Peer-to-Peer Netw. Appl. (2018) 11:334–347
DOI 10.1007/s12083-016-0506-8

Efficient privacy-preserving online medical primary diagnosis scheme on naive bayesian classification

Xiaoxia Liu[1] · Hui Zhu[1] · Rongxing Lu[2] · Hui Li[1]

Received: 1 April 2016 / Accepted: 8 September 2016 / Published online: 7 October 2016

Abstract With the advances of machine learning algorithms and the pervasiveness of network terminals, online medical primary diagnosis scheme, which can provide the primary diagnosis service anywhere anytime, has attracted considerable interest recently. However, the flourish of online medical primary diagnosis scheme still faces many challenges including information security and privacy preservation. In this paper, we propose an efficient and privacy-preserving medical primary diagnosis scheme, called PDiag, on naive Bayes classification. With PDiag, the sensitive personal health information can be processed without privacy disclosure during online medical primary diagnosis service. Specifically, based on an improved expression for the naive Bayes classifier, an efficient and privacy-preserving classification scheme is introduced with lightweight polynomial aggregation technique. The encrypted user query is directly operated at the service provider without decryption, and the diagnosis result can only be decrypted by user. Through extensive analysis, we show that PDiag ensures users' health information and service provider's prediction model are kept confidential, and has significantly less computation and communication overhead than existing schemes. In addition, performance evaluations via implementing PDiag on smartphone and computer demonstrate PDiag's effectiveness in term of real environment.

Keywords Online medical primary diagnosis · Privacy-preserving · Naive Bayes classifier · Polynomial aggregation

1 Introduction

Online medical primary diagnosis system, which can provide the pre-diagnosis service anywhere anytime and guide users' behaviors, has attracted considerable interest. Due to the lack of medical doctors, the waiting time of seeing doctors for patients increased in many countries [1–3]. It is true that most people are unlikely to be familiar with medical departments in hospitals and certainly not familiar with the symptoms associated with the different diseases. If an individual goes to hospital without any preparation, it may be a frustrating and time-wasting exercise in case she/he consults a doctor not trained or specialized in the particular disease. As ubiquitous healthcare services are becoming more and more popular, especially under the urgent demand of the global aging issue, primary diagnosis scheme should be developed to help people get a primary disease diagnosis knowledge conveniently [4]. As one of the most popular machine learning techniques, naive Bayes classification, which is a simple and effective probabilistic classification method, has been widely used for predicting various diseases in medical informatics [5–7]. For instance, through

Xiaoxia Liu
17802929568@163.com

✉ Hui Zhu
zhuhui@xidian.edu.cn

Rongxing Lu
rxlu@ntu.edu.sg

Hui Li
lihui@mail.xidian.edu.cn

[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China

[2] School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang 639798, Singapore

building a prediction model upon existing clinical datasets by using naive Bayes classification algorithm, an untrusted third-party organization can provide online medical primary diagnosis service, and users can request the primary diagnosis service through Internet before they go to a hospital for diagnosis [8].

However, due to the sensitivity of users' health data, users are usually reluctant to offer their medical information to an untrusted third-party organization for obtaining online medical primary diagnosis service. Meanwhile, as the predication model is also private and valuable asset, the third-party organization may be reluctant to reveal the information of prediction model as well. Therefore, the flourish of online medical primary diagnosis scheme still hinges upon how to fully understand and manage these challenges including information security and privacy preservation. Thus, it is of great importance to develop adequate security techniques for protecting privacy of medical users as well as the prediction model of third-party organization. We can image that, if a stronger protection is available, users may be more willing to receive many services provided by the third-party through internet [9–11]. To address these challenges, different homomorphic encryption techniques are introduced in the medical diagnosis system [12–14]. However, most of the homomorphic encryption schemes are not very efficient and not quite appropriate for providing online medical primary diagnosis service.

In order to address the above-mentioned challenges, we propose an efficient and privacy-preserving online medical primary diagnosis scheme, called PDiag, on naive Bayes classification, which preserves users' query information and the service provider's diagnosis model. With PDiag scheme, users can achieve privacy-preserving online medical primary diagnosis service by themselves according to the diagnosis model stored at the service provider. The service provider will provide medical prediction service without the leakage of disease diagnosis model. Specifically, the main contributions of this paper are threefold.

First, the proposed PDiag is secure and privacy-preserving. With PDiag, the user can keep her/his query information and the final primary diagnosis result secret from the service provider, meanwhile the service provider can also keep the diagnosis model secret from the user. In our novel PDiag scheme, the user first preprocesses query information by introducing different random numbers, and the service provider calculates the preprocessed query information by naive Bayes classification standard with lightweight polynomial aggregation technique, then the user will obtain the primary diagnosis result.

Second, PDiag can provide the online medical primary diagnosis service with a high accuracy. To evaluate the accuracy of the proposed PDiag scheme, we construct an improved expression for the naive Bayes classifier, verify the correctness of PDiag, and do experiments over two real datasets from the UCI machine learning repository. Final experimental results show that PDiag can achieve a high accuracy.

Third, PDiag is efficient in terms of computation and communication overhead. Since the user only interacts with the service provider for once during the process of medical diagnosis, different from other time-consuming homomorphic encryption techniques, all of the encryption operations are based on lightweight polynomial aggregation techniques. Meanwhile, we also develop a custom simulator built in Java, and implement PDiag over smartphone and computer in real environment. Performance evaluation demonstrates that our proposed PDiag can provide an efficient online medical primary diagnosis service in real life.

The remainder of this paper is organized as follows. In Section 2, we formalize the system model, security requirements, and identify our design goal. In Section 3, we review the bilinear pairing and the naive Bayes classifier, and propose an improved expression for naive Bayes classification standard as the preliminaries. Then, we present our PDiag scheme in Section 4, followed by the security analysis and performance evaluation in Section 5 and Section 6, respectively. We also review some related works in Section 7. Finally, we draw our conclusions in Section 8.

2 System model and design goal

In this section, we formalize the system model, security requirements, and identify our design goal.

2.1 System model

In our system model, we mainly focus on how the service provider offers efficient and privacy-preserving online medical primary diagnosis services to users whose medical information is sensitive. Each user is equipped with a computer or smartphone, which can connect with the service provider for achieving online medical primary diagnosis service. Specifically, the system consists of two parts: *service provider* (SP) and *user*, as shown in Fig. 1.

– We consider an authorized data analysis organization as SP, which owns a naive Bayes classifier built upon existing clinical datasets which are initially provided by hospitals, and it provides online medical primary diagnosis service for registered medical users. SP computes the encrypted medical data by naive Bayes classification standard. Furthermore, although SP is a server with

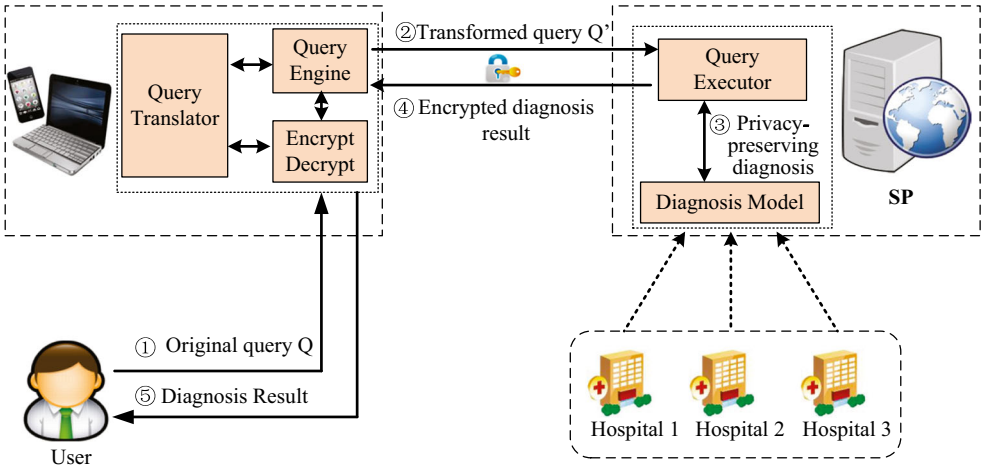

Fig. 1 System model under consideration

high performance in computation and storage, since thousands of users may query services at the same time, the efficiencies of computation and communication are still challenging.

– The medical query information is represented with a query vector, and the registered user can query the privacy-preserving medical forecasting service by the query vector. Considering the query vector may contain some sensitive information of user, and sending the query vector in plaintext to SP may lead to privacy leakage, the user should perform some encryption operations during the process of query. Moreover, in order to lower energy costs, the encryption efficiency of terminals is also very prerequisite.

2.2 Security requirements

The privacy of users' medical query information and the confidentiality of prediction model are crucial for the success of online medical primary diagnosis service. In our security model, we consider users and SP are honest-but-curious. Specifically, SP provides the online medical primary diagnosis service correctly, but it is also curious to users' medical query information; users will honestly execute the operations to achieve the final prediction result, but they also try to analyze the information of prediction model; moreover, users may try to access the online medical primary diagnosis service without registering. Therefore, to guarantee the privacy of users' medical query information and the confidentiality of prediction model, the following security requirements should be satisfied.

– *Privacy*. Protecting users' medical query information from SP, i.e., even if SP receives the encrypted query vector from the user, it cannot identify the user's medical query information in plaintext form. At the same time, though SP computes the intermediate parameters according to the encrypted query vector of user, it cannot obtain the final diagnosis result.
– *Confidentiality*. Keeping the diagnosis model secret from users, i.e., even if the user obtains the intermediate parameters calculated by SP, she/he cannot identify the parameters of diagnosis model.
– *Authentication*. Authenticating an encrypted query vector that is really sent by a legal user and has not been altered during the transmission, i.e., if an illegal user forges a data query, this malicious operation should be detected. Meanwhile, the responses from SP should also be authenticated so that the user can receive the authentic and reliable query result.

2.3 Design goal

Under the aforementioned system model and security requirements, our design goal is to develop an efficient and privacy-preserving online medical primary diagnosis scheme. Specifically, the following three objects should be achieved.

– *The security requirements should be guaranteed*. If the scheme does not consider the security, users' medical query information and the diagnosis model could be disclosed. Then, the online medical primary diagnosis scheme cannot flourish. Thus, the proposed scheme should achieve the confidentiality and authentication simultaneously.
– *Data query result's accuracy should be guaranteed*. It is important that applying the privacy-preserving strategy cannot compromise the accuracy. Therefore, the proposed scheme should also achieve a high accuracy.
– *Low communication overhead and low computation complexity should be guaranteed*. Considering the real-time

requirements of online medical primary diagnosis service and the diversity of terminals, the proposed scheme should have low overhead in terms of communication and computation.

3 Preliminaries

In this section, we review the bilinear pairing [15] and the classifier of naive Bayes [16], and then improve the naive Bayes classification standard, which will serve as the basis of our proposed PDiag scheme.

3.1 Bilinear pairing

Let $\mathbb{G}$, $\mathbb{G}_T$ be two cyclic groups of the same prime order q, and g is a generator of group $\mathbb{G}$. Suppose $\mathbb{G}$ and $\mathbb{G}_T$ are equipped with a pairing, i.e., a non-degenerated and efficiently computable bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ such that $e(g, g) \neq 1_{\mathbb{G}_T}$ and $e(u^a, v^b) = e(u, v)^{ab}$ for all $u, v \in \mathbb{G}$ and $a, b \in \mathbb{Z}_q^*$. Moreover, $e(u, v)$ can be computed efficiently for all $u, v \in \mathbb{G}$.

Definition 1 A bilinear parameter generator Gen is a probabilistic algorithm that takes a security parameter κ as input, and outputs a 5-tuple $(q, g, \mathbb{G}, \mathbb{G}_T, e)$, where q is a κ-bit prime number, $\mathbb{G}$ and $\mathbb{G}_T$ are two groups with order q, $g \in \mathbb{G}$ is a generator, and $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ is a non-degenerated and efficiently computable bilinear map.

3.2 Classifier of naive Bayes

Naive Bayes classifier is based on Bayes theorem which can be used to compute posterior probabilities given observations [17]. There are f classes which are denoted as $C = \{C_1, C_2, \ldots, C_f\}$. Let a n-dimension vector $\mathbf{x} = (x_1, \ldots, x_n)$ represent a instance and depict n measured values of the n attributes, $A_1, \ldots, A_n$. For clarity, the bold and regular symbols indicate vectors and scalar variables, respectively. The naive Bayes classifier can predict the vector $\mathbf{x}$ lies in the class $C_{j'}$ if and only if the posterior probability

$$P(C_{j'} \,|\mathbf{x}) > P(C_j \,|\mathbf{x}),$$

where $1 \leq j \leq f, j \neq j'$. The posterior probability $P(C_j \,|\mathbf{x}) = \frac{P(\mathbf{x}|C_j) \cdot P(C_j)}{P(\mathbf{x})}$ can be obtained by Bayes's theorem, where $P(C_j)$ is the prior probability of C_j. Since $P(\mathbf{x})$ is the same for all classes, $P(C_j \,|\mathbf{x}) \propto P(\mathbf{x} \,|C_j) \cdot P(C_j)$, i.e., only $P(\mathbf{x} \,|C_j) \cdot P(C_j)$ needs to be maximized. Moreover, a naive Bayes classifier assumes that all the features are conditionally independent of one another, mathematically meaning

$$P(\mathbf{x} \,|C_j) = \prod_{i=1}^{n} P(x_i \,|C_j),$$

where $P(x_1 \,|C_j)$, $P(x_2 \,|C_j), \ldots, P(x_n \,|C_j)$ can be easily estimated from the training set.

3.3 Improved naive Bayes classification standard

To predict the class label of $\mathbf{x} = (x_1, \ldots, x_n)$ by the naive Bayes model, the conditional probability $P(x_i \,|C_j)$ can be estimated as $P(x_i \,|C_j) = \frac{N_{x_i}^{(j)}}{N_j}$, where $N_{x_i}^{(j)}$ is the number of instances whose class labels are C_j and the values of attribute A_i are x_i, and N_j is the number of instances in class $C_j, i = 1, \ldots, n, j = 1, \ldots, f$. Meanwhile, the priori probability $P(C_j)$ can be computed as $P(C_j) = \frac{N_j}{N}$, where N is the total number of instances. Denote the class label of $\mathbf{x}$ by y, and y can be computed as $y = \arg\max_{C_j \in C}(P(C_j) \cdot P(\mathbf{x} \,|C_j)) = \arg\max_{C_j \in C}(P(C_j) \cdot \prod_{i=1}^{n} P(x_i \,|C_j))$. Thus, the naive Bayes classification standard can be described as

$$\begin{aligned} y &= \arg\max_{j \in \{1,\ldots,f\}} (P(C_j) \cdot \prod_{i=1}^{n} P(x_i \,|C_j)) \\ &= \arg\max_{j \in \{1,\ldots,f\}} (\frac{N_j}{N} \cdot \prod_{i=1}^{n} \frac{N_{x_i}^{(j)}}{N_j}) \\ &= \arg\max_{j \in \{1,\ldots,f\}} (\frac{1}{N \cdot N_j^{n-1}} \cdot \prod_{i=1}^{n} N_{x_i}^{(j)}). \end{aligned} \tag{1}$$

Since N is the same for all classes, Eq. (1) can be expressed as

$$y = \arg\max_{j \in \{1,\ldots,f\}} (\frac{1}{N_j^{n-1}} \cdot \prod_{i=1}^{n} N_{x_i}^{(j)}). \tag{2}$$

For convenience of calculations, we denote the lowest common multiple of $N_1^{n-1}, N_2^{n-1}, \ldots$ and N_f^{n-1} by lcm, i.e., $lcm = [N_1^{n-1}, N_2^{n-1}, \ldots, N_f^{n-1}]$, and thus Eq. (2) can be improved as

$$y = \arg\max_{j \in \{1,\ldots,f\}} (\frac{lcm}{N_j^{n-1}} \cdot \prod_{i=1}^{n} N_{x_i}^{(j)}). \tag{3}$$

4 Proposed PDiag scheme

In this section, we present our efficient and privacy-preserving online medical primary diagnosis scheme on

naive Bayesian classification which mainly consists of five phases: *system initialization*, *data preparation*, *query generation*, *privacy-preserving online medical primary diagnosis service*, *query result reading*. Specially, SP first provides registration for the user in the *system initialization* phase and does some statistical analysis in the *data preparation* phase. Then the user preprocesses query information by introducing different random numbers in the *query generation* phase. After that, SP operates the preprocessed query information with lightweight polynomial aggregation technique in the *privacy-preserving online medical primary diagnosis service* phase. Finally, the user obtains the final diagnosis result in the *query result reading* phase. To facilitate understanding, we describe the process as shown in Fig. 2. Meanwhile, for easier expression, we give the description of variables used in the following subsections in Table 1.

4.1 System initialization

SP first chooses security parameter κ to generate the bilinear parameters $(q, g, \mathbb{G}, \mathbb{G}_{\mathbb{T}}, e)$ by running $Gen(\kappa)$. Then, SP chooses a random number $s_{SP} \in \mathbb{Z}_q^*$ as its private key SK_{SP}, and computes its public key $PK_{SP} = g^{SK_{SP}}$. In addition, SP chooses a secure asymmetric encryption algorithm $E()$, i.e., ECC, and a secure cryptographic hash function $H()$, where $H : \{0,1\}^* \rightarrow \mathbb{Z}_q^*$. After that, SP keeps its private key SK_{SP} as master key secretly and publishes the system parameters $< q, g, \mathbb{G}, \mathbb{G}_{\mathbb{T}}, e, PK_{SP}, E(), H() >$.

When registering in the SP, user U_k chooses a random number $s_k \in \mathbb{Z}_q^*$ as her/his private key SK_{U_k}, and computes its corresponding public key $PK_{U_k} = g^{SK_{U_k}}$, then she/he submits her/his information and PK_{U_k} to SP through a secure channel for signature. SP first checks whether the user's information is correct or not. If it is correct, SP makes a signature for PK_{U_k} with its private key SK_{SP} and sends it back to U_k.

4.2 Data preparation

In general, SP has plenty of medical instances (we assume the number of medical instances is N). Each instance can be represented by $(\mathbf{x}^{(k)}, C_j)$ where $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$, C_j is the class label of $\mathbf{x}^{(k)}$, $j = 1, \ldots, f$, $k = 1, \ldots, N$. SP first checks all the instances and groups the instances by their class labels, i.e., all instances whose class labels are C_j belong to one group named group C_j. For convenience of calculations, each symptom $x_i^{(k)}, i = 1, \ldots, n$, is normalized into binary, i.e., $x_i^{(k)}$ should be converted to $(b_{i,1}, \ldots, b_{i,m_i})$ instead, where $x_i^{(k)}$ is numeric, $b_{i,1}, \ldots, b_{i,m_i}$ are binary, and m_i is the value range of

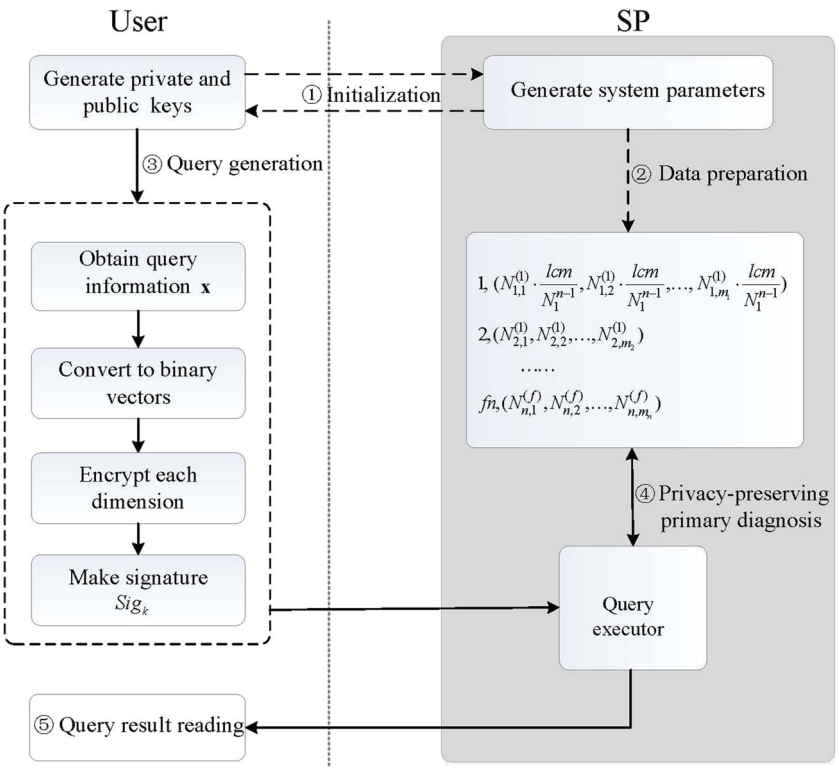

Fig. 2 The conceptual architecture of PDiag

Table 1 Variables and their descriptions in PDiag

Variables	Description
κ	the secure parameter chosen by SP
q, g	parameters of bilinear groups
$H()$	the secure cryptographic hash function
$E()$	the secure asymmetric encryption algorithm
$A_1, \ldots, A_n$	n symptom attributes
$C_1, \ldots, C_f$	f disease classes
N	the number of instances in the dataset
N_j	the number of instances in class C_j
$N_{x_i}^{(j)}$	the number of instances whose class labels are C_j and the values of attribute A_j are x_i
$\mathbf{x}$	the query vector of user
$\mathbf{x}^{(k)}$	an instance of the dataset
x_i^k	the value of attribute A_i for instance $\mathbf{x}^{(k)}$
m_i	the value range of attribute A_i
n	the dimension of query vector
$\mathbf{b}_i^{(j)}$	the statistics vector of class C_j
lcm	the lowest common multiple of $N_1^{n-1}, N_2^{n-1}, \ldots, N_f^{n-1}$
p, α	two big primes chosen by U_k
s, c_{i,m_t}	random numbers chosen by U_k
$r_i^{(j)}$	random numbers chosen by SP
k_1, k_2, k_3, k_4	secure parameters chosen by U_k

$x_i^{(k)}$. For example, the attribute age $x_i^{(k)}$ ranges from 1 to 130, which should be converted into $(b_{i,1}, \ldots, b_{i,130})$. If $x_i^{(k)} = 35$, then the converted attributes $b_{i,35} = 1$, and $b_{i,m_t} = 0$, where $m_t = 1, \ldots, 130$ and $m_t \neq 35$. Then SP does some statistical analysis of each b_{i,m_t}, where $i = 1, \ldots, n$, $m_t = 1, \ldots, m_i$. To reduce storage consumption, SP only stores the number of $b_{i,m_t} = 1$ in class C_j, which is denoted by $N_{i,m_t}^{(j)}$ and equals $N_{x_i}^{(j)}$. The final instance statistics information stored in SP is shown in Fig. 3.

4.3 Query generation

User U_k wants to query online medical primary diagnosis service through $\mathbf{x} = (x_1, \ldots, x_n) \in F_{q_1}^n$, which is sensitive and needed to be protected during the process of query. After registration, user U_k converts each $x_i, i = 1, \ldots, n$, to binary $(a_{i,1}, a_{i,2}, \ldots, a_{i,m_i})$ according to the corresponding range of attribute value and numerical interval. We define $q_1 \leq 2^{32}$, $n \leq 2^{32}$, some security parameters k_1, k_2, k_3, k_4, where $k_1 > {k_2}^2$, $k_2 \cdot k_3 < k_1$, $k_2 \cdot k_4 < k_1$ and $k_3 \cdot k_4 < {k_2}^2$. U_k chooses two large primes p and α such that $|p| = k_1$, $|\alpha| = k_2$. Then she/he sets $a_{i,m_i+1} = a_{i,m_i+2} = 0$ and chooses a large random number $s \in Z_p^*$, and executes the following operations for each a_{i,m_t},

$$W_{i,m_t} = \begin{cases} s(\alpha \cdot a_{i,m_t} + c_{i,m_t}) \bmod p, & if\ a_{i,m_t} \neq 0; \\ s \cdot c_{i,m_t} \bmod p, & if\ a_{i,m_t} = 0; \end{cases} \quad (4)$$

where each c_{i,m_t}, $i = 1, \ldots, n$, $m_t = 1, \ldots, m_i + 2$, is a random number chosen by U_k with $|c_{i,m_t}| = k_3$. U_k keeps $s^{-1} \bmod p$ secret, computes $Q = E_{PK_{SP}}(\alpha||p||W_{1,1}||W_{1,2}||\ldots||W_{1,m_1+2}||\ldots||W_{n,1}||W_{n,2}||\ldots||W_{n,m_n+2})$ and creates a signature $Sig_k = (H(Q||TS_1))^{SK_{U_k}}$ by using her/his private key SK_{U_K}, where TS_1 is the current time stamp, which can resist the potential replay attack. Finally, U_k sends $< Q||TS_1||Sig_k >$ to SP.

4.4 Privacy-preserving medical primary diagnosis service

After receiving $< Q||TS_1||Sig_k >$, SP first checks the time stamp TS_1 and the signature Sig_k to verify its validity, i.e., verify whether $e(g, Sig_k) = e(PK_{U_k}, H\ (Q||TS_1))$. If it does hold, the signature is accepted, since $e(g, Sig_k) = e(g, H\ (Q||TS_1)^{SK_{U_k}}) = e(PK_{U_k}, H(Q||TS_1))$. Then, SP decrypts Q with its private key SK_{SP} to obtain the encrypted query. To facilitate understanding, we denote statistics vectors of class C_j stored in SP by $\mathbf{b}_i^{(j)} = (b_{i,1}^{(j)}, b_{i,2}^{(j)}, \ldots, b_{i,m_i}^{(j)})$, $i = 1, \ldots, n$. For every $\mathbf{b}_i^{(j)}$, SP chooses a random number $r_i^{(j)}$ with $|r_i^{(j)}| = k_4$, sets

Class label	Statistics vector
C_1	$(N_{1,1}^{(1)} \cdot \frac{lcm}{N_1^{n-1}}, N_{1,2}^{(1)} \cdot \frac{lcm}{N_1^{n-1}}, \ldots, N_{1,m_1}^{(1)} \cdot \frac{lcm}{N_1^{n-1}})$ $(N_{2,1}^{(1)}, N_{2,2}^{(1)}, \ldots, N_{2,m_2}^{(1)})$ $\cdots$ $(N_{n,1}^{(1)}, N_{n,2}^{(1)}, \ldots, N_{n,m_n}^{(1)})$
C_2	$(N_{1,1}^{(2)} \cdot \frac{lcm}{N_2^{n-1}}, N_{1,2}^{(2)} \cdot \frac{lcm}{N_2^{n-1}}, \ldots, N_{1,m_1}^{(2)} \cdot \frac{lcm}{N_2^{n-1}})$ $(N_{2,1}^{(2)}, N_{2,2}^{(2)}, \ldots, N_{2,m_2}^{(2)})$ $\cdots$ $(N_{n,1}^{(2)}, N_{n,2}^{(2)}, \ldots, N_{n,m_n}^{(2)})$
$\cdots$	$\cdots$
C_f	$(N_{1,1}^{(f)} \cdot \frac{lcm}{N_f^{n-1}}, N_{1,2}^{(f)} \cdot \frac{lcm}{N_f^{n-1}}, \ldots, N_{1,m_1}^{(f)} \cdot \frac{lcm}{N_f^{n-1}})$ $(N_{2,1}^{(f)}, N_{2,2}^{(f)}, \ldots, N_{2,m_2}^{(f)})$ $\cdots$ $(N_{n,1}^{(f)}, N_{n,2}^{(f)}, \ldots, N_{n,m_n}^{(f)})$

Fig. 3 Statistics vectors stored in SP

$b_{i,m_i+1}^{(j)} = b_{i,m_i+2}^{(j)} = 0$, and executes the following operations.

$$D_{i,m_t}^{(j)} = \begin{cases} \alpha \cdot r_i^{(j)} \cdot b_{i,m_t}^{(j)} \cdot W_{i,m_t} \bmod p, & if\ b_{i,m_t}^{j} \neq 0; \\ r_i^{(j)} \cdot W_{i,m_t} \bmod p, & if\ b_{i,m_t}^{j} = 0; \end{cases} \quad (5)$$

where $m_t = 1, \ldots, m_i + 2$, and $\prod_{i=1}^{n} r_i^{(1)} = \prod_{i=1}^{n} r_i^{(2)} = \ldots = \prod_{i=1}^{n} r_i^{(f)} = A$, $i = 1, \ldots, n$, $j = 1, \ldots, f$. SP computes $D_i^{(j)} = \sum_{m_t=1}^{m_i+2} D_{i,m_t}^{(j)} \bmod p$ for each $\mathbf{b}_i^{(j)}$, $i = 1, \ldots, n$, $j = 1, \ldots, f$, as follows.

$$\begin{aligned} D_i^{(j)} &= \sum_{m_t=1}^{m_i+2} D_{i,m_t}^{(j)} \bmod p \\ &= \alpha sr_i^{(j)} \sum_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}\neq 0}} b_{i,m_t}^{(j)}(\alpha a_{i,m_t} + c_{i,m_t}) + sr_i^{(j)} \sum_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}= 0}} (\alpha a_{i,m_t} + c_{i,m_t}) \\ &\quad + \alpha sr_i^{(j)} \sum_{\substack{a_{i,m_t}= 0,\\ b_{i,m_t}^{(j)}\neq 0}} c_{i,m_t} b_{i,m_t}^{(j)} + sr_i^{(j)} \sum_{\substack{a_{i,m_t}= 0,\\ b_{i,m_t}^{(j)}= 0}} c_{i,m_t} \\ &= \alpha^2 sr_i^{(j)} \sum_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}\neq 0}} a_{i,m_t} b_{i,m_t}^{(j)} + \alpha sr_i^{(j)} \sum_{b_{i,m_t}^{(j)}\neq 0} b_{i,m_t}^{(j)} c_{i,m_t} \\ &\quad + sr_i^{(j)} \sum_{b_{i,m_t}^{(j)}= 0} c_{i,m_t} + \alpha sr_i^{(j)} \sum_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}= 0}} a_{i,m_t}. \end{aligned} \quad (6)$$

Then SP computes $R = E_{PK_{U_k}}(D_1^{(1)}||\ldots||D_n^{(1)}||D_1^{(2)}||\ldots||D_n^{(2)}||\ldots||D_1^{(f)}||\ldots||D_n^{(f)})$ and creates a signature $Sig_{SP} = (H(R||TS_2))^{SK_{SP}}$ by its private key SK_{SP}, then sends $< R||TS_2||Sig_{SP} >$ to U_k.

4.5 Query result reading

After receiving $< R||TS_2||Sig_{SP} >$, U_k first checks TS_2 and the signature Sig_{SP} to verify its validity, i.e., verify whether $e(g, Sig_{SP}) = e(PK_{SP}, H(R||TS_2))$. If it does hold, the signature is accepted, since $e(g, Sig_{SP}) = e(g, H(R||TS_2)^{SK_{SP}}) = e(PK_{SP}, H(R||TS_2))$. U_k decrypts R with her/his SK_{U_k} to obtain $D_1^{(1)}, \ldots, D_n^{(1)}$, $D_1^{(2)}, \ldots, D_n^{(2)}, \ldots, D_1^{(f)}, \ldots, D_n^{(f)}$. Then she/he computes $M_i^{(j)} = s^{-1} \cdot D_i^{(j)} \bmod p$, $i = 1, \ldots, n$, $j = 1, \ldots, f$. After that, U_k puts them into the formula

$$T = \arg\max_{j\in\{1,\ldots,f\}} \left(\prod_{i=1}^{n} \frac{M_i^{(j)} - (M_i^{(j)} \bmod \alpha^2)}{\alpha^2}\right) \quad (7)$$

If $T = j' \in \{1, \ldots, f\}$, U_k knows she/he more likely suffer from the disease marked by $C_{j'}$.

Correctness Traditional naive Bayes classifier judges which class the query vector lies in according to Eq. (1), while in Eq. (7), considering the above setting, i.e., $k_1 > k_2^2$, $k_2 \cdot k_3 < k_1$, $k_2 \cdot k_4 < k_1$ and $k_3 \cdot k_4 < k_2^2$, the components in Eq. (7) should meet the following constraints

$$\begin{cases} \alpha^2 r_i^{(j)} \sum\limits_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}\neq 0}} a_{i,m_t} b_{i,m_t}^{(j)} + \alpha r_i^{(j)} \sum\limits_{b_{i,m_t}^{(j)}\neq 0} b_{i,m_t}^{(j)} c_{i,m_t} + r_i^{(j)} \sum\limits_{b_{i,m_t}^{(j)}= 0} c_{i,m_t} \\ + \alpha r_i^{(j)} \sum\limits_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}= 0}} a_{i,m_t} < p, \\ \alpha r_i^{(j)} \sum\limits_{b_{i,m_t}^{(j)}\neq 0} b_{i,m_t}^{(j)} c_{i,m_t} + r_i^{(j)} \sum\limits_{b_{i,m_t}^{(j)}= 0} c_{i,m_t} + \alpha r_i^{(j)} \sum\limits_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}= 0}} a_{i,m_t} < \alpha^2. \end{cases} \quad (8)$$

Under the aforementioned constraints, T is implicitly formed by

$$\begin{aligned} T &= \arg\max_{j\in\{1,\ldots,f\}} \left(\prod_{i=1}^{n} \frac{M_i^{(j)} - (M_i^{(j)} \bmod \alpha^2)}{\alpha^2}\right) \\ &= \arg\max_{j\in\{1,\ldots,f\}} \left(\prod_{i=1}^{n} r_i^{(j)} \cdot \left(\sum_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}\neq 0}} a_{i,m_t} \cdot b_{i,m_t}^{(j)}\right)\right) \\ &= \arg\max_{j\in\{1,\ldots,f\}} \left(\prod_{i=1}^{n} r_i^{(j)} \cdot \prod_{i=1}^{n} \left(\sum_{\substack{a_{i,m_t}\neq 0,\\ b_{i,m_t}^{(j)}\neq 0}} a_{i,m_t} \cdot b_{i,m_t}^{(j)}\right)\right) \\ &= \arg\max_{j\in\{1,\ldots,f\}} \left(A \cdot \frac{lcm}{N_j^{n-1}} \cdot \prod_{i=1}^{n} N_{x_i}^{(j)}\right) \end{aligned} \quad (9)$$

Obviously, since A is the same for all classes, when $y = j'$, $T = j'$. Therefore, T in Eq. (9) equals y in Eq. (3), and the correctness of PDiag is verified.

5 Security analysis

In this section, we analyze security properties of the proposed PDiag scheme. Specifically, following security requirements discussed earlier, our analysis will focus on how PDiag can achieve the Bayes classifier confidentiality, as well as the user's query information privacy.

– *The user's query information is privacy-preserving in the proposed PDiag scheme.* During the query generation phase, for each a_{i,m_t}, $i = 1, 2, \ldots, n$, $m_t = 1, \ldots, m_i + 2$, we have $W_{i,m_t} = s(\alpha \cdot a_{i,m_t} + c_{i,m_t}) \bmod p$ when $a_{i,m_t} \neq 0$, and $W_{i,m_t} = sc_{i,m_t} \bmod p$ when $a_{i,m_t} = 0$. Therefore, each a_{i,m_t} is randomized by freshly generated random integers c_{i,m_t}, $i =$

$1, 2, \ldots, n$, $m_t = 1, \ldots, m_i + 2$. SP is curious to infer the user query information a_{i,m_t} from W_{i,m_t}. However, without knowing the freshly generated random numbers c_{i,m_t} and s, it is impossible to obtain a_{i,m_t}. Since the random number c_{i,m_t} is individually used for once, different c_{i,m_t} are unlinkable, and a_{i,m_t}, c_{i,m_t}, s are only known by the registered user, SP cannot obtain the user's query information according to her/his query. As described in information theory, as long as the size of user's data is less than the random integer, and the random integers are fresh, SP can only compute identical priori and posteriori probabilities [18]. Moreover, the user's encrypted data query is encrypted again by SP's public key PK_{SP} before being sent to SP, then only SP can obtain the encrypted data query. Therefore, the user's query information $\mathbf{x} = (x_1, \ldots, x_n)$ is privacy-preserving during the computation.

– *The proposed PDiag scheme can also achieve confidentiality of the Bayes classifier.* In the proposed PDiag scheme, SP stores the statistics information of all instances, which is considered as its own asset and should be kept privately. For every statistics vector $\mathbf{b}_i^{(j)} = (b_{i,1}^{(j)}, b_{i,2}^{(j)}, \ldots, b_{i,m_i}^{(j)})$, SP sets $b_{i,m_i+1}^{(j)} = b_{i,m_i+2}^{(j)} = 0$ to ensure that at least two random numbers are included in $D_i^{(j)}$, which prevents the user from obtaining $\mathbf{b}_i^{(j)} = (b_{i,1}^{(j)}, b_{i,2}^{(j)}, \ldots, b_{i,m_i}^{(j)})$. Then SP computes $D_{i,m_t}^{(j)}$ according to Eq. (5). Therefore, each $D_{i,m_t}^{(j)}$ is randomized by a random number $r_i^{(j)}$. When all $D_{i,m_t}^{(j)}$ are summated into $D_i^{(j)}$, $b_{i,m_t}^{(j)}$ will hide the operation in it. After the user receives $D_1^{(1)}, \ldots, D_n^{(1)}, D_1^{(2)}, \ldots, D_n^{(2)} \ldots, D_1^{(f)}, \ldots, D_n^{(f)}$, she/he can compute $M_i^{(j)} = s^{-1} \cdot D_i^{(j)} \bmod p$, $i = 1, \ldots, n$, $j = 1, \ldots, f$. Then the user can obtain $\frac{M_i^{(j)} - (M_i^{(j)} \bmod \alpha^2)}{\alpha^2} = r_i^{(j)} \sum_{\substack{a_{i,m_t} \neq 0, \\ b_{i,m_t}^{(j)} \neq o}} a_{i,m_t} \cdot b_{i,m_t}^{(j)}$. As explained previously, since the the size of random integer $r_i^{(j)}$ is greater than the size of $\sum_{\substack{a_{i,m_t} \neq 0, \\ b_{i,m_t}^{(j)} \neq 0}} a_{i,m_t} \cdot b_{i,m_t}^{(j)}$, SP ensures the security of statistics vectors information-theoretically, that is, without knowing freshly generated random numbers $r_i^{(j)}$, $i = 1, \ldots, n$, $j = 1, \ldots, f$, the user can't extract the statistics vector $\mathbf{b}_i^{(j)}$ according to $D_i^{(j)}$. On the other hand, owing to the constraints in Eq. (7), T equals y, i.e., when $y = j'$, $T = j'$. Meanwhile, the user can achieve T according to Eq. (8), but cannot obtain the exact value of $\frac{lcm}{N_j^{n-1}} \cdot \prod_{i=1}^{n} N_{x_i}^{(j)}$, $j = 1, \ldots, f$ without knowing A. Moreover, SP's response is encrypted again by the user's public key before being sent to the user, then only the right user can obtain the encrypted response. Therefore, the components of Bayes classifier are also privacy-preserving during the computation.

– *The authentication of data query request is achieved in the proposed PDiag scheme.* In the proposed eDiag scheme, each registered user's request is signed by Boneh-Lynn-Shacham (BLS) short signature [19]. Since the BLS short signature is provably secure under the computational Diffie-Hellman problem in the random oracle model, the source authentication can be guaranteed. Moreover, for any unregistered user, since she/he doesn't have the secret key, she/he also cannot submit a valid query request to SP. As a result, the query request from the unregistered user can be detected in the proposed PDiag scheme.

From the given analysis, we can conclude that the proposed PDiag scheme is secure and privacy-preserving for the user as well as SP, and can achieve our security goal.

6 Performance evaluation

In this section, we first evaluate the performance of PDiag in terms of accuracy and computational complexity. Then, we implement PDiag and deploy it in real environment to evaluate its integrated performance.

6.1 Evaluation environment

In order to measure the integrated performance of PDiag in real environment, we implement PDiag on smartphone and computer with different datasets. Specifically, a smartphone with 1.2 GHz, 2GB RAM, Android 4.4.2, and a computer with 2.9 GHz, 4GB RAM, Windows 7, are chosen to evaluate user and SP respectively, which are connected through 802.11g WLAN. Based on PDiag scheme, an application built in Java, named PDiag.apk, is installed in the smartphone, and the simulator for SP is deployed in the computer. Users who registered in SP can obtain online medical primary diagnosis by PDiag.apk. In particular, when the user inputs the medical data by PDiag.apk, the smartphone will send a query request to the computer and get the response through WLAN. To obtain the correct classification result under the above setting, we can just set $k_1 = 512$, $k_2 = 200$, $k_3 = 128$, and $k_4 = 64$. In addition, we consider two real datasets to evaluate the accuracy of our proposed scheme. They are from the UCI machine learning repository called the Wisconsin Breast Cancer (WBC) [20] and Heart Disease (HD) datasets [21].

6.2 Accuracy evaluation

The WBC dataset contains 699 instances where 241 instances are malignant and 458 instances are benign, while the HD dataset contains 270 instances where 150 instances are absence of heart disease and 120 instances are presence of heart disease. The number of features used for training the naive Bayes classifier for each instance in WBC and HD datasets are nine and ten, respectively (excluding class label attribute). Before training the classifiers, each of the features can be normalized into a binary vector. Then we use the two datasets to train two different naive Bayes classifiers. Meanwhile, we choose 200 instances where 100 instances are malignant and 100 instances are benign from the WBC dataset, and 200 instances where 100 instances are absence of heart disease and 100 instances are presence of heart disease from the HD dataset. The chosen instances are used for testing the success rate of the two classifiers. From Table 2 we can see that the total number of correctly classified malignant instances in WBC dataset is 94 out of 100 (94 %) and that of benign instances is 92 out of 100 (92 %). In total, 200 samples are correctly classified out of 186(93 %). Similarly, the total number of correctly classified instances in HD dataset is 194 out of 200 (97 %). Obviously, our privacy-preserving algorithm can achieve a high accuracy.

6.3 Computation complexity

The proposed PDiag scheme can offer efficient online medical primary diagnosis service to medical users. Specifically, we assume the average dimension of statistic vectors is k, the number of symptom attributes is n and the number of disease classes is f. When the user generates the encrypted information $W_{1,1}, W_{1,2}, \ldots, W_{1,m_1+2}, \ldots, W_{n,1}, W_{n,2}, \ldots W_{n,m_n+2}$, it requires $n(2k+2)$ multiplication operations for $\mathbf{x} = (x_1, \ldots, x_n)$. After receiving the ciphertext from the user, it will cost SP $fn(3k+2)$ multiplication operations to generate $D_1^{(1)}, \ldots, D_n^{(1)}, D_1^{(2)}, \ldots, D_n^{(2)}, \ldots, D_1^{(f)}, \ldots, D_n^{(f)}$. To obtain the final diagnosis result, it will cost the user $(3fn - f)$ multiplication operations. Denote the computational costs of an exponentiation operation and a multiplication operation by C_e and C_m, respectively. Then totally for the user and SP, the computational cost will be $(2kn + 2n + 3fn - f) * C_m$ and $(3fkn + 2fn) * C_m$ in PDiag.

Different from many of time-consuming fully and partially homomorphic encryption techniques, the proposed PDiag uses lightweight polynomial aggregation technique, which can provide efficient online medical primary diagnosis service for users while preserves the privacy of medical users' data and the Bayes classifier efficiently with low overhead in computation. In the following, for the comparison with PDiag, we selected a clinical decision support system [13], which we call CDSS in the rest of paper for the sake of simplicity, and it also preserves the privacy of patient data as well as the Bayes classifier by using Paillier encryption technique and secure multiplication protocol. We assume the number of symptom attributes is n and the number of disease classes is f too. And the corresponding computational costs of the user and SP are $(4fkn + 2kn - 4) * C_e + (5fkn + nk) * C_m$ and $(11fkn - 7) * C_e + (7fkn - 5) * C_m$, respectively, in CDSS.

We present the computation complexity comparison of PDiag and CDSS in Table 3. It is obvious that our proposed PDiag scheme can achieve efficient medical diagnosis with low computation complexity in the user and SP. To further demonstrate the advantage of the proposed PDiag over CDSS, we evaluate the computation overhead in the environment described in Section 6.1. Figures 4 and 5 depict the computation overhead varying with the number of disease classes in user and SP, and we assume the dimension of query vector is 20 and the average dimension of statistic vectors is 100. Comparing Figs. 4 and 5, we can find that with the increase of the numbers of disease classes, the computation overhead of CDSS is much higher than that of our proposed PDiag scheme. Although the computation overhead of our proposed PDiag scheme also increases when the number of disease classes is large, it is still much lower than that of CDSS. In addition, the user needs to interact with SP many times in CDSS and only once in PDiag to achieve the final diagnosis result. In conclusion, our proposed PDiag scheme can achieve better efficiency in terms of computation overhead in user and SP.

Table 2 Accuracy of PDiag

Accuracy	WBC	HD
Yes(100)	94(94 %)	96(96 %)
No(100)	92(92 %)	98(98 %)
Overall(200)	186(93.0 %)	194(97.0 %)

6.4 Efficiency evaluation

In order to test the factors that may affect the efficiency of our proposed PDiag, different naive Bayes classifiers are randomly generated. We evaluate the computation complexity of our proposed PDiag in the user and SP, respectively.

6.4.1 SP

In our proposed PDiag scheme, the factors which may impact the computation complexity in SP are the dimension of the query vector and the number of diseases to be

Table 3 Comparison of Computation Complexity

	PDiag	CDSS
User	$(2kn + 2n + 3fn - f) * C_m$	$(4fkn + 2kn - 4)*C_e + (5fkn + nk)*C_m$
SP	$(3fkn + 2fn) * C_m$	$(11fkn - 7)*C_e + (7fkn - 5)*C_m$

classified. Therefore, different dimensions of query vectors and different numbers of diseases are chosen to illustrate the computation cost of SP. The dimensions of query vectors are selected from 5 to 30, and the numbers of diseases are chosen from 2 to 12. Then, we execute 1000 times with different dimensions and numbers. As shown in Fig. 6, we can learn that the computation overhead of SP increases with the increasing of query vectors' dimension and diseases' number. The reason is that, when SP intends to offer online medical primary diagnosis service to the user, statistics vectors will be operated to compute $D_1^{(1)}, \ldots, D_n^{(1)}, D_1^{(2)}, \ldots, D_n^{(2)}, \ldots, D_1^{(f)}, \ldots, D_n^{(f)}$, which will cost more time with the increase of diseases' number and query vectors' dimension. But due to the fact that basic operations are based on lightweight polynomial aggregation technique, which are very quick in speed, the maximum time required for SP is less than 80 milliseconds.

6.4.2 The user

The query response time of user (i.e. smartphone) is an important result illustrating the feasibility of our proposed PDiag scheme. Therefore, different dimensions of data queries are chosen to illustrate the computation cost of smartphone. To observe the computation cost of smartphone, the dimensions of query vectors are selected from 5 to 30, and the numbers of diseases are chosen from 2 to 12. Then, we execute 1000 times with different dimensions and numbers. Specifically, Fig. 7 shows the computation overhead of smartphone increases with the increasing of query vectors' dimension and diseases' number. The reason is that smartphone needs to compute more encrypted parameters with increase of query vectors' dimension and diseases' number. Similarly, due to the fact that basic operations are based on lightweight polynomial aggregation technique, which are very quick in speed, the maximum time required for user is less than 2 seconds.

6.4.3 Integrated performance in real environment

In order to evaluate the integrated performance of our proposed PDiag scheme, the PDiag scheme is deployed in real environment. The integrated performance is measured by the overhead in computation and communication, i.e., the average response time. The dimensions of query vectors are selected from 5 to 30, and the number of diseases is fixed at 6. Then we execute 1000 times with different dimensions. In particular, Fig. 8 shows the average response time of PDiag increases with the increasing of query vectors' dimension. We can find that the entire overhead for once whole privacy-preserving online medical primary diagnosis service query is consistent with the results in the simulation environment, and all the query response time is less than 2 seconds in real environment.

From the above analysis, our proposed PDiag scheme is indeed efficient in terms of computation and

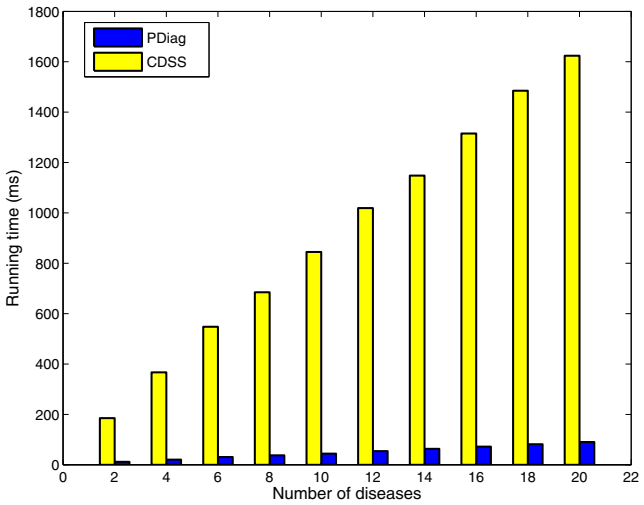

Fig. 4 Average running time of SP in PDiag and CDSS

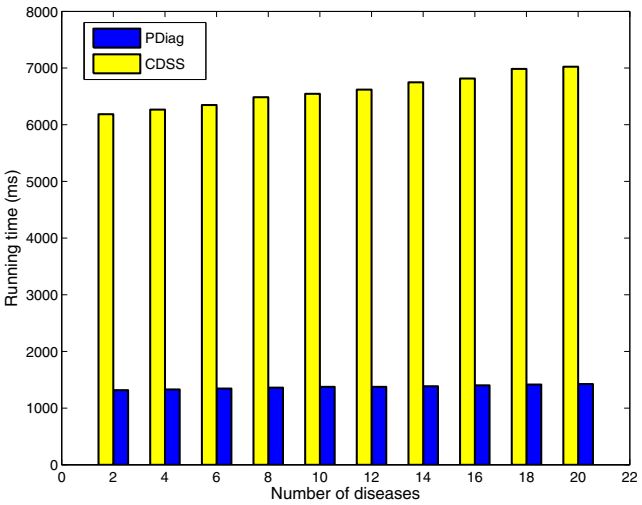

Fig. 5 Average running time of user in PDiag and CDSS

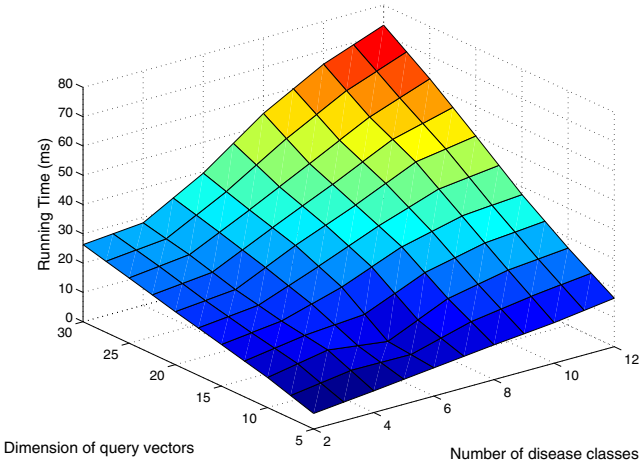

Fig. 6 Computation cost of SP in PDiag

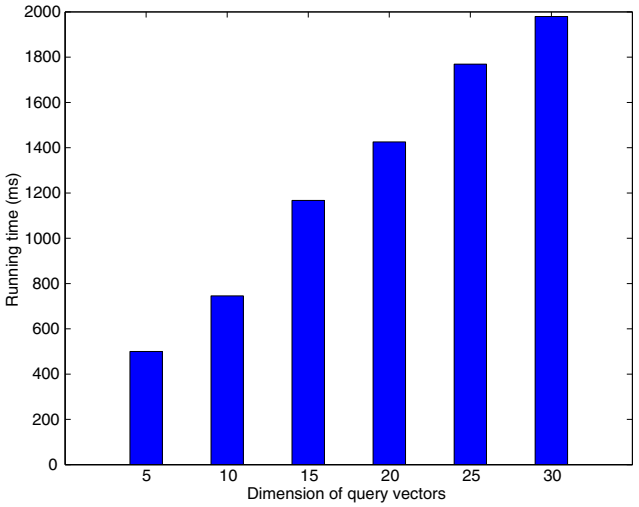

Fig. 8 Query response time in real environment

communication cost, which is suitable for providing online medical primary diagnosis service in real environment.

7 Related works

In this section, we briefly discuss some related works on medical diagnosis and privacy-preserving naive Bayes algorithm.

As the evolution of machine learning techniques, various diseases prediction models were built in biomedical engineering [22–30]. By training a Naive Bayes classifier in MR images, Zhou et al. [22] proposed a new approach to improve the brain diagnosis accuracy. Ajemba et al. [24] developed a fast predictive tool to predict the risk of progression of adolescent idiopathic scoliosis by employing a support vector classifier approach. In order to diagnose pancreatic cancer, Wang et al. [25] presented a risk prediction model by using Bayesian classification. Moreover, many naive Bayes prediction models were built to predict whether a patient has the heart disease or not and showed a good performance in accuracy [28–30]. However, all these medical diagnosis schemes are all in the plain domain, and they can't be used for providing disease prediction service via the Internet due to the fact that privacy has become a major concern.

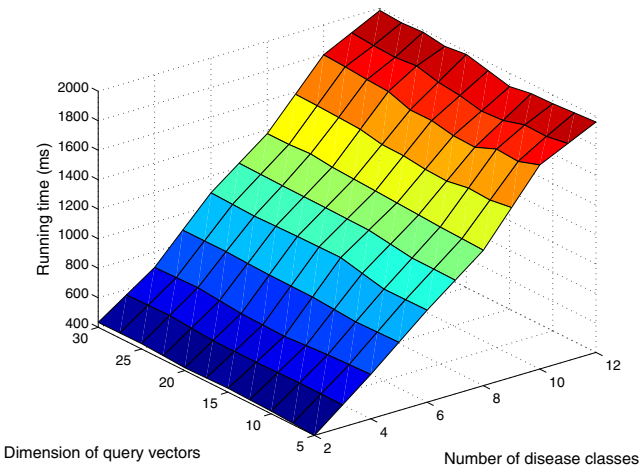

Fig. 7 Computation cost of user in PDiag

Aiming at the privacy concern, Mathew and Obradovic [31] proposed a privacy-preserving scheme for constructing a useful clinical tool in the form of a decision tree, which can be used for diagnosing disease, while it can only protect the privacy of the training dataset. Bos et al. [12] presented a working implementation of a prediction service to diagnose the likelihood of contracting a disease by using logistic regression and the Cox proportional hazard model. While in their setting, the predictive model is publicly known, and their proposed scheme can only protect the patient's information. Liu et al. [13] proposed a privacy-preserving clinical diagnosis system using naive Bayesian classifier, which can also help clinician complementary to diagnose the risk of patients' disease in a privacy-preserving way. Similarly, a privacy-preserving system using the support vector machine was proposed by Rahulamathavan et al. [14], and it can help to diagnose the patients without compromising the privacy of patients and third part. Since all the encrypted operations are based on homomorphic encryption technique, their efficiencies are not very high.

In the following, we detail the works of privacy-preserving naive Bayes algorithm. When it comes to constructing the global Bayes model without revealing their private databases, many existing approaches are considered instances of a secure multi-party problem. To build a privacy-preserving naive Bayes classifier on horizontally partitioned data, Kantarcioglu and Vaidya [32] first

used secure summation and logarithm method to make it, but the proposed protocol is vulnerable to collusion and eavesdropping attack. Then Yang et al. [33] proposed a privacy-preserving classification protocol by using additive homomorphic property of a modified version of ElGamal. Afterwards, Yi et al. [34] improved Kantarcioglu-Vaidya protocol [32] in terms of efficiency and security by using Paillier cryptosystem. Similarly, Sumana et al. [35] designed an improved privacy-preserving distributed naive Bayesian classifier by using the homorphic property of Paillier. To produce a a privacy preserving naive Bayes classifier without using a trusted third party, Gangrade et al. [36] proposed a three layer protocol by introducing a un-trusted third party, but the assumption that the communication networks used by the input parties to communicate with the UTP are secure seems rather restrictive. As for as data are vertically partitioned, Vaidya et al. [37] developed a privacy-preserving naive Bayes classifier on vertically partitioned data by using homomorphic public key encryption system. Then Keshavamurthy and Toshniwal [38] constructed a global classification model by using naive Bayes classification, which addressed various fragmentation issues such as horizontal, vertical and arbitrary distribution require format, while the need of a trusted third party seems rather restrictive. To resist both collusion and eavesdropping attacks during the distributed privacy-preserving of naive Bayes learning, Huai et al. [39] constructed differentially private protocols where data are either horizontally or vertically partitioned. However, they cannot be used for the user and service provider scenario considered in this paper. Meanwhile, most of homomorphic encryption schemes require massive resource-consuming computation, which makes them not quite suitable for providing efficient classification service via the Internet.

Different from all of the above works, our proposed PDiag scheme aims at the efficiency and privacy issues, and based on lightweight polynomial aggregation technique, we develop an efficient privacy-preserving online medical primary diagnosis scheme on naive Bayes classification. In particular, our proposed PDiag scheme can protect users' medical data privacy as well as ensure the confidentiality of diagnosis model, and can be easily implemented in smartphone and computer because of its high efficiency.

8 Conclusion

In this paper, we have proposed an efficient and privacy-preserving online medical primary diagnosis scheme, called PDiag, on naive Bayes classification. Based on an improved expression for naive Bayes classification, PDiag is introduced with lightweight polynomial aggregation technique. With PDiag, users' medical data privacy and the confidentiality of naive Bayes classifier can be protected with low overhead in computation and communication. Specifically, for a data query from a registered user, the response is directly performed on ciphertext at the service provider without decryption, and the diagnosis result can also only be decrypted by the registered user. Meanwhile, this scheme can achieve a high accuracy of disease prediction. Thus, the user can get efficient online medical primary diagnosis service without compromising privacy and accuracy. Detailed security analysis shows its security strength and privacy-preserving ability, and extensive experiments are conducted to demonstrate its efficiency.

9 Availability

The implementation of the proposed PDiag scheme and relevant information can be downloaded at http://xdzhuhui.com/demo/PDiag.

Acknowledgments This work was financially supported by the National Natural Science Foundation of China under Grant 61303218, Grant 6167241 and Grant U1401251, National Key Research and Development Program of China under Grant 2016YFB0800804, Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2016JM6007, Research Foundations for the Central Universities of China under Grant JB161507, and China 111 Project under Grant B16037. We would like to thank the anonymous reviewers for their insightful comments and suggestions.

References

1. Mattio R (2014) Shortest average wait time for doctors in major cities increased one minute year over year. http://www.businesswire.com/news/home/20140326005955/en/Shortest-Average-Wait-Time-Doctors-Major-Cities
2. news B (2016) Waiting lists: Increase in number for ni outpatient appointments. [Online]. Available: http://www.bbc.com/news/uk-northern-ireland-35661496
3. Messenger S (2016) Breast cancer patient waits in wales shocking. [Online]. Available: http://www.bbc.com/news/uk-wales-35778888
4. Chenguang H, Xiaomao F, Ye L (2013) Toward ubiquitous healthcare services with a novel efficient cloud platform. IEEE transactions on bio-medical engineering 60(1):230–234
5. Anderson MP, Dubnicka SR (2014) A sequential naïve bayes classifier for dna barcodes. Stat Appl Genet Mol Biol 13(4):423–434
6. Bellazzi R, Zupan B (2008) Predictive data mining in clinical medicine: Current issues and guidelines. Int J Med Inform 77(2):81–97
7. Blanco R, Inza I, Merino M, Quiroga J, Larrañaga P. (2005) Feature selection in bayesian classifiers for the prognosis of survival of cirrhotic patients treated with tips. J Biomed Inform 38(5):376–388
8. Ko EJ, Lee HJ, Lee JW (2007) Ontology-based context modeling and reasoning for u-healthcare. IEICE Trans Inf Syst 90(8):1262–1270
9. Lu R, Lin X, Shen X (2013) Spoc: A secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency. IEEE Trans Parallel Distrib Syst 24(3):614–624

10. Zhu H, Lu R, Huang C, Chen L, Li H (2015) An efficient privacy-preserving location based services query scheme in outsourced cloud. IEEE Trans Veh Technol PP(99):1–1. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7327242
11. Lu R, Zhu H, Liu X, Liu J, Shao J (2014) Toward efficient and privacy-preserving computing in big data era. IEEE Netw 28(4):46–50
12. Bos JW, Lauter K, Naehrig M (2014) Private predictive analysis on encrypted medical data. J Biomed Inform 50:234–243
13. Liu X, Lu R, Ma J, Chen L, Qin B (2015) Privacy-preserving patient-centric clinical decision support system on naive bayesian classification. IEEE Journal of Biomedical and Health Informatics 99:1–1
14. Rahulamathavan Y, Veluru S, Phan R-W, Chambers J, Rajarajan M (2014) Privacy-preserving clinical decision support system using gaussian kernel-based classification. IEEE Journal of Biomedical and Health Informatics 18(1):56–66
15. Boneh D, Franklin MK (2001) Identity-based encryption from the weil pairing. In: Proceedings of the 21st Annual International Cryptology Conference on Advances in Cryptology, ser CRYPTO '01. Springer-Verlag, London, UK, pp 213–229
16. Leung KM (2007) Naive bayesian classifier, Polytechnic University Department of Computer Science/Finance and Risk Engineering
17. Ren J, Lee SD, Chen X, Kao B, Cheng R, Cheung D (2009) Naive bayes classification of uncertain data. In: Data Mining, 2009. ICDM'09 Ninth IEEE International Conference on. IEEE, pp 944–949
18. Rahulamathavan Y, Rajarajan M (2015) Efficient privacy-preserving facial expression classification. IEEE Trans Dependable Secure Comput 7516:1
19. Boneh D, Shacham H (2001) Short signatures from the weil pairing. In: Advances in Cryptology 2001. Springer, pp 514–532
20. Wolberg DWH (1995) UCI machine learning repository. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+
21. To GB, Brown G, To GB, Brown G (2004) Diversity in neural network ensembles. University of Birmingham
22. Zhou X, Wang S, Xu W, Ji G, Phillips P, Sun P, Zhang Y (2015) Detection of pathological brain in mri scanning based on wavelet-entropy and naive bayes classifier. In: Bioinformatics and biomedical engineering. Springer, pp 201–209
23. Güler I, Beyli EDÜ (2007) Multiclass support vector machines for eeg-signals classification. IEEE Trans Inf Technol Biomed 11(2):117–126
24. Ajemba P, Ramirez L, Durdle N, Hill D, Raso V (2005) A support vectors classifier approach to predicting the risk of progression of adolescent idiopathic scoliosis. IEEE Trans Inf Technol Biomed 9(2):276–282
25. Wang W, Chen S, Brune KA, Hruban RH, Parmigiani G, Klein AP (2007) Pancpro: risk assessment for individuals with a family history of pancreatic cancer. J Clin Oncol 25(11):1417–1422
26. Barakat MNH, Bradley AP (2010) Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed 14(4):1114–1120
27. Huang C-L, Liao H-C, Chen M-C (2008) Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Systems with Applications 34(1):578–587
28. Sundar NA, Latha PP, Chandra MR (2012) Performance analysis of classification data mining techniques over heart disease database. IJESAT] International Journal of engineering science & advanced technology ISSN:2250–3676
29. Pattekari SA, Parveen A (2012) Prediction system for heart disease using naïve bayes. International Journal of Advanced Computer and Mathematical Sciences 3(3):290–294
30. Medhekar DS, Bote MP, Deshmukh SD (2013) Heart disease prediction system using naive bayes. Int J Enhanced Res Sci Technol Eng 3:2
31. Mathew G, Obradovic Z (2011) A privacy-preserving framework for distributed clinical decision support. In: Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on IEEE, pp 129–134
32. Kantarcıoglu M, Vaidya J, Clifton C (2003) Privacy preserving naive bayes classifier for horizontally partitioned data. In: IEEE ICDM workshop on privacy preserving data mining, pp 3–9
33. Yang Z, Zhong S, Wright RN (2005) Privacy-preserving classification of customer data without loss of accuracy. In: SDM. SIAM, pp 92–102
34. Yi X, Zhang Y (2009) Privacy-preserving naive bayes classification on distributed data via semi-trusted mixers. Inf Syst 34(3):371–380
35. Sumana M, Hareesha KS (2014) Privacy preserving naive bayes classifier for horizontally partitioned data using secure division. International Journal of Network Security and Its Applications 6:6
36. Gangrade A, Patel R (2012) Privacy preserving naïve bayes classifier for horizontally distribution scenario using un-trusted third party. IOSR Journal of Computer Engineering (IOSRJCE) ISSN:2278–0661
37. Vaidya J, Clifton C (2004) Privacy preserving naïve bayes classifier for vertically partitioned data. In: SDM. SIAM, pp 522–526
38. Toshniwal D (2011) Privacy preserving naïve bayes classification using trusted third party computation over distributed progressive databases. Advances in Computer Science and Information Technology:24–32
39. Huai M, Huang L, Yang W, Li L, Qi M (2015) Privacy-Preserving Naive Bayes Classification. Springer International Publishing

Xiaoxia Liu received the B.Sc. degree from Henan Agricultural University, Zhengzhou, China, in 2014. She is current working toward the Master's degree with the School of Telecommunications Engineering, Xidian University. Her research interests are in the areas of applied cryptography, cyber security and privacy.

Hui Zhu (M'13) received his B.Sc. degree from Xidian University in 2003, M.Sc. degree from Wuhan University in 2005, and Ph.D. degrees from Xidian University in 2009. In 2013, he was with School of Electrical and Electronics Engineering, Nanyang Technological University as a research fellow.

Since 2014, he has been with the school of Cyber Engineering, Xidian University, China, as an associate professor. His research interests are in the areas of applied cryptography, cyber security and privacy.

Rongxing Lu (S'09-M'11-SM'15) received the Ph.D degree in computer science from Shanghai Jiao Tong University, Shanghai, China in 2006 and the Ph.D. degree (awarded Canada Governor General Gold Medal) in electrical and computer engineering from the University of Waterloo, Waterloo, Ontario, Canada, in 2012.

Since 2013, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, as an Assistant Professor. His research interests include computer, network and communication security, applied cryptography, security and privacy analysis for vehicular network, e-Healthcare system, and smart grid communications.

Dr. Lu received the IEEE Communications Society Asia-Pacific Outstanding Young Researcher Award in 2013 and the Canada Governor General Gold Medal.

Hui Li (M'10) Received his B.Sc. degree from Fudan University in 1990, M.Sc. and Ph.D. degrees from Xidian University in 1993 and 1998, respectively. Since 2005, he has been the professor in the school of Telecommunication Engineering, Xidian University, China. His research interests are in the areas of cryptography, wireless network security, information theory and network coding.

Dr. Li served as a Technical Program Committee Cochair of the 2009 International Conference on Information Security Practice and Experience and the 2009 IEEE Industrial Applications Society Meeting and as the General Cochair for the 2010 International ICST Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia; the 2011 Provable Security Conference; and the 2011 International Supercomputing Conference.