doi:10.11959/j.issn.1000-436x.2015328

面向医疗大数据的风险自适应的访问控制模型

惠榛，李昊，张敏，冯登国
（中国科学院 软件研究所，北京 100080）

摘 要：面对医疗大数据，策略制定者难以预测医生的访问需求，进而制定准确的访问控制策略。针对上述问题，提出一种基于风险的访问控制模型，能够适应性地调整医生的访问能力，保护患者隐私。该模型通过分析医生的访问历史，使用信息熵和 EM 算法量化医生侵犯隐私造成的风险。利用量化的风险，监测和控制对于医疗记录的过度访问以及特殊情况下的访问请求。实验结果表明，该模型是有效的，并且相比于其他模型能更为准确地进行访问控制。

关键词：风险访问控制；大数据；隐私；自适应
中图分类号：TP309 **文献标识码**：A

Risk-adaptive access control model for big data in healthcare

HUI Zhen, LI Hao, ZHANG Min, FENG Deng-guo
(Institute of Software, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: While dealing with the big data in healthcare, it was difficult for a policy maker to foresee what information a doctor may need, even to make an accurate access control policy. To deal with it, a risk-based access control model that regulates doctors' access rights adaptively was proposed to protect patient privacy. This model analyzed the history of access, applies the EM algorithm and the information entropy technique to quantify the risk of privacy violation. Using the quantified risk, the model can detect and control the over-accessing and exceptional accessing of patients' data. Experimental results show that this model is effective and more accurate than other models.

Key words: risk-based access control; big data; privacy; adaptive

1 引言

近年来，大数据越来越受到学术界和产业界的关注。如今，大数据不再是一个空泛的概念，而已经开始逐步落地，进入到产业化阶段。目前大数据的应用在电子商务、出行、医疗、养老等方面都已有呈现。例如，“淘宝大数据”从在线商铺中捞取假冒侵权商品[1]；“春运大数据”为交通调度提供了人流信息作为依据，并指出假日安全的可预测性[2]；医疗大数据被用来辅助临床决策、指导临床治疗和新药研发[3]。

大数据带来便利的同时也存在着安全隐患。如果管理不慎，造成对大数据的访问权限被滥用，就可能会引起严重的隐私泄漏。例如，美国犹他州卫生机构的数据被盗，造成超过 18 万人的信息泄露[4]。因此，对于大数据需要进行妥善的访问控制。在大数据时代，访问控制面临新的挑战，提出了自动化和自适应的需求[5,6]。由于访问控制策略的制定与应用环境紧密相关，需要结合相关的专业知识进行分析。面对大量且复杂的数据，人工制定策略是劳动密集型的工作，力有未逮。因此，大数据时代的访问控制需要一定的自动化能力来提高效率。另一方面，大数据不断更新的特性导致了访问控制场景的复杂，事先制定的访问控制策略在使用中可能需要调整。这就要求访问控制具有一定的自适应性，能够根据环境的变化（如数据使用情况的改变）来调整用户的访问能力。

收稿日期：2015-03-30；**修回日期**：2015-05-29
基金项目：国家自然科学基金资助项目（61402456）
Foundation Item: The National Natural Science Foundation of China (61402456)

医疗信息系统中存储着患者的所有医疗信息，缺少访问的控制必然导致患者的隐私毫无保障。虽然患者全部的医疗信息对于医疗信息系统都是可用的，但是本文希望访问控制系统能确保医生仅仅能够访问那些和她当前工作任务相关的信息，即满足 need–to-know 原则[7]。实现该原则要求策略制定者能够准确预计医生在不同情况下将会使用什么信息。这些信息不仅包括了一般情况下完成工作目标通常会使用的信息，还包括特殊情况下需要访问的额外信息。然而在大数据场景下，患者数量的庞大、医疗信息的丰富和复杂使策略制定者难以进行准确预测，并进而制定准确的访问控制策略。传统的访问控制[8~10]采用静态的授权方式，使用固定的策略，在访问控制实施时会过于严格，有碍于医生的正常工作，不适合处理上述困难。实际上，大部分的医院不得不授权医生对于其患者所有重要信息的访问权限，这导致了患者医疗数据的大量风险。

为了解决上述问题，已有学者提出了基于风险进行访问控制的方法[11]，并针对不同系统提出了一些风险量化和风险管理方法[11~22]。其中，Wang 等[12]的研究给出了处理医疗系统访问控制的一种较为实用的方案，但是在风险量化和敌手模型构建等方面仍然存在不足。

本文提出一种面向医疗大数据的风险访问控制模型，能够自动化地配置和适应性地调整用户的访问控制能力，拒绝侵犯隐私的用户的访问，同时允许意外情况下的正常访问。

通过分析医生历史访问记录自动确定其操作风险，从而自适应地调整其访问控制策略，降低其过度访问病人信息的威胁，使医生的访问尽量符合 need-to-know 原则。本文所述方法的主要步骤包括：1）通过信息熵来描述医生的访问行为；2）根据诚实医生和好奇医生在信息访问行为上的统计差异，采用 EM 算法将二者区分出来，计算出诚实医生访问行为先验分布，作为风险量化的基准（即诚实医生访问行为的熵）；3）依据医生行为偏离基准的程度对医生访问风险进行量化，并使用风险配额进行风险的管理，实施访问控制。

此外，本文修改了医生的行为模型，使其更为合理。现有模型中，医生为随机给定的工作目标选择医疗记录，而本文定义医生先选择工作目标（即对病人病情的估计），再选择医疗记录，该模型更接近真实情况。

2　相关工作

将风险引入访问控制是实现访问控制自动化和自适性的一种主要方法。近些年，风险访问控制逐渐受到研究者的关注。通过计算风险来进行访问控制判决，能够满足特殊情况下的访问需求，达成适应性；同时，利用风险对访问控制的动态调整减轻了安全管理员在策略制定和授权管理方面的工作量，一定程度上实现了自动化。JASON 计划办公室的研究报告[11]是较早将风险引入访问控制领域的工作，给出了风险量化和风险配额的概念，指出可以用风险来进行访问控制判定，从而处理意外情况。随后，开始涌现许多相关的研究工作，其中大多数关注如何扩展传统的访问控制模型来引入风险，以及扩展模型中的风险量化和管理的方法。

Cheng 等[13]提出一种模糊的多级安全访问控制模型。基于客体的安全标签和主体的信任度来估计授权用户读取数据所产生的风险，并且给出了一种用于多级安全系统的风险管理方法，设置可接受的风险区间，并且对风险区间进行划分形成风险带，根据用户风险所处的风险带，确定其访问能力。类似地，Ni 等[14]进一步使用模糊推理分析多级安全访问控制模型。研究结果显示模糊推理对于该模型下的风险估计是很合适的方法。同时，他们为主体和客体引入了模糊隶属函数，使用预定义的规则来实现基于风险的访问控制。这些方法都是对于多级安全访问控制模型即 MAC 模型的扩展。

此外，有部分学者研究对 RBAC 进行扩展，引入风险。Nissanke 等[15]假设存在一个风险等级的偏序集合，并且将这些风险等级分配给 RBAC 中的权限。从而，使用一个权限就可能比使用另一个权限造成更多的风险。此外，他们还提供了一种用风险来重构角色层级的方法。Celiker 等[16]提出了一种概率的风险管理框架来计算 RBAC 系统中用户的风险。Chen 等[17]提出了风险感知的 RBAC 模型，通过用户的信任度，用户执行角色的能力等来量化风险，并且能够返回多种类型的访问控制决策。Bijon 等[18]提出了一个用户风险感知的 RBAC 模型框架。定义了风险感知的 RBAC 组件，以及它们需要进行的改变。此外，还分析了 RBAC 中风险感知传统方法和量化方法的区别。

上述的方案与传统的访问控制方案是紧耦合的，相互之间具有紧密的联系。它们扩展了用户的

访问能力，使用户能够在意外情况下访问额外资源，然而用户访问同一资源的风险不会随着时间以及使用情况而有所调整。这使方案的自适应能力不足，需要较高的管理代价，影响了其应用价值。

针对以上问题，McGraw[19]提出了一种风险自适应的访问控制（RAdAC）机制。系统首先计算访问的风险和访问的需求程度，然后通过验证需求是否大于风险来进行访问判定。RAdAC 给出了一个实现自适应风险访问控制的高层框架，然而并没有给出具体的风险和需求的量化方法。Santos 等[20]提出了一个动态的基于风险的访问控制模型。该系统使用量化和集合的风险矩阵来管理用户对于云端资源的访问。谢文冲等[21] 对访问控制中的操作需求进行分析，提出静态和动态操作需求相结合的方式对访问风险进行量化。王超等[18]借鉴信息熵对事物不确定测度的思想，提出安全熵的概念，为访问控制策略的量化分析提供了科学方法。进一步，Wang 等[12]依据 RAdAC 的基本思想，提出了一种用于医疗系统的风险访问控制模型。该模型中，医生访问和目标不相关的信息会导致高风险，访问相关的信息则产生低风险。医疗记录和工作目标的相关性通过医生行为对比来近似，从而，能够拒绝高风险医生的访问，同时适应性的调节医生访问能力。然而，这个方案存在以下不足：要求使用的访问控制历史信息必须有明确的对于访问发起者是否为攻击者的标记；风险量化的基准来自访问控制历史中所有数据，当其中攻击者的访问记录较多时，会造成基准的风险量化值过高而使整个模型的有效性降低；设置的敌手模型仅仅考虑了选择医疗记录方面的风险，没有考虑通过伪造工作目标来选择医疗记录的风险。因此，需要进一步研究敌手行为，并且寻找更完善的风险量化方法。

3 问题描述

本节定义医疗信息系统中的敌手模型，包括医生类型定义及其行为模式假设。

为了完成工作职责，医生的访问行为可分为 2 个阶段，如图 1 所示。首先，医生需要为就诊患者确定一个初步目标，即医生怀疑或者确诊的病情，例如原位肿瘤；之后，医生选择与该工作目标相关的医疗记录来完成任务。当然，如果证明患者不是该病情，医生可以确定下一个目标，并且收集相关记录。该过程可以重复，直到患者被治愈或者离开。

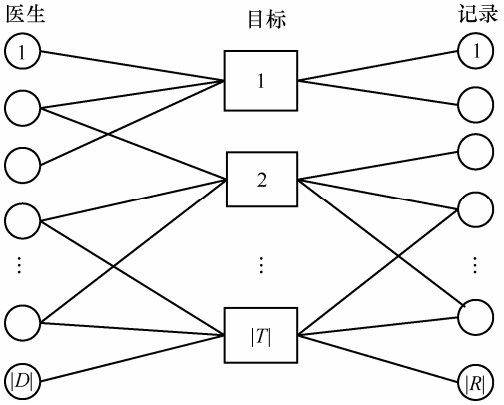


图 1 医生访问行为

系统中的所有用户是医生，都已被授权访问其患者的医疗记录。其中攻击者被称为“好奇医生”，而正常用户则称为“诚实医生”。诚实医生仅仅选择与病情相关的工作目标，并且访问与工作目标相关的医疗记录。好奇医生除此之外，还可能伪造一些工作目标来获取和该工作目标相关医疗记录，或者在完成某一个工作目标时选择和此工作目标不相关的医疗记录，以此 2 种方法来获取患者隐私。

为了形式化的描述敌手模型，本文使用如下符号。

D：医生的集合；

T：工作目标的集合；

M：医疗记录的集合；

P：患者的集合；

$\varphi: P,T \to [0,1]$：工作目标和患者病情的相关性函数，给定 $p \in P, t \in T$，返回二者的相关程度；

$\theta: T,M \to [0,1]$：医疗记录和工作目标的相关性函数，给定 $t \in T, r \in R$，返回二者的相关程度；

D_P：患者疾病的先验分布，反应了不同疾病的出现概率；

D_φ：工作目标的先验分布，和患者疾病的分布相关，由于疾病的出现概率具有规律性（服从分布 D_P），工作目标的分布具有类似的规律性；

D_θ：医疗记录和工作目标的相关性的先验分布。

假设 1 所有的医生都要履行自己的工作职责。

在医疗系统中，医生完成其工作职责是基本要求，也是符合实际的。也就是说，对于一次诊断而言，医生选择的工作目标集合中都必须是包含有和该疾病相关的工作目标，并且选择的医疗记录集合中也包含和该工作目标相关的医疗记录。

形式化地，给定 $p_i \in P$，记 $T_i \subseteq T$ 为医生选择的工作目标，则 $\sum_{t \in T_i} \varphi(p_i, t) \geqslant 1$；对于 $t_j \in T_i$，令

$R_j \subseteq R$ 是医生为完成该工作目标选择的医疗记录，则 $\sum_{r\in R_j}\theta(t_j,r)\geqslant 1$。

假设 2　诚实医生的行为满足：可选的工作目标应与患者疾病相关；其次，可选的医疗记录应与工作目标相关。考虑到可能出现的例外情况，这些选择不一定要完全符合先验分布。

形式化地，给定 $p_i\in P$，记 $T_i\subseteq T$ 为诚实医生选择的工作目标。对于 $\forall t\in T_i$ 的选择，以概率 $1-\alpha_1$ 服从分布 D_φ，以 α_1 的概率服从均匀分布；对于 $t_j\in T_i$，令 $R_j\subseteq R$ 是医生为完成该工作目标选择的医疗记录。对于 $\forall r\in R_j$ 的选择，以概率 $1-\beta_1$ 服从分布 D_θ，以 β_1 的概率服从均匀分布。

假设 3　好奇医生可能试图获得更多用户隐私信息，包括尝试更多的工作目标，或者在同一个工作目标下尝试获取更多医疗记录。

形式化地，给定 $p_i\in P$，记 $T_i'\subseteq T$ 为好奇医生选择的工作目标。对于 $\forall t\in T_i'$ 的选择，以概率 $(1-\alpha_1)(1-\alpha_2)$ 服从分布 D_φ，以 $\alpha_1+\alpha_2-\alpha_1\alpha_2$ 的概率服从均匀分布；对于 $t_j\in T_i$，令 $R_j'\subseteq R$ 是医生为完成该工作目标选择的医疗记录。对于 $\forall r\in R_j$ 的选择，以概率 $(1-\beta_1)(1-\beta_2)$ 服从分布 D_θ，以 $\beta_1+\beta_2-\beta_1\beta_2$ 的概率服从均匀分布。

诚实医生对于工作目标的选择，以及给定工作目标对于医疗记录的选择，都不是完全服从先验分布的。这表示医生需要额外资源的特殊情况：例如治疗复杂的病情需要参考多方面的医疗记录，甚至包括相关性较低的医疗记录。这种场景在医疗中是确实存在的。但好奇医生的选择更倾向于获取病人的其他信息，因此较诚实医生偏离先验分布更多，从而 2 类医生的行为存在一定差异。

相比于 Wang 等[12]描述的敌手模型，本文的敌手模型还考虑了医生在工作目标的选择中可能造成的隐私危害。如果好奇医生伪造一些和患者疾病不相关的工作目标，然后获取该工作目标相关的所有医疗记录，在 Wang 等[12]的模型中不会检测到这种行为，也就不会产生风险。而在本文的模型中，大量这种行为会导致该医生的工作目标后验分布与工作目标先验分布 D_φ 存在较大差异，从而被识别出来。

2 类医生的行为都在一定程度上符合先验分布，如果能够获得先验分布 D_φ 和 D_θ，就能够制定出准确的风险基准，并且根据医生访问行为偏离先验分布的程度进行风险量化和访问控制。然而，在现实中，很难直接获得先验分布。因此，本文从历史访问记录入手，获得近似的先验分布。具体而言，可以将诚实医生的访问历史记录近似看作先验分布。这就需要从无标记的历史访问记录中区分 2 类医生。依据假设可知，相比于诚实医生，好奇医生在工作目标或者医疗记录的选择上都具有更复杂的多样性，从而选择医疗记录的熵以及选择工作目标的熵也就更高，这为区分 2 类医生提供了依据。同时，也可以用熵作为量化风险的基础。在下一节中将会对区分方法和风险量化进行详细描述。

4　访问控制模型

本节描述访问控制模型，首先介绍整个模型的主要思想，然后具体介绍风险量化的细节，最后给出一个简单的案例进行说明。

4.1　概述

本文的访问控制模型对已有的基础模型[16,17]进行完善，首先利用传统的模型进行粗粒度的访问控制，再根据历史记录进行访问能力的调整。由第 3 节可知，需要解决的问题主要是：如何在先验分布 D_φ 和 D_θ 未知的情况下确定风险基准和量化风险以及如何进行风险的管理。为此，模型采用熵来描述医生访问行为特征，从而区分 2 类医生，进一步确定风险基准，并且进行风险量化；同时，采用风险配额的机制管理访问控制能力。访问控制模型如图 2 所示。该模型有 3 个核心部分。

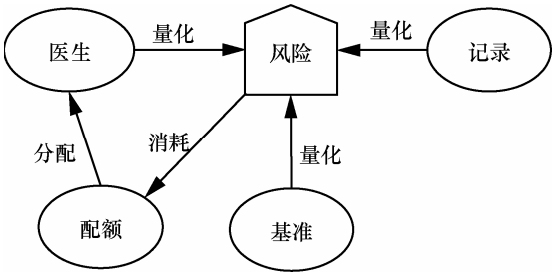


图 2　访问控制模型

1) 风险基准的确定。量化风险需要确定一个基准，以此基准作为阈值确定用户访问是否产生了风险。确定风险阈值就是寻找 2 类医生的访问行为熵的分界值，也就是区分 2 类医生。假设 2 类医生访问行为的熵值分别服从 2 个具有不同均值的高斯分布，而全体数据就是一个混合高斯分布。而采用 EM 算法[24]就可以求解这个混合高斯分布，从而完成医

生的分类。在获得 2 个高斯分布之后，根据它们均值的差值可以确定一个阈值作为基准。详细内容见 4.2 节。

2) 风险量化。对于医生的访问，仅仅通过单纯的一次访问行为难以确定他是否属于好奇医生，因此允许医生先进行访问，在一段时间之后，考察医生的行为来确定风险。具体地，本文分析医生在一段时间内的工作目标选择和医疗记录选择。由好奇医生和诚实医生的假设可知，不论在工作目标的选择还是给定工作目标下对于医疗记录的选择上，好奇医生的选择更加多样，即混乱程度更高，因此使用信息熵作为风险量化基本依据。熵值越高，说明其访问造成的风险越大。将医生访问行为的熵值比基准高的部分作为量化的风险。详细内容见 4.3 节。

3) 风险配额的计算和管理。使用风险配额的机制来管理基于风险的访问控制是通用的手段。风险配额表示系统对于每个医生所造成风险的容忍程度。为了进行访问控制，系统会定期为每位医生分配一定数目的风险配额。医生的每次访问造成一定量的风险，同时消耗等量的风险配额。如果医生的风险配额大于零，那么他能够继续访问；否则他的访问请求都会被拒绝，直到获得了新的风险配额。配额的分配是定期的，每次分配的风险配额应该满足诚实医生在下次分配前不会耗尽，即保证诚实医生的访问请求能够顺利通过。在本文中，每阶段配额的供给量根据好奇医生历史消耗的风险用统计学的方法计算。记 A_k 为第 k 个风险配额分配阶段，$Q_d(A_k)$ 为医生 d 在此阶段使用的访问配额，$E(A_k)=\dfrac{Q_d(A_k)}{|D|}$ 为平均值。在第 k+1 个风险配额分配阶段，需要分配的配额是通过前 k 个阶段的配额消耗平均值来确定的。认为前 k 个阶段的配额消耗平均值是正态分布的采样，那么可以得到该分布的均值 μ 和方差 σ。所要分配的配额就是在 $[\mu-n\sigma,\mu+n\sigma]$ 的范围内，其中，n 依据系统情况选择。

4.2 风险基准确定

由于医生的访问分为 2 个阶段，使用相似的方法，分别为 2 个阶段的访问行为计算基准。

记医生的一次访问行为信息为三元组 $\langle d,t,M_d\rangle$，其中 $d\in D,t\in T,M_d\subseteq M$。假设访问行为信息以其生成的时间顺序排列。为了量化风险基准需要引入如下符号。

$f(MS,e)$：该返回多集 MS 中元素 e 的出现次数。

ST_d：包含医生 d 选择的工作目标的多集。

SM_d^t：包含医生 d 在给定工作目标下，选择医疗记录的多集。

首先对于医生选择工作目标这个阶段，依据医生的访问历史信息，计算其工作目标的选择熵。医生 d 选择工作目标 t 的概率为

$$p_d(t)=\frac{f(ST_d,t)}{\sum_{t\in ST_d}f(ST_d,t)}$$

定义 1 工作目标熵。工作目标熵描述了医生 d 选择的工作目标的混乱程度，形式化地描述如下。

$$H^T(d)=-\sum_{t\in ST_d}p_d(t)\log(p_d(t))$$

记 $\{H^T(d_1),H^T(d_2),\cdots\}$ 为所有医生的工作目标熵，将其作为 EM 算法的输入，可获得 2 个高斯分布：(μ_H^T,σ_H^T) 和 (μ_M^T,σ_M^T)，以及它们的比例 a_H^T、a_M^T。分别表示诚实医生和好奇医生选择工作目标的熵的分布。

定义 2 工作目标熵基准。工作目标熵基准描述了从工作目标的选择上区分 2 类医生的阈值，形式化地描述如下。

$$H_B^T=a_H^T\mu_M^T+a_M^T\mu_H^T$$

下面计算医生选择医疗记录时产生的风险基准。对于相同的工作目标，诚实医生仅仅选择所需的医疗记录，而好奇医生还会选择其他的医疗记录来获得患者隐私。因此，好奇医生选择医疗记录的混乱程度更高。对于每个工作目标，量化医生选择医疗记录所可能产生的风险；将其总体值作为最终医生选择医疗记录的风险。

对于给定的工作目标 t，医生选择医疗记录 r 的概率为

$$p_{d,t}(r)=\frac{f(SM_d^t,r)}{\sum_{r\in SM_d^t}f(SM_d^t,r)}$$

定义 3 医疗记录熵。医疗记录熵描述了对于一个特定的工作目标，医生选择医疗记录的混乱程度。形式化地描述如下

$$H^R(d,t)=-\sum_{r\in SM_d^t}p_{d,t}(r)\log(p_{d,t}(r))$$

记 $\{H^R(d_1,t),H^R(d_2,t),\cdots\}$ 为对于给定的工作目标 t，所有医生的医疗记录熵，将其作为 EM 算法的

输入，获得 2 个高斯分布：$(\mu^R_{H,t},\sigma^R_{H,t})$ 和 $(\mu^R_{M,t},\sigma^R_{M,t})$，以及它们的比例 $a^R_{H,t}$、$a^R_{M,t}$，分别表示诚实医生和好奇医生选择医疗记录的熵的分布(给定工作目标 t)。

定义 4　医疗记录熵基准。医疗记录熵基准描述了对于给定的工作目标，从医疗记录的选择上区分 2 类医生的阈值，形式化地描述如下

$$H^R_{B,t}=a^R_{H,t}\mu^R_{M,t}+a^R_{M,t}\mu^R_{H,t}$$

4.3　风险量化

与风险基准的制定类似，分别为 2 个阶段的访问行为量化风险。

定义 5　工作目标选择风险。工作目标选择风险描述了医生选择工作目标所造成的访问风险，形式化地描述如下

$$Risk^T(d)=\max\{0,H^T(d)-H^T_B\}$$

由 H^T_B 定义可知，医生的工作目标熵可以高于、低于或者等于 H^T_B，在本文的模型中不处理直接负值，因此设置风险的最小值为 0。对于诚实医生而言，其工作目标熵通常小于 H^T_B，偶尔由于需要处理的特殊情况较多时才会高于 H^T_B，也就在此时才会产生风险。对于好奇医生而言，其工作目标熵通常高于 H^T_B，因此其工作目标选择风险总是会存在，并且积累的速度也会比诚实医生快。

定义 6　医疗记录选择风险。医疗记录选择风险描述了医生对于给定的工作目标，选择医疗记录所造成的访问风险，形式化地描述如下

$$Risk^R(d,t)=\max\{0,H^R(d,t)-H^R_{B,t}\}$$

与工作目标选择风险类似，本文模型认为最小医疗记录选择风险为 0。同时，可以知道好奇医生的该风险值通常会比诚实医生高，并且积累的速度也会比诚实医生快。

定义 7　医生访问总风险。医生访问总风险描述了医生访问行为造成的总风险，形式化地描述如下

$$totalRisk(d)=\omega_1 Risk^T(d)+\omega_2\sum_{t\in ST_d}Risk^R(d,t)$$

其中，$\omega_1,\omega_2\in[0,1],\omega_1+\omega_2=1$，用来表明 2 个阶段风险危险程度的不同。

直观上易知，好奇医生选择工作目标的多样性和选择医疗的多样性都会使其访问行为的风险高于诚实医生。

4.4　案例效果分析

本节通过具体的案例从理论上分析本文模型相较于以往的设计具有更好的效果。

案例 1　假设 t 是与某个罕见病相关性最高的工作目标。如果医生 d 的工作目标选择历史中多次出现 t（即通过伪造 t 来获取相关的医疗记录），而其他医生的工作目标选择历史中很少出现。那么，应有 $Risk^R(d,t)$ 和其他医生相差不多，但是 $Risk^T(d)$ 较大，从而 $totalRisk(d)$ 较大。医生 d 的风险配额会快速消耗完，其后续的访问都会被拒绝。

该案例说明了 $Risk^T(d)$ 的作用。对于一些罕见病，如果某一个医生经常提交需要治疗这些病的工作目标，那么其风险值会比较高，从而更快地消耗风险配额，被限制访问。

案例 2　假设存在 4 个医生 d_1,d_2,d_3,d_4，其中 d_1 是诚实医生，只会访问资源 r_1,r_2；d_2,d_3,d_4 都是好奇医生，他们会访问资源 r_1,r_2,r_3,r_4。那么在均匀选择记录的假设下，诚实医生的记录选择熵为 1，好奇医生的记录选择熵为 2。在以往的模型中[9,14]，当好奇医生比例较大时，会影响基准值。而在本文所述模型中风险基准更接近于低值。例如使用 Wang 方法[12]确定的熵基准为 1.75，本文模型的熵基准为 1.25，更加接近诚实医生的记录选择熵。因此具有更好的区分能力。

该案例说明了使用 EM 算法确定基准的优势。通过 EM 算法区分 2 类医生之后建立的基准，更加接近诚实医生的行为。在原始数据中好奇医生较多时，仍然能保证模型有效。

5　实验和评估

实验的目标是检测本文访问控制模型的有效性，即是否能够有效地抑制好奇医生的访问。由前文可知，当医生的风险过高时，其后续访问请求会被拒绝。因此，可以通过研究好奇医生的风险高低来测试访问控制模型的有效性：当好奇医生的风险都很高，并且风险高的医生都是好奇医生时，可以认为此访问控制模型是有效的。本文以 Wang 的方法[12]为参考，来检查本文模型的有效性。

5.1　实验准备

本文参照 Wang 所述方法[12]，模拟生成 600 名医生的访问历史信息。并按照医疗体系惯例，采用 ICD-10 分类标注工作目标和医疗记录。诚实医生与好奇医生所访问的资源满足前文的假设 1~假设 3。

此外，假设选择工作目标和选择医疗记录 2 部分的风险具有同等的威胁性，即令 $\omega_1=\omega_2=0.5$，$\alpha_2=\beta_2,\alpha_1=\beta_1$。

5.2　实验结果及说明

在每次仿真中，计算每个医生的风险值，并且按风险高低排序。为了检验方法的有效性，本文考查 3 个指标。准确率@K 表示风险最高的前 K 个医生中好奇医生的比率；召回率@K 表示风险最高的前 K 个医生中好奇医生占全体好奇医生的比例；F1 分数@K 表示准确率和召回率的几何平均。

1) 不同的记录请求数目的实验结果

第一类实验用来测试该模型对于不同的记录请求数目的效果。实验表明对于不同的记录请求数目，本文的模型是有效的。模型的总体性能随着访问请求数目的增加而提升。此外，本文模型的效果相较以往 Wang 的方案[8]有所提升。

在每个实验中，有 600 名医生，其中 10%（即 60 名）是好奇医生，其他的是诚实医生。过度访问的比例为 5%。实验结果如图 3 和表 1 所示。

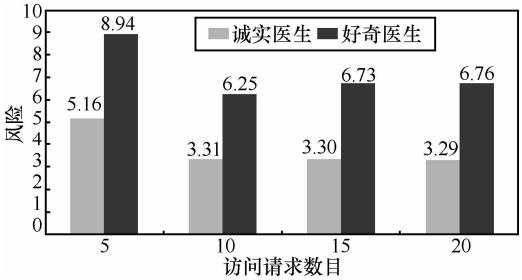


图 3　不同记录访问数目下的风险值

由图 3 可知，好奇医生的平均风险是诚实医生的 1.7 到 2.1 倍。这说明 2 类医生是明显可区分的，本文的模型是有效的。由表 1 可知，在所有的实验中，风险值最高的前 20 名医生都是好奇医生，风险值最高的前 50 名医生中准确率也保持在 88%以上；对于召回率而言，当访问数目为 10 或者更高的时候，至少有 88%的好奇医生的风险值在前 50 名内；模型总体的性能随着访问数目的增加而提升。这是因为更多的信息有助于算法准确的学习医生的行为，更准确地分类医生以及量化风险值。在访问数目达到 10 之后，再更多地增加访问数目，

表 1　　不同的记录访问数目下的性能指标

访问数目	K	准确率@K		召回率@K		F1 分数@K	
		本文	WANG	本文	WANG	本文	WANG
5	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	0.95	0.33	0.32	0.50	0.47
	30	0.97	0.88	0.33	0.32	0.64	0.59
	40	0.93	0.80	0.62	0.53	0.72	0.64
	50	0.88	0.78	0.73	0.65	0.80	0.67
10	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	1.00	0.33	0.33	0.50	0.50
	30	1.00	0.97	0.50	0.48	0.67	0.64
	40	1.00	0.78	0.67	0.52	0.80	0.62
	50	0.96	0.74	0.80	0.62	0.87	0.67
15	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	1.00	0.33	0.33	0.50	0.50
	30	1.00	0.97	0.50	0.48	0.67	0.64
	40	1.00	0.88	0.67	0.59	0.80	0.70
	50	0.94	0.76	0.78	0.63	0.85	0.69
20	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	1.00	0.33	0.33	0.50	0.50
	30	1.00	0.93	0.50	0.47	0.67	0.62
	40	1.00	0.85	0.67	0.57	0.80	0.68
	50	0.96	0.80	0.80	0.67	0.87	0.72

并不会明显提升性能。这说明本文的方案在访问数目很小的时候就达到了最优。

2) 不同的过度访问比例的实验结果

第二类实验测试对于不同的过度访问比例，该模型的有效性。实验表明本文的风险访问控制模型是有效的。模型的总体性能随着过度访问比例上的增加而提升。此外，本文模型的效果相较以往的 Wang 的方案[12]有所提升。

在每个实验中，有 600 名医生，其中 10%（即 60 名）是好奇医生，其他的是诚实医生。对于每个工作目标，记录的访问请求的数目都是 10。实验结果如图 4 和表 2 所示。

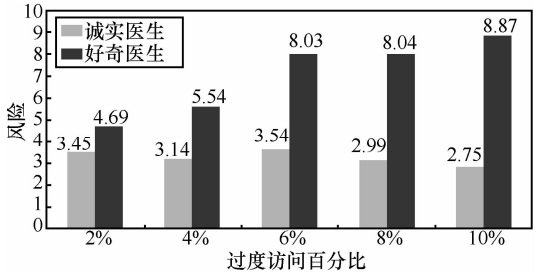


图 4　不同过度访问比例下的风险值

由图 4 可知，好奇医生的风险值是诚实医生的 1.4 到 3.23 倍，说明 2 类医生是明显可区分的，本文模型是有效的。由表 2 可知，实验的性能随着过度访问的比例增加而提升。这是因为当好奇医生的

表 2　　**不同的过度访问比例下的性能指标**

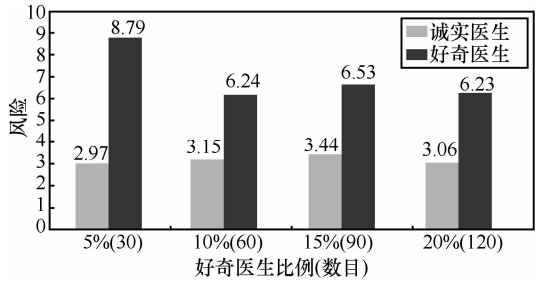
比例	*K*	准确率@*K*		召回率@*K*		F1 分数@*K*	
		本文	WANG	本文	WANG	本文	WANG
2%	10	0.70	0.70	0.12	0.12	0.20	0.20
	20	0.65	0.65	0.22	0.21	0.33	0.33
	30	0.67	0.63	0.33	0.32	0.44	0.42
	40	0.63	0.60	0.42	0.40	0.50	0.58
	50	0.54	0.52	0.45	0.43	0.49	0.47
4%	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	0.95	0.33	0.32	0.50	0.48
	30	0.97	0.87	0.48	0.44	0.64	0.58
	40	0.93	0.72	0.62	0.48	0.74	0.58
	50	0.84	0.72	0.70	0.60	0.76	0.65
6%	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	1.00	0.33	0.33	0.50	0.50
	30	1.00	0.97	0.50	0.49	0.67	0.65
	40	1.00	0.92	0.67	0.61	0.80	0.74
	50	1.00	0.90	0.83	0.75	0.90	0.81
8%	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	1.00	0.33	0.33	0.50	0.50
	30	1.00	1.00	0.50	0.50	0.67	0.67
	40	1.00	0.98	0.67	0.65	0.80	0.78
	50	1.00	0.92	0.83	0.77	0.90	0.84
10%	10	1.00	1.00	0.17	0.17	0.29	0.29
	20	1.00	1.00	0.33	0.33	0.50	0.50
	30	1.00	1.00	0.50	0.50	0.67	0.67
	40	1.00	0.98	0.67	0.65	0.80	0.78
	50	1.00	0.92	0.83	0.77	0.90	0.84

过度访问越多，会使其行为和诚实医生的差距越大，也就越容易区分以及确定风险量化的基准。具体地，本文模型在好奇医生的过度访问比例仅为6%时就已经十分有效。当过度访问比例高于 6%时，准确率@50 始终保持在 100%，召回率@50 最低为83%。当过度访问比例过低时，例如 2%，要区分 2 类医生就相对困难许多，本文模型此时的准确率@50 为 54%，召回率@50 为 45%。然而，不论过度访问比例处于何种情况，本文模型的性能始终优于 Wang[12]的方法。

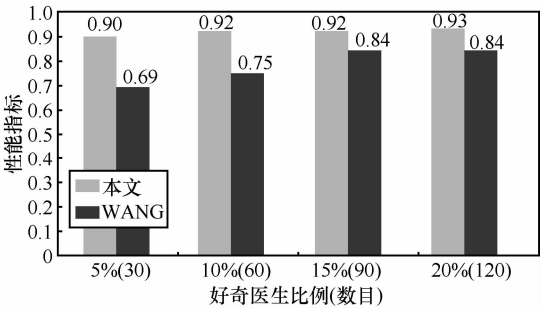
3) 不同的好奇医生比例的实验结果

第三类实验测试对于不同的好奇医生比例，该模型的有效性。实验表明本文的模型对于不同比例的好奇医生都是有效的。模型总体的性能随着好奇医生比例的增加而提升。此外，本文模型的效果相较以往 Wang 的方案[12]有所提升。

在每个实验中，有 600 名医生，对于每个工作目标，访问请求的数目都是 5，过度访问的比例为 5%。实验中计算准确率@x、召回率@x 和 F1 分数@x，其中，x 是好奇医生的数目。此时，这 3 个性能指标值相同，实验结果如图 5 所示。



(a) 不同的好奇医生比例下的风险



(b) 不同的好奇医生比例下的性能指标

图 5　不同的好奇医生比例下的风险和性能指标

如图 5(a)所示，好奇医生的风险是诚实医生的 2.1 到 3.3 倍。这说明 2 类医生是明显可区分的，本文的模型是有效的。有图 5(b)可知，模型的性能随着好奇医生的增加而上升。这是因为，在医生总数相同的情况下，好奇医生越多则诚实医生越少，高风险的医生中好奇医生的数目也就越多。总体性能上，本文模型依然优于 Wang 的方法[12]。此外，相比与 Wang 的模型在好奇医生比率高于 15%后性能没有变化的特点，本文模型具有一定的改善。

6　结束语

本文提出了一种面向医疗大数据的风险访问控制模型，能够自动化地配置访问控制策略并且适应性地调整医生的访问能力，从而保护患者的隐私信息不会因医生滥用访问权限而泄露。该模型利用传统访问控制模型的配置进行分析，采取动态学习的方式建立资源和用户的授权关系。用户的访问能力取决于他的历史访问行为，及其访问的资源和当前工作的相关程度；依据医生在工作目标选择和医疗记录选择上的差异建立敌手模型；使用 EM 算法和信息熵完成风险基准的确定和风险的量化，区分好奇医生和诚实医生。本文还给出了风险管理的方法和依据风险进行访问控制判定的方案。最后，本文实现了一个原型系统，并且使用人工数据进行了对比实验。实验结果显示该模型对于抑制好奇医生的访问是有效的，同时相比于其他方案效果更好。

参考文献：

[1] http://gd.qq.com/a/20150316/021748.htm[EB/OL].

[2] http://www.chnrailway.com/html/20150317/867407.shtml[EB/OL].

[3] http://news.ifeng.com/a/20150121/42985231_0.shtml[EB/OL].

[4] http://www.eweek.com/c/a/Health-Care-IT/Utah-Health-Care-Data-Breach-Exposed-About-780000-Patient-Files-189084[EB/OL].

[5] SANDHU R. The future of access control: attributes, automation, and adaptation[A]. Computational Intelligence, Cyber Security and Computational Models[C]. 2014. 45-45.

[6] 冯登国, 张敏,李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1):246-258.
FENG D G, ZHANG M, LI H. Big data security and privacy protection [J]. Chinese Journal of Computers, 2014, 37(1):246-258.

[7] SANDHU R, PIERANGELA S. Access control: principle and practice [J]. Communications Magazine, 1994, 32(9): 40-48.

[8] GRAHAM G S, DENNING P J. Protection: principles and practice [A]. Proceedings of the Conference[C].ACM, 1972.417-429.

[9] HARRISON M A, RUZZO W L, ULLMANJ D. Protection in operating systems[J]. Communications of the ACM, 1976, 19(8): 461-471.

[10] SANDHU R, COYNE E, FEINSTEIN H, et al. Role-based access control models [J]. Computer, 1996, 29(2):38-47.

[11] JASONPROGRAMOFFICE. Horizontal Integration: Broader Access Models for Realizing Information Dominance[R]. The MITRE Corporation 2004.

[12] WANG Q, JIN H. Quantified risk-adaptive access control for patient privacy protection in health information systems[A]. Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security[C]. ACM, 2011. 406-410.

[13] CHENG P C, ROHATGI P, KESER C, et al. Fuzzy multi-level security: an experiment on quantified risk-adaptive access control[A]. IEEE Symposium on Security and Privacy(SP '07)[C]. IEEE, 2007. 222-230.

[14] NI Q, BERINNO E, LOBO J. Risk-based access control systems built on fuzzy inferences[A]. In Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security[C]. ACM. 2010. 250-260.

[15] NISSANK N, KHAYAT E J. Risk based security analysis of permissons in RBAC [A]. Proceedings of the 2nd International Workshop on Security in Information Systems, 2004[C]. IEEE, 2004. 332-341.

[16] CELIKEL E, KANTARCIOGLU M, THURAISINGHAM B M, et al. A risk management approach to RBAC [J]. Risk and Decision Analysis, 2009, 1(1): 21-33.

[17] CHEN L, JASON C. Risk-aware role-based access control[A]. Security and Trust Management[C]. Springer Berlin Heidelberg, 2012. 140-156.

[18] BIJON K Z, KRISHNAN R, SANDHU R. A framework for risk-aware role based access control[A]. Proceedings of the 6th Symposium on Security Analytics and Automation 2013[C]. IEEE, 2013. 462-469.

[19] MCGRAW R. Risk-adaptable access control RADAC [R]. Privilege (Access) Management Workshop. 2009.

[20] SANTOS D R, WESTPHALL C M, WESTPHALL C B. A dynamic risk-based access control architecture for cloud computing [A]. IEEE Network Operations & Management Symposium[C]. IEEE, 2014. 1-9.

[21] 谢文冲, 杨英杰, 汪永伟,等. 基于风险的访问控制操作需求计算方法研究[J]. 计算机工程与设计, 2013, 34(7): 2281-2285.
XIE W C, YANG Y J, WANG Y W, et al. Research on computing method of operational need in risk-based access control[J]. Computer Engineering and Design, 2013, 34(7): 2281-2285.

[22] 王超, 陈性元. 基于加权熵的访问控制策略安全性分析研究[J]. 电子学报, 2013, 41(1): 47-51.
WANG C, CHEN X Y. An approach for security analysis to access control policy based on entropy-weigh[J]. Acta Electronica Sinica, 2013, 41(1): 47-51.

[23] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B (Methodological), 1977: 1-38.

作者简介：

惠榛（1987-），男，青海西宁人，中国科学院软件研究所博士生，主要研究方向为大数据访问控制、数据隐私保护。

李昊（1983-），男，河南固始人，中国科学院软件研究所副研究员，主要研究方向为网络与系统安全。

张敏（1975-），女，安徽萧县人，中国科学院软件研究所高级工程师，主要研究方向为数据隐私保护、可信计算和云存储安全。

冯登国（1965-），男，陕西靖边人，中国科学院软件研究所研究员、博士生导师，主要研究方向为信息安全和密码学。