基于局部学习的差分隐私集成特征选择算法

刘中锋
(南京邮电大学 计算机学院、软件学院、网络空间安全学院,江苏 南京 210000)

摘 要:面对海量数据,特征选择在数据挖掘和机器学习领域上通常是不可或缺的一步。目前,机器学习安全领域受到了越来越多的关注,尤其是隐私保护方面。然而,对于隐私保护的特征选择仍然是一个比较新的课题,特别是与集成学习相关的集成特征选择。差分隐私是一种有着严格理论基础的隐私保护方法,因此提出了一种基于局部学习的差分隐私集成特征选择算法。该算法的主要思想是基于一种输出干扰策略,即向输出结果添加噪声从而保护隐私,而且该噪声依赖于原始算法的隐私度和敏感度。除了严格的理论证明之外,也从实验中展现了算法的性能。实验采用 KNN 和 SVM 作为分类器,分别分析了隐私度和特征数量的影响。结果显示随着隐私度的降低,提高了隐私保护程度。

关键词:特征选择;集成;差分隐私;隐私度;敏感度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2018)10-0079-04
doi:10.3969/j.issn.1673-629X.2018.10.016

An Ensemble Feature Selection Algorithm with Differential Privacy Based on Local Learning

LIU Zhong-feng
(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210000, China)

Abstract: When confronting massive data, feature selection is usually a necessary step for data mining and machine learning. Currently, secure machine learning, especially in privacy preservation, has attracted much attention. However, feature selection with privacy preservation is still a new issue, especially for feature selection related to ensemble learning. In this paper, we present a differentially private ensemble feature selection algorithm, of which the basic idea is the output perturbation where the density of perturbation noise depends on the privacy degree and sensitivity of original feature selection algorithm. Besides the theoretical proof, the experimental results also demonstrated their high performance under certain privacy preservation degree. In the experiment, KNN and SVM are selected as classifiers and the privacy degree and the number of features are researched. The results show that the privacy preserving degree is better, along with the decline of privacy degree.

Key words: feature selection; ensemble; differential privacy; privacy degree; sensitivity

0 引 言

特征选择是机器学习和数据挖掘等领域的一个关键问题,是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。特征选择不仅能够降低特征维数,也能够加快机器学习或者特征选择算法的执行速度,同时提高算法的准确率,使算法更具有可理解性。在之前的工作中,对于特征选择的研究主要集中在特征选择算法的稳定性[1]以及分类准确率等方面[2-3],然而隐私保护也是一个非常重要的研究方向,比如医院电子病历记录病人基本信息、疾病信息以及药品购买记录等,这些信息的泄露会对人身安全造成威胁[4]。虽然关于隐私保护的分类和回归等应用[5]都已着重研究过,但是对于隐私保护的特征选择算法的研究却很少[6-7]。已研究过的隐私保护仅仅是单特征选择算法,未涉及多个算法的领域。

与集成学习类似,集成特征选择算法也分为两个步骤:第一步是构造基特征选择器[8],第二步是通过某种组合集成每个基特征选择器的输出结果。文中采取 Bagging 集成策略,利用 bootstrap 抽样方法对原始数据集进行抽样,在抽样后的数据集上基于局部学习来训练基特征选择器[9],并且采取线性组合的方式对结果进行集成。为了使集成特征选择具有隐私保护的效

收稿日期:2017-11-30 **修回日期**:2018-04-03 **网络出版时间**:2018-05-28
基金项目:国家自然科学基金(61603197,91646116);江苏省自然科学基金(BK20140885)
作者简介:刘中锋(1993-),男,硕士研究生,研究方向为机器学习、特征选择、差分隐私。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20180525.1610.066.html

果,利用输出干扰策略,提出了先对结果集成后添加噪声的集成特征选择算法 FELP。证明该算法满足差分隐私模型[10-11]的定义,并通过实验证明其效用性。

1 基于局部学习的差分隐私集成特征选择算法

1.1 基于局部学习的特征选择

基于局部学习的特征权重算法的主要思想是将任意一个复杂的非线性问题转化为一组局部线性问题。在存在大量不相关特征的数据集上,采用基于局部学习的特征权重算法可以获得比较理想的特征选择结果。对于给定的包含 n 个样本的样本集 D,其中 $D=\{X,Y\}=\{\boldsymbol{x}_i,y_i\}_{i=1}^{n}$,$\boldsymbol{x}_i$ 是样本集中的第 i 个样本,y_i 是第 i 个样本对应的标签,并且每个样本都是一个 d 维的向量,即 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{id})$。由于逻辑回归损失具有简单效率高的特点,且是二阶可微和强凸的,这种强凸性对于快速求解最优值是非常有效的[6]。对于样本 $\boldsymbol{x}_i$,逻辑损失函数的定义如下:

$$l(\boldsymbol{w}^{\mathrm{T}}z_i)=\log(1+\exp(-\boldsymbol{w}^{\mathrm{T}}z_i)) \tag{1}$$

其中,$\boldsymbol{w}$ 为特征权重向量;$z_i=|\boldsymbol{x}_i-NM(\boldsymbol{x}_i)|-|\boldsymbol{x}_i-NH(\boldsymbol{x}_i)|$。$z_i$ 可以看作是 $\boldsymbol{x}_i$ 的变换点,$\boldsymbol{w}^{\mathrm{T}}z_i$ 可以看作是局部间隔,属于假设间隔[12]。此外,为了防止过拟合,在公式中加入了正则化项。由于 L_2 正则化项具有旋转不变性[13],同时也具有很强的稳定性,所以评价准则定义为:

$$L(\boldsymbol{w},D)=\frac{1}{n}\sum_{i=1}^{n}l(\boldsymbol{w}^{\mathrm{T}}z_i)+\lambda\|\boldsymbol{w}\|^2 \tag{2}$$

其中,λ 为正则化参数。

基于局部学习的特征权重算法 FWELL 的内容见文献[14]。

1.2 差分隐私模型和敏感度

文中采用差分隐私模型[10-11]作为隐私风险的一个度量。差分隐私算法定义如下:

定义 1(ε - 差分隐私):对于任意给定的数据集 D 和 D^i(其中 D 和 D^i 最多只有一个元素不同),以及任意的输出子集 $S\subseteq \mathrm{Range}(F)$,如果有:

$$P[F(D)\in S]\leqslant e^{\varepsilon}\times P[F(D^i)\in S] \tag{3}$$

则算法 F 满足差分隐私。

因为加入噪声的多少会影响算法的性能,所以基于差分隐私算法的输出干扰策略和算法的敏感度相关。文献[4,15]中对敏感度进行了定义。

定义 2:对于任意带有 n 个输入值的算法 F,全局敏感度 ΔQ 定义为对所有的输入值,当算法 F 的某个输入值变化时函数值的最大变化的 L_2 范数,即:

$$\Delta Q=\max_{D,D^i}\|F(D)-F(D^i)\| \tag{4}$$

式 4 的敏感度定义和文献[16]中算法稳定性的定义类似,该稳定性定义为:

$$\Delta\mathrm{St}=\max_{D,D^{\prime i}}\|F(D)-F(D^{\prime i})\| \tag{5}$$

式 4 和式 5 的区别在于,敏感度的定义旨在改变一个样本,而稳定性的定义旨在移除一个样本。根据三角不等式,能得到两者之间的关系,结论如下:

$$\begin{aligned}\Delta Q=\|F(D)-F(D^{\prime i})-(F(D^i)-F(D^{\prime i}))\|\leqslant\\ \|F(D)-F(D^{\prime i})\|+\|F(D^i)-\\ F(D^{\prime i})\|=2\Delta\mathrm{St}\end{aligned} \tag{6}$$

1.3 先集成后扰动策略的差分隐私特征选择算法(FELP)

文中采用 bootstrap 抽样策略对数据集 D 进行抽样,抽取 k 次得到 k 个不同的样本子集。因为 bootstrap 是统计学习中一种重采样技术,且每个样本子集的大小是$\lceil\beta n\rceil$,趋近于 βn,其中 β 为抽样比例。假定 $r_t=\{r_t(1),r_t(2),\cdots,r_t(\lceil\beta n\rceil)\}(t=1,2,\cdots,k)$ 是随机序列,同时指定 k 个样本子集 $D(r_t)=\{\boldsymbol{x}_{n(j)},\boldsymbol{y}_{n(j)}\}_{j=1}^{\lceil\beta n\rceil}$ 都是独立同分布的。此外,在每个样本子集上,基分类器 FWELL 在第 t 个样本子集 $D(r_t)$ 上的输出结果是 $\boldsymbol{w}_{D(r_t)}$,从而 k 个基分类器的输出结果集为 $\{\boldsymbol{w}_{D(r_1)},\boldsymbol{w}_{D(r_2)},\cdots,\boldsymbol{w}_{D(r_k)}\}$。接着把 k 个基分类器的输出结果通过线性组合取平均的方式作为最终的输出结果 $\boldsymbol{w}_D^{\prime}$。

以上介绍的是基于局部学习的传统的集成特征选择(FWELL - EN)过程,为了使该算法具有隐私保护的功能,对最终集成后的结果 $\boldsymbol{w}_D^{\prime}$ 添加一个噪声向量。该噪声向量与 FWELL - EN 算法的敏感度有关,见引理 1。

引理 1:FWELL - EN 算法的敏感度是 $\frac{2}{n\lambda}$。

证明:在数据集 D 和 $D^{\prime i}$ 上,FWELL - EN 算法的输出结果分别是 $\boldsymbol{w}_D^{\prime}$ 和 $\boldsymbol{w}_{D^{\prime i}}^{\prime}$,为了证明算法的敏感度,根据定义 2 和式 6,可以得到:

$$\begin{aligned}\Delta Q=\|\boldsymbol{w}_D^{\prime}-\boldsymbol{w}_{D^i}^{\prime}\|\leqslant 2\|\boldsymbol{w}_D^{\prime}-\boldsymbol{w}_{D^{\prime i}}^{\prime}\|=\\ 2\|\frac{1}{k}\sum_{t=1}^{k}\boldsymbol{w}_{D(r_t)}-\frac{1}{k}\sum_{t=1}^{k}\boldsymbol{w}_{D^{\prime i}(r_t)}\|\end{aligned} \tag{7}$$

因为 $r_1,r_2,\cdots,r_k$ 是独立同分布的,所以 $r_1,r_2,\cdots,r_k$ 与 r 有相同的分布,根据三角不等式,则

$$\begin{aligned}\Delta Q\leqslant\frac{2}{k}\sum_{t=1}^{k}\|\boldsymbol{w}_{D(r_t)}-\boldsymbol{w}_{D^{\prime i}(r_t)}\|=\\ 2\|\boldsymbol{w}_{D(r)}-\boldsymbol{w}_{D^{\prime i}(r)}\|\end{aligned} \tag{8}$$

故在数据集 $D(r)$ 上,根据式 5 得到 FWELL-EN 的稳定性是 $\|\boldsymbol{w}_{D(r)}-\boldsymbol{w}_{D^{\prime i}(r)}\|$。因此可以得到:

$$\begin{aligned}\Delta Q\leqslant 2\|\boldsymbol{w}_{D(r)}-\boldsymbol{w}_{D^{\prime i}(r)}\|(I(i\in r)+I(i\notin r))=\\ 2[\|\boldsymbol{w}_{D(r)}-\boldsymbol{w}_{D^{\prime i}(r)}\|I(i\in r)+\|\boldsymbol{w}_{D(r)}-\\ \boldsymbol{w}_{D^{\prime i}(r)}\|I(i\notin r)]\end{aligned} \tag{9}$$

如果该索引 r 与 i 无关,也就意味着样本 x_i 不在样

本子集 $D(r)$ 中,即满足 $D(r)=D^{/i}(r)$,于是有 $\boldsymbol{w}_{D(r)}=\boldsymbol{w}_{D^{/i}(r)}$ 和 $\|\boldsymbol{w}_{D(r)}-\boldsymbol{w}_{D^{/i}(r)}\|I(i\notin r)=0$。因此可得:

$$\Delta Q\leqslant 2\|\boldsymbol{w}_{D(r)}-\boldsymbol{w}_{D^{/i}(r)}\|I(i\in r) \tag{10}$$

由于数据子集 $D(r)$ 的大小是 $\lceil\beta n\rceil$,于是有 $I(i\in r)=\lceil\beta n\rceil/n\approx\beta$,进一步有 $\|\boldsymbol{w}_{D(r)}-\boldsymbol{w}_{D^{/i}(r)}\|\leqslant\frac{1}{\lceil\beta n\rceil\lambda}$。所以可以得到:

$$\Delta Q\leqslant\frac{2}{\lceil\beta n\rceil\lambda}\beta\approx\frac{2}{\lambda n} \tag{11}$$

根据文献[11]中的噪声定义可知,敏感度为 $2/n\lambda$ 的 FWELL-EN 算法的噪声向量 $\boldsymbol{b}_D$ 定义如下:

$$v(\boldsymbol{b}_D)=\frac{1}{a}\mathrm{e}^{-\frac{n\lambda\varepsilon}{2}\|b\|} \tag{12}$$

其中, a 为一个常量。

FELP 算法的伪代码如下所述。

输入:训练集 $D=\{\boldsymbol{x}_i,y_i\}_{i=1}^n,\boldsymbol{x}_i\in R^d$,式 2 中的正则化参数 λ ,隐私度 ε ,参数 a

输出:特征权重向量 $\boldsymbol{w}_D^*$

第一步:采取 bootstrap 抽样策略重复抽取 k 次(抽样参数为 β),得到 k 个不同的样本子集,并且每个样本子集的大小是 $\lceil\beta n\rceil$ 。

第二步:在每个样本子集上,根据算法 FWELL 得到 k 个输出结果 $\boldsymbol{w}_D$ 。

第三步:将 k 个输出结果 $\boldsymbol{w}_D$ 利用线性求和取平均的方法得到结果 $\boldsymbol{w}_D^{'}$ 。

第四步:根据 FWELL-EN 算法的敏感度,计算 $\boldsymbol{w}_D^*=\boldsymbol{w}_D^{'}+\boldsymbol{b}_D$,并且输出最终结果 $\boldsymbol{w}_D^*$ 。

因为 FELP 算法是基于差分隐私的算法,所以该算法一定要满足差分隐私的定义,见定理 1。

定理 1:FELP 算法满足差分隐私。

证明:因为数据集 D 和 D^i 只有一个样本不同,并且 $\boldsymbol{w}_D^*$ 和 $\boldsymbol{w}_{D^i}^*$ 分别是算法 FELP 在数据集 D 和 D^i 上的输出结果, $\boldsymbol{w}_D^{'}$ 和 $\boldsymbol{w}_{D^i}^{'}$ 分别是算法 FWELL-EN 在数据集 D 和 D^i 上的输出结果,而且 $\boldsymbol{b}_D$ 和 $\boldsymbol{b}_{D^i}$ 分别是各自对应的噪声向量,其中 $\boldsymbol{w}_D^*$ 和 $\boldsymbol{w}_{D^i}^*$ 分别满足: $\boldsymbol{w}_D^*=\boldsymbol{w}_D^{'}+\boldsymbol{b}_D$, $\boldsymbol{w}_{D^i}^*=\boldsymbol{w}_{D^i}^{'}+\boldsymbol{b}_{D^i}$ 。因此由文献[4]可得:

$$\frac{P(\boldsymbol{w}_D^*)}{P(\boldsymbol{w}_{D^i}^*)}=\frac{v(\boldsymbol{b}_D)}{v(\boldsymbol{b}_{D^i})}=\mathrm{e}^{\frac{n\varepsilon\lambda}{2}(\|\boldsymbol{b}_{D^i}\|-\|\boldsymbol{b}_D\|)} \tag{13}$$

其中, $P(\boldsymbol{w}_D^*)(P(\boldsymbol{w}_{D^i}^*))$ 是算法 FELP 在数据集为 D (D^i)时输出结果为 $\boldsymbol{w}_D^*(\boldsymbol{w}_{D^i}^*)$ 时的概率。

因为满足在数据集的某个敏感样本改变时最终的输出结果却没变,因此有 $\boldsymbol{w}_D^{'}+\boldsymbol{b}_D=\boldsymbol{w}_{D^i}^{'}+\boldsymbol{b}_{D^i}$,即 $\boldsymbol{w}_D^{'}-\boldsymbol{w}_{D^i}^{'}=\boldsymbol{b}_{D^i}-\boldsymbol{b}_D$,于是根据三角不等式,可得:

$$\|\boldsymbol{b}_{D^i}\|-\|\boldsymbol{b}_D\|\leqslant\|\boldsymbol{b}_{D^i}-\boldsymbol{b}_D\|=\|\boldsymbol{w}_D^{'}-\boldsymbol{w}_{D^i}^{'}\| \tag{14}$$

根据式 11、13、14,可以得到:

$$\frac{P(\boldsymbol{w}_D^*)}{P(\boldsymbol{w}_{D^i}^*)}=\mathrm{e}^{\frac{n\varepsilon\lambda}{2}(\|\boldsymbol{b}_{D^i}\|-\|\boldsymbol{b}_D\|)}\leqslant\mathrm{e}^{\frac{n\varepsilon\lambda\,2}{2\,n\lambda}}=\mathrm{e}^{\varepsilon} \tag{15}$$

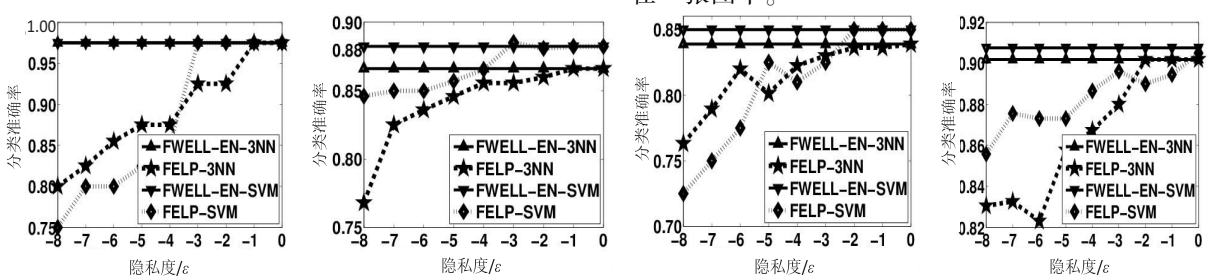
由上可知,算法 FELP 满足差分隐私。

2　实　验

文中采用 FWELL-EN 算法和 FELP 算法进行实验对比。整个实验包括两部分:验证隐私度参数 ε 的影响以及在某个特定隐私度的情况下,验证不同特征数量时的分类性能。在该实验中,选取支持向量机(SVM)和 k 近邻(kNN)作为分类器, SVM 中参数 $C=1$, k 近邻分类器中的参数 $K=3$。采用十次交叉验证,将数据集分为 10 等份,9 份作为训练数据,1 份作为测试数据。使用 bootstrap 抽样策略将训练数据集分为 20 个样本子集($k=20$),并且每份抽样比例是 $\beta=0.9$,所有实验中使用的参数 λ 将根据交叉验证调节。选取四个不同大小、不同维度的数据集作为实验数据,包括 Arcene、Soybean、Wdbc 和 Breast,其中 Arcene 是一个典型的高维度的小样本数据集。

2.1　隐私度实验

该实验中选定的特征维数是原始数据集中特征维数的 10% 。FELP 算法中的隐私度由 ε 衡量, ε 值的增加意味着隐私度的降低,保护效果也越差。实验结果见图 1。为了节省空间,两个分类器的结果共同显示在一张图中。



(a)Soybean　(b)Breast　(c)Arcene　(d)Wdbc

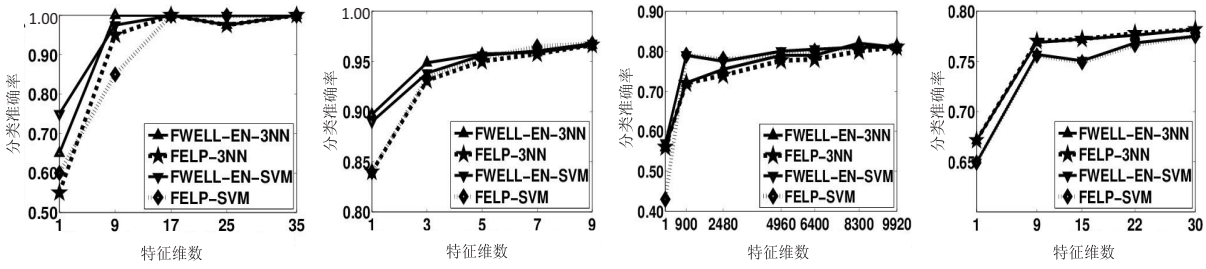
图 1　隐私度实验结果 3NN-SVM

从实验结果可以看出,在没有隐私保护的情况下(即 $\varepsilon=10$)FWELL-EN 算法的分类准确率和具有隐私保护效果的 FELP 算法有相同的值。但是随着隐私度 ε 的减小,算法 FELP 的分类准确率也随之减小,

而隐私保护性能逐步提高。并且从整体上看，SVM 分类器的准确率比 3NN 要高。虽然当隐私度越小时，隐私保护效果越好，但同样也面临可用性的降低，所以考虑到隐私保护和可用性的平衡，$\varepsilon = 0.01$ 是一个效果不错的选择。

2.2 特征维数实验

该实验主要研究的是特征维数和分类准确率的情况，此时选定的隐私度 $\varepsilon = 0.01$。特征维数是根据数据集的特征维数来选取的。分类结果见图 2。



(a)Soybean (b)Breast (c)Arcene (d)Wdbc

图 2 特征维数实验结果 3NN-SVM

从实验结果可以看出，在特定的隐私度 $\varepsilon = 0.01$ 时，算法 FELP 的分类准确率接近算法 FWELL - EN，说明算法 FELP 的分类性能和算法 FWELL - EN 非常接近，证明了算法 FELP 的有效性。

3 结束语

在安全类机器学习中，具有隐私保护性能的特征选择是一个热门话题。文中提出了一种基于局部学习的带有输出干扰策略的差分隐私集成特征选择算法 FELP，并且从理论上证明了该算法满足差分隐私，同时通过实验也证明在特定隐私度下，该算法是有效实用的。

参考文献：

[1] 李 云. 稳定的特征选择研究[J]. 微型机与应用，2012，31(15)：1-2.

[2] WOZNICA A，NGUYEN P，KALOUSIS A. Model mining for robust feature selection[C]//Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. Beijing，China：ACM，2012：913-921.

[3] LI Yun，SI J，ZHOU Guojing，et al. FREL：a stable feature selection algorithm[J]. IEEE Transactions on Neural Networks and Learning Systems，2015，26(7)：1388-1402.

[4] 蔡红云，田俊峰. 云计算中的数据隐私保护研究[J]. 山东大学学报：理学版，2014，49(9)：83-89.

[5] CHAUDHURI K，MONTELEONI C，SARWATE A D. Differentially private empirical risk minimization[J]. Journal of Machine Learning Research，2011，12：1069-1109.

[6] TAN Mingkui，TSANG I W，WANG Li. Minimax sparse logistic regression for very high-dimensional feature selection[J]. IEEE Transactions on Neural Networks & Learning Systems，2013，24(10)：1609-1622.

[7] YANG J，LI Y. Differential privacy feature selection[C]// Proceedings of international joint conference on neural networks. Beijing，China：ACM，2014：4182-4189.

[8] SAEYS Y，ABEEL T，PEER Y. Robust feature selection using ensemble feature selection techniques[C]//Proceedings of the European conference on machine learning and knowledge discovery in databases. Antwerp，Belgium：Springer，2008：313-325.

[9] SUN Yijun，TODOROVIC S，GOODISON S. Local learning based feature selection for high dimensional data analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2010，32(9)：1610-1626.

[10] DWORK C. Differential privacy[C]//Proceedings of the 33rd international conference on automata，languages and programming. Venice，Italy：Springer-Verlag，2006：1-12.

[11] DWORK C，ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science，2014，9(3-4)：211-407.

[12] CRAMMER K，BACHRACH R G，NAVOT A，et al. Margin analysis of the LVQ algorithm[C]//Proceedings of advances in neural information processing systems. [s. l.]：[s. n.]，2002：462-469.

[13] NG A Y. Feature selection，l1 vs. l2 regularization，and rotational invariance[C]//Proceedings of international conference on machine learning. Banff，Alberta，Canada：ACM，2004.

[14] LI Yun，YANG Jun，JI Wei. Local learning-based feature weighting with privacy preservation[J]. Neurocomputing，2016，174：1107-1115.

[15] DWORK C，MCSHERRY F，NISSIM K，et al. Calibrating noise to sensitivity in private data analysis[C]//Proceedings of the third conference on theory of cryptography. New York，NY：Springer，2006：265-284.

[16] 熊 平，朱天清，王晓峰，等. 差分隐私保护及其应用[J]. 计算机学报，2014，37(1)：101-122.