基于B-ISVM算法的物联网云存储数据伪装不良信息检测

贾长云 梁海军
(淮海工学院信息中心 连云港 222005)

摘 要 针对物联网云存储数据伪装不良信息隐蔽性造成的信息量预处理困难、深层次语义理解不准确和样本不均衡等问题,提出了一种基于B-ISVM(Boundary-Incremental SVM)算法的物联网云存储数据不良信息检测算法。在该算法中,首先采取基于均值和标准差的K均值初始聚类分析对云存储数据信息量进行样本空间训练分类;然后将所有样本类进行欧氏距离遍历计算,得到类间子聚类中心距离矩阵和各聚类中心的邻界子聚类区;再通过信息量伪装与筛选原理进行云存储信息真伪筛选,以不良信息在伪信息中发生的概率为指数、以数据安全度阈值和不良伪装信息模板向量集的相似度阈值为指标,对云存储信息量进行识别;最后进行增量模式学习,得到各分类样本最终的最优分类超平面,并将各类检测出的不良伪装信息进行输出。系统测试证明,该算法能快速有效地对物联网云存储数据中的伪装信息进行检测。

关键词 真伪信息,不良信息伪装,信息过滤,相似度计算,SVM
中图法分类号 TP393 **文献标识码** B

Things Networking Cloud Storage Data Bad Information Detection Based on Boundary-incremental SVM Algorithm

JIA Chang-yun LIANG Hai-jun
(Data Center, Huaihai Institute of Technology, Lianyungang 222005, China)

Abstract In order to solve the problem that things networking cloud storage data network camouflage bad information concealment causes information preprocessing difficulties, deep semantic understanding not accurate and sample disequilibrium, etc., this paper put forward the things networking cloud storage data bad information detection based on Boundary-Incremental SVM algorithm. This algorithm firstly takes mean initial clustering analysis based on mean and standard deviation of cloud storage data information for sample space training classification, and then puts all the sample classes euclidean distance traversal calculation to get son clustering center distance matrix between class and the clustering center adjacent boundary son clustering region, and then through the information content camouflage and selection principle of cloud storage information makes authenticity screening, using probability occurred in bad information in the false information as index, the data safety threshold and bad camouflage information template vector set of similarity threshold value as indexes, identifies the cloud storage information pseudo, finally carries on the incremental mode study, obtains each classification sample final optimal separating hyper plane, and will detect all kinds of bad camouflage information output. System test proves that this algorithm can fast effective in things networking cloud storage data of camouflage information detection

Keywords True bogus information, Bad information camouflage, Information filtering, Similarity calculation, SVM

1 引言

伪装不良信息过滤是物联网云存储数据净化的主要方式。伪装不良信息的过滤与一般文本信息过滤相比,物联网云存储不良信息过滤的目标和对象明确而且稳定,待过滤的内容常呈现多种方式和多变性及隐蔽性,样本分布不均衡且负面样本较少[1]。同时,针对文本内容进行的预处理是不良信息检测过程中最重要的环节,会直接影响到信息内容的识别。不良信息的隐蔽性信息特征使得其过滤比一般的信息数据处理更加困难,甚至目前许多不良信息均以伪装信息进行网络传播,与性教育和医疗健康信息的区分度越来越模糊,而当前常见的文本内容相似匹配算法模型都是某种形式上的概念匹配,系统难以理解文本的深层语义涵义[2]。潜在语义索引模型在进行文本语义理解时,具有较好效果,但其忽略了深层次的语义关联关系,同时该算法的执行效率较慢,难以实现实时过滤[3]。其他的文本内容匹配算法模型都依赖于先验知识,当前每类中心向量或每个词在类中的几率,对于传统的类别特征较为稳定的信息过滤效果较好,但是由于物联网云存

到稿日期:2012-05-13 返修日期:2012-09-11 本文受国家自然科学基金(61103017)资助。
贾长云(1960-),男,博士,副教授,主要研究方向为云计算、数据库技术,E-mail:jcy819@gmail.com;梁海军(1977-),男,硕士,主要研究方向为计算机网络。

储数据伪装不良信息的隐蔽性和灵活多变性，其中心向量极不稳定，每个词在类中的发生概率也难以预测[4]。虽然人工神经网络算法可以避免这类问题，但是其执行速度较慢[5]。基于此，本文提出了一种基于邻界区和增量学习的SVM(Boundary-Incremental SVM，B-ISVM)算法来对物联网云存储数据伪装不良信息进行检测，通过对网络信息量进行样本聚类和真伪信息筛选，以数据安全程度阈值和伪信息中不良信息发生的概率及相似度阈值为判决规则进行信息量伪装不良信息识别，并以邻界区和增量学习为尺度对整个信息量的样本空间进行降维和尺度检测，以序列风险最小化为准则进行不良伪装信息最优超平面的分割。系统测试证明，该算法具有较好的检测效率。

2 物联网云存储数据的信息伪装与筛选的基础定义

定义 1(信息伪装) 若 $\forall x_k \in (x)$ 为 (x) 的一个信息元，且 (x) 具有属性集合 $\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_k\}$，则设 $(x)\subset U$ 为 U 的一个有限信息，且 $(x)=\{x_1,x_2,\cdots,x_q\}$。同时有 $(x)^F\subset U$ 是 (x) 生成的一个 $\bar{F}$-信息伪装载体，且 $(x)^F=\{x_1,x_2,\cdots,x_p\}$；$(x)^-$ 为 (x) 的 $\bar{F}$-信息伪装损失，且 $(x)^-=\{x_{p+1},x_{p+2},\cdots,x_q\}$。

若 $(x)^F$ 具有属性集 α^F，且 $\alpha^F=\alpha\cup\{\alpha'\mid f(\beta)=\alpha'\in\alpha, f\in F\}$，且 $p\leqslant q$。

则将由信息伪装载体 $(x)_i^F$ 和 $\bar{F}$-伪装损失 $(x)_i^-$ 构成的信息对 $((x)_i^F,(x)_i^-)$ 称为 (x) 的 $\bar{F}$-信息伪装。其中 $f\in F$ 是元素迁移量，$F=\{f_1,f_2,\cdots,f_m\}$ 是元素迁移族，在实际应用中，$f\in F$ 是构造的一个具体函数，用于进行信息量的传递。

定义 2(伪装模) 设 λ_i^F 为 $(x)_i^F$ 关于 (x) 的 $\bar{F}$-信息伪装载体模数，且 $\lambda_i^F=\dfrac{\|(y)_i^F\|}{\|(y)\|}$。

其中 $(y)_i^F=\{y_1,y_2,\cdots,y_p\}$ 和 $(y)=\{y_1,y_2,\cdots,y_q\}$ 分别是 $(x)_i^F$，(x) 的信息量集合；$\|(y)_i^F\|$ 和 $\|(y)\|$ 分别是信息量集合构成的向量 $(y)_i^F=\{y_1,y_2,\cdots,y_p\}^{\mathrm{T}}$，$(y)=\{y_1,y_2,\cdots,y_q\}^{\mathrm{T}}$ 的 2-范数，其中 $\|y_i^F\|=(\sum_{j=1}^{p}y_j^2)^{\frac{1}{2}}$，$\|y\|=(\sum_{j=1}^{q}y_j^2)^{\frac{1}{2}}$，$(x)_i^F=\{x_1,x_2,\cdots,x_p\}$，$(x)=\{x_1,x_2,\cdots,x_q\}$，$p\leqslant q$。

同理，设 λ_i^- 为 $(x)_i^-$ 关于 (x) 的 $\bar{F}$-信息伪装损失模数，且 $\lambda_i^-=\dfrac{\|(y)_i^-\|}{\|(y)\|}$。

其中 $(y)_i^-=\{y_1,y_2,\cdots,y_s\}$ 和 $(y)=\{y_1,y_2,\cdots,y_q\}$ 分别是 $(x)_i^-$，(x) 的信息量集合；$\|(y)_i^-\|$ 和 $\|(y)\|$ 分别是信息量集合构成的向量 $(y)_i^-=\{y_1,y_2,\cdots,y_s\}^{\mathrm{T}}$，$(y)=\{y_1,y_2,\cdots,y_q\}^{\mathrm{T}}$ 的 2-范数，其中 $\|y_i^-\|=(\sum_{j=1}^{s}y_j^2)^{\frac{1}{2}}$，$\|y\|=(\sum_{j=1}^{q}y_j^2)^{\frac{1}{2}}$，$(x)^-=\{x_1,x_2,\cdots,x_s\}$，$p+s=q$，$p\leqslant s$，$p$ 为 $(y)_i^F$ 中信息量的个数。

则将由 $\{\lambda_i^F,\lambda_i^-\}$ 构成的数对称为 $\bar{F}$-信息伪装 $((x)_i^F,(x)_i^-)$ 关于 (x) 的 $\bar{F}$-伪装模。

定义 3(颗粒度) 设 $GRD((x)^F)$ 是信息 $(x)^F$ 关于 (x) 的颗粒度，则 $GRD((x)^F)=\dfrac{card((x)^F)}{card((x))}$，其中 $card((x)^F)$ 是 $(x)^F$ 的基数。

定义 4(筛选度) 设 $SFD((x)^F)$ 是 $(x)^F$ 关于 (x) 的筛选度，则 $SFD((x)^F)=\dfrac{SUR((x)^F)}{card((x))}=1-GRD((x)^F)$，且称 $SUR((x)^F)$ 是 $(x)^F$ 关于 (x) 的筛选剩余度，$SUR((x)^F)=card((x))-card((x)^F)$。

定理 1(信息伪装载体筛选第一定理) 设 $\{GRD((x)_1^F),GRD((x)_2^F),\cdots,GRD((x)_n^F)\}$ 是 $\{(x)_1^F,(x)_2^F,\cdots,(x)_n^F\}$ 的颗粒度，且有信息序列 $(x)_n^F\subseteq(x)_{n-1}^F\subseteq\cdots\subseteq(x)_1^F$，则 $GRD((x)_n^F)\leqslant GRD((x)_{n-1}^F)\leqslant\cdots\leqslant GRD((x)_1^F)$。

定理 2(信息伪装载体筛选第二定理) 设 $\{SFD((x)_1^F),SFD((x)_2^F),\cdots,SFD((x)_n^F)\}$ 是 $\{(x)_1^F,(x)_2^F,\cdots,(x)_n^F\}$ 的筛选度，且有信息序列 $(x)_n^F\subseteq(x)_{n-1}^F\subseteq\cdots\subseteq(x)_1^F$，则 $SFD((x)_1^F)\leqslant SFD((x)_2^F)\leqslant\cdots\leqslant SFD((x)_n^F)$。

$\bar{F}$-信息伪装载体筛选准则：$\bar{F}$-信息伪装载体 $\{(x)_1^F,(x)_2^F,\cdots,(x)_n^F\}$ 中，属性集合中属性数量最小的 $\bar{F}$-信息伪装载体 $(x)_u^F$ 满足 $card((x)_u^F)=\sum_{i=1}^{n}\max(card(x)_i^F)$。

$\bar{F}$-信息伪装损失筛选准则：$\bar{F}$-信息伪装损失 $\{(x)_1^-,(x)_2^-,\cdots,(x_n^-)\}$ 中，属性集合中属性数量最多的 $\bar{F}$-信息伪装损失 $(x)_r^-$ 满足 $card((x)_r^-)=\sum_{i=1}^{n}\min(card(x)_i^-)$。

定理 3(真信息筛选第一定理) 设 $\{(x)_1^F,(x)_2^F,\cdots,(x)_n^F\}$ 是 $\bar{F}$-信息伪装载体，$\{(x)_1^-,(x)_2^-,\cdots,(x_n^-)\}$ 是 $\bar{F}$-信息伪装损失，若 $card((x)_u^F)=\sum_{i=1}^{n}\max(card(x)_i^F)$，$card((x)_r^-)=\sum_{i=1}^{n}\min(card(x)_i^-)$，则 (x) 是由 $(x)_u^F$ 和 $(x)_r^-$ 构成的真信息，且 $(x)=(x)_u^F\cup(x)_r^-$。

定理 4(真信息筛选第二定理) 设 $\{(x)_1^-,(x)_2^-,\cdots,(x)_n^-\}$ 是 $\bar{F}$-信息伪装损失，若 $card((x)_v^F)=\sum_{i=1}^{n}\min(card(x)_i^F)$，$card((x)_s^-)=\sum_{i=1}^{n}\max(card(x)_i^-)$，则 (x) 是由 $(x)_v^F$ 和 $(x)_s^-$ 构成的真信息，且 $(x)=(x)_v^F\cup(x)_s^-$[6]。

3 物联网云存储数据中的不良信息伪装与识别

设信息序列由某一文档 D 的信息量组成，将其训练分解为特征向量 $\{x\mid x_1,x_2,\cdots,x_n\}$ 和决策变量 $\{c\mid c_1,c_2,\cdots,c_n\}$，并设特征向量的各分量间相对独立，则信息量可分为真信息和伪信息两类 (c_1,c_2)。在进行新文档信息检测时，计算该文档 X 属于真信息和伪信息的概率为：

$$p(c_j\mid x)=\frac{p(x\mid c_j)p(c_j)}{p(x)}=\prod_{i=1}^{n}\frac{p(x_i\mid c_j)p(c_j)}{p(x)}$$

则，对于任意信息量而言，数据的安全程度 $A(k)$ 可表示为：

$$\begin{cases}A(k)=card((x)_u^F)*p(c_1\mid x)\geqslant\theta, & \text{真信息}\\ A(k)=card((x)_i^-)*p(c_2\mid x)\leqslant\sigma, & \text{伪信息}\end{cases}$$

介于阈值间的数据安全度未知，当 $\theta=\sigma$ 时，信息量被准确分割为真伪信息簇，否则，信息量不能被完全分割，在进行伪装不良信息检测时，将除去真信息的剩余信息量作为伪信息量进行伪装不良信息检测。

设已知的经过训练和分类的伪装不良信息模板向量为 $\{P\mid P_1,P_2,\cdots,P_n\}$，而对于信息量 x，可通过计算求得其信息向量为伪信息的概率，则对于任意信息量 x_i，其与伪装不良信息的相似度 $\mathrm{Sim}(x_i,P_j)$ 为：

$$\mathrm{Sim}(x_i,P_j)=\frac{\sum_{i=1,j=1}^{n}\|\lambda_i^F*p(c_2\mid x)\|*\|P_j\|}{\sqrt{\sum_{i=1}^{n}\|\lambda_i^F*p(c_2\mid x)\|^2}*\sqrt{\sum_{j=1}^{n}\|P_j\|}}$$

设伪装不良信息判定的阈值为 ϕ，则：

$$\begin{cases}\mathrm{Sim}(x_i,P_j)\geqslant\phi, & \text{伪装不良信息}\\ \mathrm{Sim}(x_i,P_j)<\phi, & \text{其他信息}\end{cases}$$

且对于伪信息而言，其中伪装不良信息的占有比 δ 为：

$$\delta=\frac{card((x)_i^-)*p(c_2|x)}{\sum_{i=1}^{n}\max((x)_i^-)}$$

则在进行伪装不良信息检测时，计算信息量的真伪信息函数，依据伪装信息中不良信息发生的概率，计算伪信息中不良伪装信息的占有比重，并依据不良信息模板向量进行伪装不良信息相似度计算，通过阈值比较进行不良信息检测。其中漏检率 L 为：

$$L=|\mathrm{Sim}-\delta|=\left|\frac{\sum_{i=1,j=1}^{n}\|\lambda_i^F*p(c_2|x)|*\|P_j\|}{\sqrt{\sum_{i=1}^{n}\|\lambda_i^F*p(c_2|x)|^2}*\sqrt{\sum_{j=1}^{n}\|P_j\|}}-\frac{card((x)_i^-)*p(c_2|x)}{\sum_{i=1}^{n}\max((x)_i^-)}\right|$$

4 基于 B-ISVM 算法的物联网云存储数据伪装不良信息检测方法

SVM 算法是基于有序风险最小化的统计学习方法，其算法的核心思想是通过简单的线性分类器进行样本空间划分，通过在特征向量空间构造具有最佳间隔的超平面对样本进行划分，并使期望风险的上界最小。对于当前特征空间中先行不可分的分类模式，使用核函数将样本映射到高维空间，并在新空间体系内进行点积运算，使得复杂样本线性可分。但是 SVM 算法在针对海量、非平衡样本的信息检测时仍存在训练速度慢。针对样本规模大时如何减少训练样本数目的问题，本文提出了基于邻界区[7]的快速增量[8] SVM(B-ISVM)的不良信息检测方法，即首先对样本集进行聚类分析，以临界子聚类区来确定邻界区，将其作为检测算法的时序尺度，同时利用接近度因子进行邻界区筛选，并进行 SVM 训练，以完成信息量和不良信息模板量的超平面构造，在此基础上，计算样本的分散度和筛选因子，然后进行基于 KKT 条件的增量模式学习，完成 SVM 的分类器构造。

B-ISVM 算法的聚类分析算法采取基于均值和标准差的 K 均值初始聚类分析中心点的选取，与传统的 K 均值聚类分析方法相比，该算法可以在保证聚类精度的同时提高聚类的时间效率，具体的算法步骤如下：

Input：N 个 d 维聚类样本$\{x_1,x_2,\cdots,x_n\}$，其中 $x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}$，聚类个数为 K，迭代的终止条件为最大迭代次数 L 和收敛系数 T

Output：K 个聚类中心，并使数据与类中心相异度总和最小

Step1　计算 K 个初始聚类的中心$\{c_1,c_2,\cdots,c_K\}$，其中 $c_j=\{c_{j1},c_{j2},\cdots,c_{jd}\}$，聚类中心 m_{il} 为：

$$m_{il}=(u_l-\sigma_l)+\frac{2\sigma_l}{K};i=1,2,\cdots,K;l=1,2,\cdots,d$$

其中 m_{il} 是第 i 初始聚类中心的第 l 维分量，u_l 和 σ_l 分别为待聚类样本的第 l 维分量的平均值和标准差。

Step2　依据欧氏距离计算 i 个数据到聚类中心的距离，将数据分到具有最小距离的类别中，距离的计算公式为：

$$d(x_i,m_j)=\sqrt{\sum(x_{it}-m_{jt})^2};i=1,2,\cdots,n;j=1,2,\cdots,K$$

其中 $d(x_i,m_j)$ 为第 i 个数据向量到第 j 个聚类中心的距离。

Step3　更新 K 个聚类中心$\{m_1,m_2,\cdots,m_K\}$，计算公式为：

$$m_{jt}=\frac{1}{n}\sum x_{it};m_j=\{m_{j1},m_{j2},\cdots,m_{jd}\};l=1,2,\cdots,d$$

其中 m_j 为第 j 个聚类中心。

Step4　依据迭代条件判断每个聚类的收敛距离是否小于参数 T，若是，则结束聚类，否则迭代次数加 1，转向 Step2。其中收敛距离的计算公式为：

$$t_{jt}(k)=\sqrt{\sum(m_{jt}(k)-m_{jt}(k-1))^2};j=1,2,\cdots,K;l=1,2,\cdots,d$$

利用该聚类方法可以将文档的信息量进行基于均值和标准差的 K 均值初始聚类分析方法的真伪信息量向量样本空间构造和不良伪装信息模板向量分类。

在进行基于 B-ISVM 算法的物联网云存储数据伪装不良信息检测时，以邻界区增量为尺度进行目标信息量向量空间与不良伪装信息模板向量的相似度匹配，在单元尺度内，若当前信息量的空间指数降低，则可将该过程看作是信息量多维空间的降维。经过降维的信息量空间通过信息量伪装筛选与识别，将样本类中的每个样本分为真信息与伪信息向量信息集，通过真伪信息的发生概率进行数据安全度阈值比较，将伪信息中的每个信息向量集与不良信息模板向量集进行相似度计算，将伪信息向量集进行分割，构造超平面，最后以增量模式进行下一个尺度的样本匹配与分类。

基于 B-ISVM 算法的物联网云存储数据伪装不良信息检测算法描述如下：

Input：信息量(x)=$\{x_1,x_2,\cdots,x_q\}$，迁移函数 F=$\{f_1,f_2,\cdots,f_m\}$，伪装不良信息模板向量集$\{P|P_1,P_2,\cdots,P_n\}$，数据安全程度阈值(θ,σ)，伪装不良信息判定阈值 ϕ，聚类个数 K。

Output：不良信息集合 M。

Step1　利用 B-ISVM 算法的聚类分析算法进行信息量样本训练和聚类分析。

Step2　通过 $D=(D_{ij})_{c_+\times c_-}$ 计算样本类间聚类中心距离矩阵。其中 D_{ij} 为第 i 个正类中心到第 j 个负类中心的空间距离，将所有样本类进行遍历计算可得类间子聚类中心距离矩阵。

Step3　设 $1<t<c_+$，从 Step2 中得到的距离矩阵 D 中提取负类中心和正类的子聚类中心距离最近的 t 个子聚类中心，将其对应的子聚类区作为正类的子聚类中心 v_+^i 的 t-邻界子聚类区 $Z(v_+^i)$，同理可得到负类的各聚类中心的 t-邻界子聚类区。

Step4　依据信息伪装筛选第一、第二定理和真信息筛选第一、第二定理对信息量样本类进行真伪信息筛选与分类，并依据数据安全度阈值及不良伪装信息模板向量集的相似度阈值，对伪信息中发生不良伪装信息概率的信息向量集进行识别与分类。从而对各样本子类进行基于不良伪装信息检测的超平面 $g_i=0$ 的构造。

Step5　进行增量模式学习，得到各分类样本最终的最优分类超平面 $g^*=0$，并将各类的不良伪装信息存储于集合 M。

通过该算法，使得信息量 x 在增量模式下可以有效地对各样本的信息真伪进行训练，并以数据安全度阈值和不良信息伪装的发生概率及不良伪装信息向量集的相似度阈值为参考，对网络信息量进行过滤，有效地克服了网络信息海量、不良伪装信息检测难度大、基于 SVM 的训练时间过长等问题。

5 系统测试与分析

通过对比测试的方法，以传统的基于敏感词汇的方法和文献[9]的算法为参考，对本文算法的有效性进行了实验仿真。实验时，选取 1280 个词汇为基准作为网络信息传输信息量，其中包括了 195 个不良信息。实验开始时，对初始条件进

（下转第 138 页）

digms Workshop. Charlottesville, VA, USA: [s. n.], 1998
[3] Sheyner O, Haives J, Jha S. Automated generate on and analysis of attack graphs[C]// Proc 2002 IEEE Symposium on Security and Privacy. Oakland, California, USA, 2002: 254-265
[4] Ammann P, Wijesekera D, Kaushik S. Scalable, graph-based network vulnerability analysis [C]// Proc the 9th ACM Conference on Computer and Communications Security. Wa-shington, DC, USA, 2002: 217-224
[5] Lippmann R, Ingols K, Scott C, et al. Validating and Restoring Defense in Depth Using Attack Graphs[C]// Proc the 2006 Military communications Conference. Washington, DC, USA, IEEE Press, 2006: 1-10
[6] 汪渊, 蒋凡, 陈国良. 基于图论的网络安全分析方法的研究与实现[J]. 小型微型计算机系统, 2003, 24(10): 1865-1869
[7] 张涛, 胡铭曾, 云晓春, 等. 计算机网络安全性分析建模研究[J]. 通信学报, 2005, 26(12): 100-109
[8] 张维明, 毛捍东, 陈锋. 一种基于图论的网络安全分析方法研究[J]. 国防科技大学学报, 2008, 30(2)
[9] 陈秀真, 郑庆华, 管晓宏, 等. 层次化网络安全威胁态势量化评估方法[J]. 软件学报, 2006, 17(4): 885-897

(上接第 97 页)

行初始化，其中数据安全程度阈值(θ,σ)中，取$\theta=\sigma=\sum_{i=1}^{n}\min(card(x)_i^F)=\sum_{i=1}^{n}\max(card(x)_i^-)$，伪装不良信息判定阈值$\phi$取$\phi=0.99$，伪装不良信息模板向量为$\{P|P_1,P_2,\cdots,P_n\}$。

利用传统的基于敏感词汇的方法、文献[9]的算法和本文算法对实验目标数据集进行不良伪装信息检测，具体仿真效果图如图 1 所示。

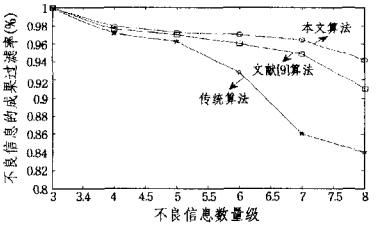

图 1　3 种算法的不良信息检测效果对比仿真图

通过图 1 可以看出，随着敏感词汇数据集合信息量级数的增加，文献[9]算法的过滤效果要优于传统效果，这是由于基于语义分析的物联网云存储数据伪装不良信息更能从语义关联的角度对信息量进行检测，检测面更广，效果更优。但同时可以看到，文献[9]的算法依赖于对词汇语义关联关系的提取，其提取信息的正确率和敏感词汇语义矩阵的抽取直接影响着其语义关联度的计算，且其算法基于先验概率来进行敏感词汇关联度的计算。与本文基于 B-ISVM 算法相比，本文的算法基于序列风险最小化，将信息量样本集基于均值和标准差的 K 均值方法进行聚类分析，并以数据安全度阈值和不良伪装信息相似度阈值的计算为参数在增量模式下进行信息量判定识别，其效果要优于传统的基于敏感词汇的算法和文献[9]的算法。具体的统计结果如表 1 所列。

表 1　3 种算法的检测效果对比

算法	敏感词汇数	检测个数
传统算法	195	98
文献[9]算法	195	177
本文算法	195	192

通过计算，3 种算法的漏检率如表 2 所列。

表 2　3 种算法的漏检率对比

算法	漏检率
传统算法	3
文献[9]算法	0.62
本文算法	0.13

通过漏检率可以看出，本文提出的基于 B-ISVM 算法的物联网云存储数据伪装不良信息检测算法对敏感词汇的检测效果要优于传统算法和文献[9]的算法，其不仅通过信息量样本空间分类构造对信息量进行真伪识别和伪信息中不良信息判别，而且利用增量模式进行尺度检测，在确保检测精度的同时提高了 SVM 的训练速度和不良伪装信息的检测效率，有利于物联网云存储数据伪装不良信息的检测。

结束语　通过对物联网云存储数据信息量的伪装与筛选原理进行基础知识定义，对信息量真伪信息筛选的定理进行了研究，并对信息量的伪信息发生不良信息的概率进行了探讨，以此为基础进行了数据安全度阈值计算和不良伪装信息模板向量集的相似度阈值计算。基于此，提出了基于 B-ISVM 算法的物联网云存储数据伪装不良信息检测算法，以样本信息量空间最优分割面构造为目的对信息量进行基于邻界区和增量模式的 SVM 算法检测，在相似度阈值范围内对各样本集进行不良伪装信息分类，从而得到不良信息向量集。实验证明该算法具有较好的检测效果和准确率。下一步的研究将集中于对阈值的修正和初始参数的统计验证，以确保算法检测效率的最佳与优化。

参 考 文 献

[1] 彭昱忠, 元昌安, 王艳. 基于内容理解的不良信息过滤技术研究[J]. 计算机应用研究, 2009, 26(2): 433-438, 447
[2] 季秀兰, 熊拥军. 基于网络安全的网页过滤模型及其关键算法[J]. 中南林业科技大学学报, 2011, 12: 197-201
[3] 李连, 朱爱红, 苏涛. 一种改进的基于向量空间文本相似度算法的研究与实现[J]. 计算机应用与软件, 2012, 2: 282-284
[4] 袁鼎荣, 钟宁, 张师超. 文本信息处理研究述评[J]. 计算机科学, 2011, 2: 9-13
[5] 唐云, 罗俊松. 基于粗糙集和 BP 神经网络的文本分类研究[J]. 计算机仿真, 2011, 6: 219-222, 283
[6] 耿红琴, 张冠宇, 史开泉. F-信息伪装与伪装-还原辨识[J]. 计算机科学, 2011, 38(2): 241-245
[7] 牟琦, 陈艺坤. 一种基于快速增量 SVM 的入侵检测方法[J]. 计算机工程, 2012, 12: 92-94
[8] 丁文军, 薛安荣. 基于 SVM 的 Web 文本快速增量分类算法[J]. 计算机应用研究, 2012, 4: 1275-1278
[9] 邵听, 徐倩漪. 物联网云存储数据伪装不良信息检测方法的研究与仿真[J]. 计算机仿真, 2012, 29(2): 135-138
[10] 邬贺铨. 物联网的应用与挑战综述[J]. 重庆邮电大学学报: 自然科学版, 2010, 22(5): 526-531