journal of Software, 2013, 24(7): 1601-1613.

[11] Quinlan J R. C4.5: Programs for machine learning[M]. San Mateo, CA: Morgan Kaufmann, 1993.

[12] Zhou Z H, Jiang Y. NeC4.5: Neural ensemble based C4.5[J]. Knowledge and Data Engineering, IEEE Transactions on, 2004, 16(6): 770-773.

[13] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees[M]. Florida: CRC Press, 1984.

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

[15] Kargupta H, Datta S, Wang Q, et al. On the privacy preserving properties of random data perturbation techniques[C]//Data Mining, 2003. Third IEEE International Conference on. New York: IEEE, 2003: 99-106.

[16] Bapna S, Gangopadhyay A. A wavelet-based approach to preserve privacy for classification mining[J]. Decision Sciences, 2006, 37(4): 623-642.

[17] 胡文军, 王士同. 隐私保护的SVM快速分类方法[J]. 电子学报, 2012, 40(2): 280-286.
Hu Wenjun, Wang Shitong. Fast classification approach of support vector machine with privacy preservation[J]. Acta Electronica Sinica, 2012, 40(2): 280-286.

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

[19] Duda R O, Hart P E, Stork D G. Pattern classification[M]. New York: John Wiley & Sons, 2012.

[20] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]// IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, 2: 1137-1145.

[21] 董春曦. 支持向量机及其在入侵检测中的应用研究[D]. 西安: 西安电子科技大学, 2004.
Dong Chunxi. Study of support vector machines and its application in intrusion detection systems[D]. Xi'an: Xidian University, 2004.

(责任编辑　韩星明)