分布式环境下的隐私保护特征选择研究

万文强,张伶卫

(南京邮电大学 计算机学院,江苏 南京 210003)

[摘要] 在 Map-Reduce 的分布式环境框架下,基于微分隐私与主成分分析,并与熵、误分类增益、基尼指数等统计量相结合,提出了一种新的在分布式环境下的隐私保护特征选择算法,实现了在保护数据集隐私的同时保护特征的隐私. 仿真实验结果表明,该算法具有较好的性能,能够在保护一定程度隐私信息的同时,有效地进行特征选择.

[关键词] 隐私保护,特征选择,分布式,微分隐私,主成分分析

[中图分类号] TP181 [文献标志码] A [文章编号] 1672-1292(2012)03-0060-08

Privacy Preserving Feature Selection in Distributed Environment

Wan Wenqiang,Zhang Lingwei

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Privacy preserving and feature selection are very important in data mining. Thus, how to select feature effectively based on privacy preserving is also a hot topic. Under the Map-Reduce distributed environment framework, proposed was the combination of the differential privacy and principal component analysis with the statistics including entropy, misclassification gain, and gini index, a new privacy preserving feature selection algorithm on distributed environment. The algorithm achieved the purposes of protecting privacy of both data sets and features. The simulation results on several bench-mark data sets indicated that this algorithm performed well. During the selection of the important features, it could protect privacy information to a certain extent.

Key words: privacy preserving, feature selection, distribution, differential privacy, principal component analysis

特征选择是指从一组特征中挑选出一些最有效的特征以降低特征空间维数的过程[1]. 文献[2]指出,特征选择在不显著降低分类精度和不显著改变类分布的条件下选择尽量小的特征子集. 特征选择算法一般可分为 3 类:过滤器模式、封装器模式和嵌入模式[3,4]. 然而,现有的特征选择算法大都没有考虑隐私保护问题,这使得特征选择日后将面临严峻的信息安全问题.

隐私泄露正日益成为一个新的信息安全问题,仅靠传统的访问控制并不能很好地达到隐私保护的目的. 随着数据挖掘的广泛应用和随之而来的隐私泄露问题,隐私保护已经逐渐成为一个重要且富有挑战性的课题. 只有数据中包含的隐私得到有效的保护,数据挖掘才能被有效地使用和推广. 文献[5]中首次提出了隐私保护的问题,强调了对个人信息的隐私保护. 自从隐私保护问题提出之后,产生了很多隐私保护算法:2002 年,Sweeney[6] 提出了 K-匿名(k-anonymity),Clifton[7] 等人提出了安全多方计算(SMC);2006 年,Dwork[8] 提出了微分隐私(differential privacy);2011 年,葛新景[9] 等人提出了基于博弈论的隐私保护分布式数据挖掘等. 然而,这些隐私保护方法大都只从单一方面考虑保护隐私. 本文将从保护数据集隐私和保护特征隐私两方面出发,设计新的隐私保护方法.

随着网络的迅速发展,各类应用产生的大量数据可能分布存储在多个站点上,这样自然而然就形成了分布式的环境. 本文在分布式的环境下,将新设计的隐私保护方法与特征选择相结合,提出了分布式环境下的隐私保护特征选择算法.

收稿日期:2012-07-10.
基金项目:国家自然科学基金(61073114).
通讯联系人:万文强,硕士,研究方向:数据挖掘与机器学习. E-mail:aiaiyouchou@163.com

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1 隐私保护的特征选择

1.1 基于统计理论的特征选择

近年来,越来越多的基于统计理论的方法被应用到特征选择当中. 其中比较著名的方法有熵、误分类增益、基尼指数、信息增益、互信息、χ^2 统计、相关系数等等[10]. 本文分别采用熵、误分类增益和基尼指数的方法进行特征选择.

若数据集 $\boldsymbol{D}$ 含有 $p+1$ 维向量,表示为 $\{\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_P,\boldsymbol{C}\}$,前 p 维表示样本的特征,第 $p+1$ 维表示样本的类标签. 假设样本的特征值是离散的,且只是二类问题,样本标记表示为 $\{0,1\}$(多类问题可以转换成多个两类问题). 对于 $\forall i=1,2,\cdots,p,\boldsymbol{B}_i\in\{0,1,\cdots,m_i-1\}$(其中 m_i-1 表示特征 B_i 所取到的最大值). 定义 x_{ib0} 表示为在数据集 $\boldsymbol{D}$ 中,满足 $\boldsymbol{B}_i=b$ 和 $\boldsymbol{C}=0$(此处 $b\in\{0,1,\cdots,m_i-1\}$) 的样本数目. 同理,x_{ib1} 表示满足 $\boldsymbol{B}_i=b$ 和 $\boldsymbol{C}=1$ 的样本数目,x_{ib} 表示满足 $\boldsymbol{B}_i=b$ 的样本数目. 文献[11] 已证明,基于熵、误分类增益、基尼指数的特征选择分别可以采用以下准则:

$$\boldsymbol{B}_{\text{best}}(\text{Entropy})=\underset{i\in[1\cdots p]}{\operatorname{argmax}}\left[\sum_{b=0}^{m_i-1}\left\{\left(\frac{x_{ib0}}{x_{ib}}\right)\log\left(\frac{x_{ib0}}{x_{ib}}\right)+\left(\frac{x_{ib1}}{x_{ib}}\right)\log\left(\frac{x_{ib1}}{x_{ib}}\right)\right\}\right], \tag{1}$$

$$\boldsymbol{B}_{\text{best}}(\text{MI})=\underset{i\in[1\cdots p]}{\operatorname{argmax}}\left[\sum_{b=0}^{m_i-1}|x_{ib0}-x_{ib1}|\right], \tag{2}$$

$$\boldsymbol{B}_{\text{best}}(\text{Gini})=\underset{i\in[1\cdots p]}{\operatorname{argmax}}\left[\sum_{b=0}^{m_i-1}\left\{\frac{(x_{ib0})^2+(x_{ib1})^2}{x_{ib}}\right\}\right]. \tag{3}$$

1.2 微分隐私

定义1 (ε-differential privacy) 算法 A 对于任意的数据集 $\boldsymbol{D}_1$、$\boldsymbol{D}_2$($\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 只相差一个元素),以及对于任意的输出子集 $\boldsymbol{S}\subseteq\text{Range}(A)$,若满足:

$$Pr[A(\boldsymbol{D}_1)\in\boldsymbol{S}]\leqslant\exp(\varepsilon)\times Pr[A(\boldsymbol{D}_2)\in\boldsymbol{S}], \tag{4}$$

则算法 A 被称之为满足微分隐私[12]. 其中,Range(A) 表示算法 A 输出结果的范围,Pr 表示在算法 A 上产生结果的概率.

定义2 (Sensitivity) 对于算法 $A:\boldsymbol{D}\rightarrow\mathbf{R}^d$,敏感度 ΔQ 为:

$$\Delta Q=\max_{D1,D2}\|A(\boldsymbol{D}_1)-A(\boldsymbol{D}_2)\|_1, \tag{5}$$

其中,$\boldsymbol{D}_1$ 与 $\boldsymbol{D}_2$ 最多只相差一个元素,且 $\boldsymbol{D}_1,\boldsymbol{D}_2\in\boldsymbol{D}$. 敏感度 ΔQ 的含义是指算法 A 在数据集 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 上操作所产生的最大差值.

定理1[12,13] 对算法 $A:\boldsymbol{D}\rightarrow\mathbf{R}^d$,加入满足 $\text{Lap}\left(\frac{\Delta Q}{\varepsilon}\right)$ 分布的拉普拉斯噪声后生成新的机制 $\tilde{A}$,则 $\tilde{A}$ 满足 ε-differential privacy. 具体表示为:

$$\tilde{A}(x)=A(x)+\langle\text{Lap}(\Delta Q/\varepsilon)\rangle^d. \tag{6}$$

1.3 主成分分析

主成分分析(Principal Component Analysis,PCA) 又称为 Karhunen-Loeve 变换或 Hotelling 变换,是数理统计分析、特征提取和降维的经典方法[14,15]. 它将原始特征通过线性变换映射到新的低维特征空间,而获得的主成分可以看作是所有原始特征的线性组合[16].

1.4 基于微分隐私和 *PCA* 的隐私保护特征选择

本文考虑从两个方面对特征选择进行隐私保护,分别为: 保护数据集隐私和保护特征的隐私.

保护数据集隐私,是指一批数据集 $\boldsymbol{D}_1$ 已进行特征选择产生结果 FS_1,在新的数据集 $\tilde{\boldsymbol{D}}$ 到来时,将 $\tilde{\boldsymbol{D}}$ 加入 $\boldsymbol{D}_1$ 中生成总的数据集 $\boldsymbol{D}_2$,对数据集 $\boldsymbol{D}_2$ 进行特征选择产生结果 FS_2,使得 FS_1 与 FS_2 并未发生重大的变化,甚至"近似相等",从而保护了新数据集 $\tilde{\boldsymbol{D}}$ 的隐私,这一部分是从微分隐私相关定义角度考虑的.

保护特征的隐私,是指计算出样本每个特征的信息量(amount of information, AOI),进行特征选择之后,保证所选特征子集的总信息量之和不超过某个阈值,限制其所产生的总信息量,从而减小隐私泄露的风险.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

从保护数据集隐私方面出发,本文考虑将微分隐私融入特征选择当中. 以基于基尼指数的特征选择为例,利用式(6),考虑将拉普拉斯(Lap)噪声加入到基尼指数这一统计量中. 观察式(3)不难看出,$\boldsymbol{B}_{\text{best}}$(Gini)由$x_{ib0}$和$x_{ib1}$两个变量决定. 于是,不妨将拉普拉斯噪声分别加入到x_{ib0}和x_{ib1}中,得到$\tilde{x}_{ib0} = x_{ib0} + \text{Lap}(\lambda_1)$和$\tilde{x}_{ib1} = x_{ib1} + \text{Lap}(\lambda_2)$,其中$\lambda_1 = \frac{\Delta Q_1}{\varepsilon}$,$\lambda_2 = \frac{\Delta Q_2}{\varepsilon}$. 又由于$x_{ib0}$和$x_{ib1}$分别表示在数据集$\boldsymbol{D}$中,满足$\{\boldsymbol{B}_i = b, \boldsymbol{C} = 0\}$和$\{\boldsymbol{B}_i = b, \boldsymbol{C} = 1\}$的样本数目,易得出$\Delta Q_1 = \Delta Q_2 = 1$,故$\lambda_1 = \lambda_2 = \frac{1}{\varepsilon}$. 再将修改后的$\tilde{x}_{ib0}$、$\tilde{x}_{ib1}$($\tilde{x}_{ib} = \tilde{x}_{ib0} + \tilde{x}_{ib1}$)代入式(3)中,即可得到基于基尼指数的微分隐私特征选择评价准则. 同理,亦可得出基于熵与基于误分类增益的微分隐私特征选择的评价准则.

本文考虑利用PCA对特征的隐私加以保护. 文献[17]指出,利用PCA衡量特征所具有的信息量AOI效果显著. 信息量可以很好地反映原始样本的信息,从而产生隐私泄露的风险. 因此,由人为设定一个阈值Λ,使得所选特征子集的信息量之和不超过Λ,从而达到隐私保护的效果,即:

$$\sum_{i \in \Gamma} \text{AOI}_i \leqslant \Lambda. \tag{7}$$

其中,Γ为所选的特征子集,AOI_i为第i个特征的信息量,Λ为给定的阈值. 对于AOI_i,先采用PCA方法对原始数据集进行主成分分析,利用所得的第一主成分F_1,取F_1在各原始特征上的投影值为该特征的信息量,从而得出AOI_i.

2 分布式环境下的隐私保护特征选择

随着网络的迅速发展,海量的数据已使得传统的单机系统在功能和性能上不能满足数据处理能力的需要,越来越多的分布式系统已成为当今的主流. 本文考虑将隐私保护的特征选择算法与分布式计算相结合,并利用Hadoop[18]平台下的Map-Reduce[19]编程框架实现该算法.

2.1 *Map-Reduce* 编程框架

Map-Reduce是一种分布式计算模型,也是Hadoop的核心. 它把运行在大规模集群上的并行计算过程抽象为两个函数:Map和Reduce,也即映射和化简. Map把任务分解成为多个任务,Reduce把分解后多任务处理的结果汇总起来,得到最终结果. Map-Reduce这种编程模式特别适合于非结构化和结构化的海量数据的搜索、挖掘、分析和机器智能学习等[20]. Map-Reduce处理大数据集的过程如图1所示.

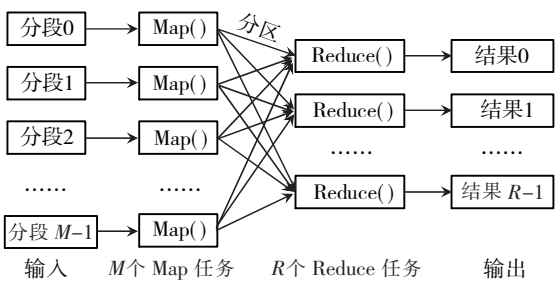

图1 Map-Reduce处理大数据集的过程

Fig.1 The process on large data sets with Map-Reduce

在Map阶段,Map-Reduce框架将用户输入的数据分割为M个片断,对应M个Map任务. Map的输出结果是一组〈key,value〉对,这是经过Map操作后所产生的中间结果. 在Reduce操作之前,系统已经将所有Map产生的中间结果进行归类处理,使得相同key对应的一系列value能够集结在一起提供给一个Reduce进行归并处理. Reduce阶段中的输入参数为(key,[value_1,…,value_m]),Reduce的工作就是对这些具有相同key的value值进行归并处理,最终形成(key,final_value)的结果.

2.2 基于隐私保护的分布式特征选择

本文将基于微分隐私和PCA的隐私保护特征选择方法与Map-Reduce编程模式相结合,实现基于隐私保护的分布式特征选择算法. 算法分为两步,第一步先由PCA得出每个特征的信息量AOI_i,第二步又分为两个阶段:Map阶段,输入参数为数据集的原始样本,输出的中间结果为〈特征维数;(特征值,类标签)〉对;Reduce阶段,以Map输出的中间结果作为输入参数,输出〈特征维数;特征统计量〉对. 然后将特征统计量进行排序,挑选出所需要的特征子集,并使所选特征子集的信息量之和不超过给定阈值Λ. 具体实现算法为:

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

算法 1　基于隐私保护的分布式特征选择算法

Step 1　由 PCA **方法得出各特征的信息量** AOI_i

Step 2

Map 阶段:

(**初始化** buffer = Null,〈**特征维数;(特征值,类标签)**〉= Null,p **为样本的特征个数**)

WHILE(**当前读取的样本行**)

　buffer ← **读取文件中的一行样本;**

　FOR($i = 1$; $i <= p$; $i++$)

　　IF(**类标签** = 0)

　　输出〈**特征维数;(特征值,**0)〉

　　ELSE

　　输出〈**特征维数;(特征值,**1)〉

　　END IF

　END FOR

　　指向下一行样本;

END WHILE

Reduce 阶段:

(**初始化** buffer = Null; $x_{ib0} = x_{ib1} = x_{ib} = 0$; **对于** B_i **特征,**$b \in \{0,1,\cdots,m_i - 1\}$; **各特征的信息量** AOI_i; **所选特征子集** Γ = Null)

用户输入: 隐私度 ε; **阈值** Λ.

FOR(**读取每一行样本**)

　b = **特征值;**

　计算出 x_{ib0}、x_{ib1};

END FOR

FOR($i = 1$; $i <= p$; $i++$)

　FOR($b = 0$; $b <= m_i - 1$; $b++$)

$$x_{1b0} = x_{1b0} + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right);$$

$$x_{1b1} = x_{1b1} + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right);$$

$$x_{1b} = x_{1b0} + x_{1b1};$$

　END FOR

　求出 B_i(Entropy)、B_i(MI)、B_i(Gini);

END FOR

Sort(B_i(Entropy)),Sort(B_i(MI)),Sort(B_i(Gini));

选出满足公式 7 **特征子集** Γ;

END

3　仿真实验

本文将在 4 个数据集上进行仿真实验,验证所提出的基于隐私保护的分布式特征选择算法. 前 3 个数据集均来自 UCI [21],最后一个采用合成数据集 $S1$,其中特征 2、1、3 设为重要特征,呈高斯分布,其余不重要特征取值随机. 所用数据集的数据特性如表 1 所示.

对这 4 个数据集分别进行基于熵、误分类增益、基尼指数的分布式隐私保护特征选择(先不考虑特征的隐私保护部分,即此处设 $\Lambda \to \infty$). 由于算法中加入了拉普拉斯随机噪声,因此分别进行了 5 次实验,结果取 5 次实验的平均值,再将所得的特征统计量进行排序,选出所需的特征子集(Monk 数据集选取了 3 个特征,Breast Cancer 数据集选取了 4 个特征,SPECT Heart 数据集选取了 5 个特征,$S1$ 数据集选取了 3 个特征). 此外,ε 由人为设定,本文分别对 ε = 1,0.1,0.01,0.001,0.000 1,0.000 01 和无噪声等 7 种情况分别

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

进行了实验. 具体的实验结果如表2 ~ 表5所示. 对于重要特征未知的数据集,本文采用无噪声情况下所选特征子集的分类准确率验证其特征选择的效果,结果见表6.

表1　实验数据集的数据特性

Table 1　Summary of the data sets

数据集	样本数	特征数	重要特征
Monk	122	6	2,5,4
Breast Cancer	683	9	未知
SPECT Heart	267	22	未知
*S*1	10 000	9	2,1,3

表2　$\Lambda \to \infty$ 时,Monk数据集特征选择结果

Table 2　The selected feature subsets for different values of ε when $\Lambda \to \infty$ on Monk data sets

ε	熵	误分类增益	基尼指数
无噪声	2,5,1	2,5,6	2,5,1
$\varepsilon = 1$	2,5,1	2,5,6	2,5,1
$\varepsilon = 0.1$	2,5,6	2,5,6	2,5,1
$\varepsilon = 0.01$	2,5,6	2,5,6	2,5,1
$\varepsilon = 0.001$	3,6,2	2,5,4	2,5,1
$\varepsilon = 0.000\,1$	6,3,1	5,2,1	5,2,4
$\varepsilon = 0.000\,01$	3,6,4	5,4,2	5,2,4

表3　$\Lambda \to \infty$ 时,Breast Cancer数据集特征选择结果

Table 3　The selected feature subsets for different values of ε when $\Lambda \to \infty$ on Breast Cancer data sets

ε	熵	误分类增益	基尼指数
无噪声	2,3,6,7	2,3,6,7	2,3,6,7
$\varepsilon = 1$	3,2,6,7	2,3,6,7	2,3,6,7
$\varepsilon = 0.1$	3,2,6,7	2,3,6,7	2,3,6,7
$\varepsilon = 0.01$	3,2,6,7	2,3,6,7	2,3,6,7
$\varepsilon = 0.001$	3,6,2,7	2,3,6,7	3,6,2,5
$\varepsilon = 0.000\,1$	2,6,5,4	6,5,2,9	2,6,3,8
$\varepsilon = 0.000\,01$	7,9,1,6	6,1,5,3	2,7,4,1

表4　$\Lambda \to \infty$ 时, SPECT Heart数据集特征选择结果

Table 4　The selected feature subsets for different values of ε when $\Lambda \to \infty$ on SPECT Heart data sets

ε	熵	误分类增益	基尼指数
无噪声	13,21,8,22,16	13,8,16,22,7	13,8,21,22,16
$\varepsilon = 1$	13,21,8,22,16	13,8,16,22,21	13,8,21,22,16
$\varepsilon = 0.1$	13,21,8,22,16	13,8,16,22,7	13,8,21,22,16
$\varepsilon = 0.01$	13,21,22,8,16	13,8,22,16,21	13,21,8,22,16
$\varepsilon = 0.001$	13,16,5,7,1	13,16,3,20,22	21,22,20,1,11
$\varepsilon = 0.000\,1$	21,7,6,1,22	8,5,1,18,19	7,10,18,15,6
$\varepsilon = 0.000\,01$	14,17,13,15,7	19,5,21,3,6	22,16,12,14,21

表5　$\Lambda \to \infty$ 时, *S*1数据集特征选择结果

Table 5　The selected feature subsets for different values of ε when $\Lambda \to \infty$ on *S*1 data sets

ε	熵	误分类增益	基尼指数
无噪声	1,3,2	2,1,3	1,3,2
$\varepsilon = 1$	2,1,3	2,1,3	1,3,2
$\varepsilon = 0.1$	2,1,3	2,1,3	1,3,2
$\varepsilon = 0.01$	2,1,3	2,1,3	1,3,2
$\varepsilon = 0.001$	2,1,3	2,1,3	1,3,2
$\varepsilon = 0.000\,1$	2,3,1	2,1,3	3,1,2
$\varepsilon = 0.000\,01$	8,7,2	8,9,3	2,1,4

表6　无噪声情况下,各数据集特征选择的分类准确率

Table 6　Accuracy rate in the case of feature selection without privacy preserving

	熵/%	误分类增益/%	基尼指数/%
Monk	80.56	80.56	80.56
Breast Cancer	95.61	95.61	95.61
SPECT Heart	80.60	80.60	80.60
*S*1	95.16	95.16	95.16

由表6可以看出,在无噪声的情况下,所选特征子集均能得到较高的分类准确率. 观察表2 ~ 表5容易得出,这些算法基本能够准确地排序重要特征. 当 $\varepsilon = 1$ 时,特征选择能够到达很高的准确率,但此时所保护的隐私程度有限. 随着 ε 的不断下降,所保护的隐私程度不断上升,特征选择的准确率随之不断下降. 当 $\varepsilon = 0.000\,01$ 时,所选的特征子集有些已不能反映出重要特征. 而对于熵、误分类增益和基尼指数之间的比较,本文采用"随着 ε 变化,所产生的特征子集与无噪声下的特征子集的相似度"作为衡量标准,结果如图2所示. 由图2知,三者之间基尼指数的相似度最高,效果最好; 误分类增益其次; 熵最差.

由于本文首次将微分隐私融入到特征选择之中,而且不同的算法对于"隐私"的定义与衡量标准不一,所以无法与现有的隐私保护特征选择算法进行直接对比实验. 表2 ~ 表5及图2验证了该算法特征选择的准确性. 对于隐私性,本文将设计一小型场景来说明其隐私保护的能力. 假设存在数据集test1,对test1进行特征选择处理后,引入新的数据集test2,生成总的数据集Test(Test = test1 + test2),此时的目的就是对Test进行特征选择的同时不泄露数据集test2的隐私信息. 这里test1和test2均采用人工合成数据集,其中test1的特征1、2、5、6设为重要特征,test2的

表7　验证隐私性的数据集

Table 7　The data sets for privacy validation

数据集	样本数	特征数	重要特征
test1	100	9	1,2,5,6
test2	100	9	1,2,3,4
Test(test1 + test2)	200	9	未知

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

特征 1、2、3、4 设为重要特征,具体的数据特性如表 7 所示. 对 test1 和 Test 的特征选择结果如表 8 所示.

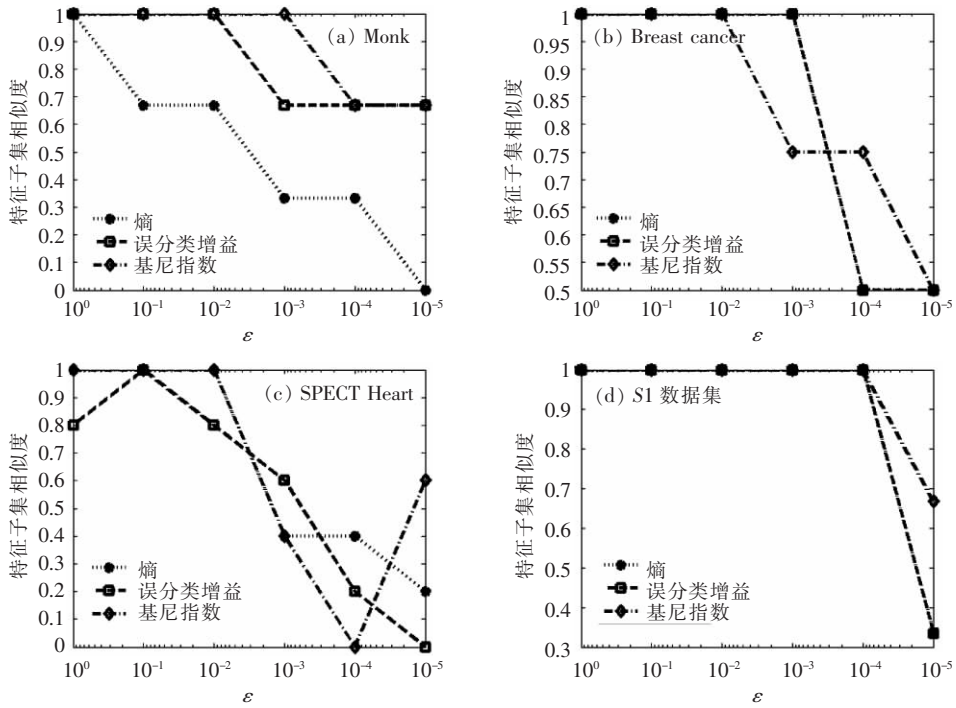

图 2 随着 ε 变化,算法在不同数据集上与原始特征选择子集的相似度

Fig.2 Similarity between subsets returned by traditional feature selection and our algorithms for different values of ε on

表 8 对 test1 和 Test 的特征选择结果

Table 8 The selected feature subsets on test1 and Test with different methods

	熵	误分类增益	基尼指数
对 test1 进行特征选择	2,1,6,5,8,3,9,4,7	1,2,6,5,8,3,9,4,7	1,2,6,5,7,9,3,4,8
无噪声情况下,对 Test 特征选择	2,1,4,3,6,5,8,9,7	1,2,3,4,6,5,7,8,9	1,2,3,4,6,7,9,5,8
ε = 1 时,对 Test 特征选择	2,1,4,3,6,5,8,9,7	1,3,4,6,5,7,8,2,9	1,3,4,6,7,9,5,2,8
ε = 0.1 时,对 Test 特征选择	2,1,4,3,6,5,8,9,7	1,3,4,6,5,7,8,2,9	1,3,4,6,7,9,5,2,8
ε = 0.01 时,对 Test 特征选择	2,1,4,3,6,5,8,7,9	1,3,4,6,5,7,8,2,9	1,3,4,6,7,5,9,2,8
ε = 0.001 时,对 Test 特征选择	2,1,5,6,4,3,8,7,9	1,6,5,3,2,4,9,7,8	1,6,5,3,4,7,9,2,8
ε = 0.000 1 时,对 Test 特征选择	2,8,7,9,3,4,5,6,1	5,2,3,4,6,8,1,7,9	1,2,7,4,5,3,6,9,8
ε = 0.000 01 时,对 Test 特征选择	8,6,1,2,3,4,9,7,5	5,4,8,1,9,6,3,7,2	5,1,3,9,8,6,4,2,7

观察表 8 可知,对 test1 的特征选择结果说明 test1 中 1、2、5、6 为重要特征. 但在无噪声的情况下(即无隐私保护的情况下),对 Test 的特征选择结果显示特征 3、4 却排在特征 5、6 之前,从而可能暴露出“test2 数据集特征 3、4 非常重要”这一隐私信息. 表 8 中加入微分隐私后,随着 ε 不断降低,隐私保护程度不断上升,特征 3、4 的排序逐渐靠后,从而达到不泄露“test2 数据集特征 3、4 非常重要”这一隐私信息,从而达到隐私保护的目的,验证了该算法的隐私保护性能.

综合表 2 ~ 表 5 中的实验结果,ε = 0.001 时效果是最好的,在能够容忍的准确率下降范围内,极大提高了算法的隐私度,很好地保护了数据集的隐私信息. 在保证数据集隐私的情况下,再考虑保护特征的隐私. 设定 ε = 0.001,通过调节不同的阈值 Λ,得出不同的特征子集,再利用所选的特征子集分别进行分类预测,得到的结果如图 3 所示.

由图 3 可以看出,当 ε 取相同的值时(这里 ε = 0.001),熵、误分类增益、基尼指数 3 个算法都能达到很好的分类准确率,从而实现了在保护数据集隐私的同时,保护特征的隐私. 观察可知,随着给定阈值的不断增大,分类准确率逐渐增加,但此时所选特征子集的信息量之和便不断增大,所保护的隐私程度也就随之下降.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

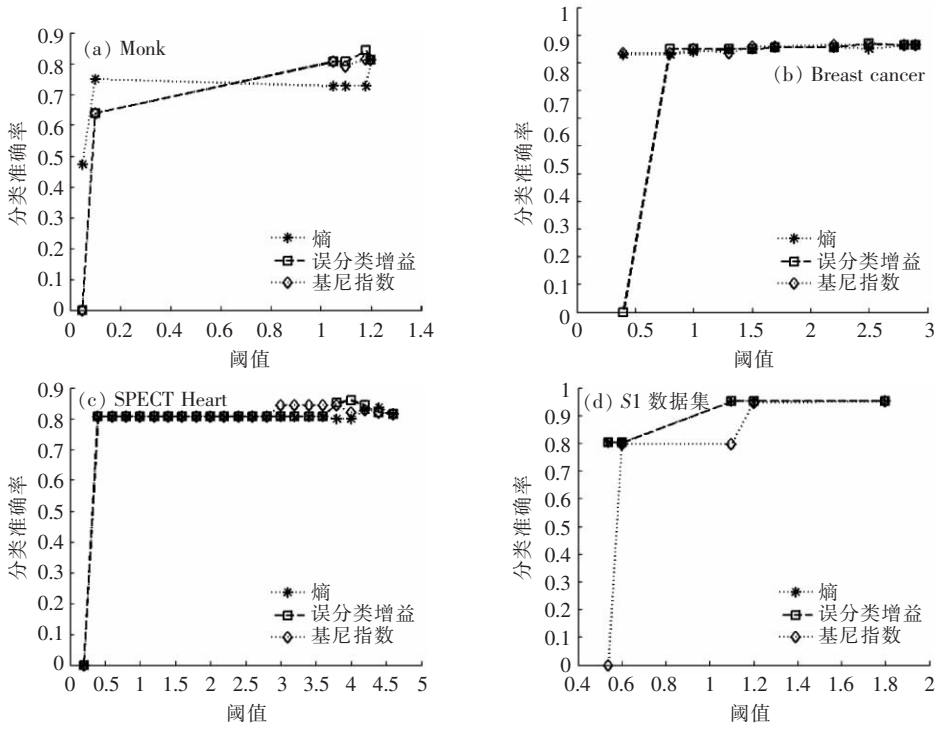

图 3　当 ε=0.001 时,算法在不同数据集上随着阈值变化的分类准确率

Fig.3　The classification accuracy varying with different value of threshold when ε=0.001

4　结语

特征选择是模式识别、机器学习和数据挖掘等领域的一个关键问题,而如何在特征选择的同时考虑隐私保护问题还鲜有相关文献. 本文在分布式的环境下,基于微分隐私和 PCA,并结合统计理论的特征选择,提出了分布式环境下的隐私保护特征选择算法,实现了同时保护数据集的隐私和特征的隐私. 实验结果表明,该算法具有较好的性能,能在保护一定隐私的同时,获得较好的准确率. 保护数据集隐私方面,对于输入参数隐私度 ε,随着 ε 的不断降低,隐私保护的程度便不断上升,所获得的准确率随之不断下降. 保护特征隐私方面,对于输入参数阈值 Λ,随着 Λ 不断增大,所保护的隐私程度不断下降,分类准确率不断上升. 在实际应用中,为了权衡隐私保护程度和准确率这两个关键点,用户需要调节选定出最优的 ε 和 Λ.

[参考文献](References)

[1] 边肇祺,张学工. 模式识别[M]. 2 版. 北京: 清华大学出版社,2000.
Bian Zhaoqi, Zhang Xuegong. Pattern Recognition [M]. 2nd ed. Beijing: Tsinghua University Press, 2000. (in Chinese)

[2] Dash M, Liu H. Feature selection for classification [J]. Intelligent Data Analysis, 1997, 1(3): 131-156.

[3] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 1-12.

[4] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003(3): 1 157-1 182.

[5] O'Leary D E. Knowledge Discovery as a Threat to Database Security Knowledge Discovery in Database [M]. Menloprk, CA: AAAI/MIF Press, 1991: 507-516.

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

[7] Clifton C, Kantarcioglu M, Vaidya J, et al. Tools for privacy preserving distributed data mining [J]. ACM SIGKDD Explorations Newsl, 2002, 4(2): 28-34.

[8] Dwork C. Differential privacy [C] // Proc of the 33rd ICALP. Venice, 2006.

[9] 葛新景,朱建明. 基于博弈论的隐私保护分布式数据挖掘 [J]. 计算机科学,2011,38(11): 161-166.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Ge Xinjing, Zhu Jianming. Privacy preserving distributed data mining based on game theory [J]. Computer Science, 2011, 38(11): 161-166. (in Chinese)

[10] Das K. Privacy preserving distributed data mining based on multi-objective optimization and algorithmic game theory [D]. Baltimore: University of Maryland Baltimore County, 2009.

[11] Das K, Bhaduri K, Kargupta H. A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to-peer networks [J]. Knowledge Information System, 2010, 24(3): 341-367.

[12] Dwork C. Differential privacy: a survey of results [C] // The 5th Annual Conference on Theory and Applications of Models of Computation. Xi'an, 2008.

[13] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] // Proceedings of the 3rd Theory of Cryptography Conference. New York, 2006: 265-284.

[14] Ding C, He Xiaofeng. Principal component analysis and effective K-means clustering [C] // Proceedings of the 4th SIAM International Conference on Data Mining. Orlando, 2004.

[15] 何晓群. 多元统计分析 [M]. 北京: 中国人民大学出版社, 2004.
He Xiaoqun. Multivariate Statistical Analysis [M]. Beijing: China Renmin University Press, 2004. (in Chinese)

[16] Mao K Z. Identifying critical variables of principal components for unsupervised feature selection [J]. IEEE Trans Systems, Man, and Cybernetics-part B: Cybernetics, 2005, 35(2): 339-344.

[17] Avidan S, Butman M. Efficient methods for privacy preserving face detection [C] // NIPS 2006. Vancouver, 2006: 57-64.

[18] Jia B, Wlodarczyk T, Rong C. Performance considerations of data acquisition in hadoop system [C] // The 2nd IEEE International Conference on Cloud Computing Technology and Science. Indianapolis, 2010: 545-549.

[19] Gunarathne T, Wu T L, Qiu J, et al. MapReduce in the clouds for science [C] // The 2nd IEEE International Conference on Cloud Computing Technology and Science. Indianapolis, 2010: 565-572.

[20] 刘鹏. 云计算 [M]. 2 版,北京: 电子工业出版社,2011.
Liu Peng. Cloud Computing [M]. 2nd ed. Beijing: Electronic Industry Press, 2011. (in Chinese)

[21] Newman D J, Hettich S, Black C L, et al. UCI Machine Learning Repository [EB/OL]. [2012-07-10]. http://archive.ics.uci.edu/ml/datasets.html.

[责任编辑: 严海琳]

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net