Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 827371, 10 pages
http://dx.doi.org/10.1155/2014/827371

Research Article

Privacy Preserving RBF Kernel Support Vector Machine

Haoran Li,[1] Li Xiong,[1] Lucila Ohno-Machado,[2] and Xiaoqian Jiang[2]

[1] *Department of Mathematics & Computer Science, Emory University, Atlanta, GA 30322, USA*
[2] *Division of Biomedical Informatics, UC San Diego, La Jolla, CA 92093, USA*

Correspondence should be addressed to Haoran Li; hli57@emory.edu

Received 16 February 2014; Accepted 8 April 2014; Published 12 June 2014

Academic Editor: Bairong Shen

Data sharing is challenging but important for healthcare research. Methods for privacy-preserving data dissemination based on the rigorous differential privacy standard have been developed but they did not consider the characteristics of biomedical data and make full use of the available information. This often results in too much noise in the final outputs. We hypothesized that this situation can be alleviated by leveraging a small portion of open-consented data to improve utility without sacrificing privacy. We developed a hybrid privacy-preserving differentially private support vector machine (SVM) model that uses public data and private data together. Our model leverages the RBF kernel and can handle nonlinearly separable cases. Experiments showed that this approach outperforms two baselines: (1) SVMs that only use public data, and (2) differentially private SVMs that are built from private data. Our method demonstrated very close performance metrics compared to nonprivate SVMs trained on the private data.

1. Introduction

Data sharing is important for accelerating scientific discoveries, especially when there are not enough local samples to test a hypothesis [1, 2]. However, medical data are sensitive as they essentially contain personal information and can reveal much about ethnicity, disease risk [3], and even family surnames [4]. To promote data sharing, it is important to develop privacy-preserving algorithms that respect data confidentiality and present data utility [5], especially when one wants to leverage cloud computing [6].

Privacy preserving data analysis and publishing [7, 8] have received considerable attention in recent years as a promising approach for sharing information while preserving data privacy. Differential privacy [9–11] has recently emerged as one of the strongest privacy guarantees for statistical data release [12–17]. A statistical aggregation or computation is DP (we shorten differentially private to DP) if the outcome is formally indistinguishable when run with and without any particular record in the dataset. The level of indistinguishability is quantified as a privacy parameter ϵ. A common mechanism to achieve differential privacy is the Laplace mechanism [18] which injects calibrated noise to a statistical measure determined by the privacy parameter ϵ and the sensitivity of the statistical measure influenced by the inclusion and exclusion of a record in the dataset. A lower privacy parameter requires larger noise to be added and provides a higher level of privacy.

General purpose algorithms for privacy protection (e.g., [19, 20]) often introduce too much perturbation error, which renders the resulting information useless for healthcare research. Our contribution is to leverage a small portion of open-consented data to maximally explore information that resides in the private data through a hybrid framework. Figure 1 shows an example of an environment in this case. We recently published differentially private distributed logistic regression using public and private biomedical datasets [21], which demonstrated advantages over pure private or public models. However, logistic regression is a generalized linear model, which has limited flexibility in classifying complex patterns. In this paper, we sought to extend our previous effort to the more powerful, RBF-kernel based support vector machines.

The remainder of the paper is organized as follows. Section 2 reviews background knowledge of differential privacy and SVM and RBF kernel. Section 3 describes the framework and details for our hybrid SVM mechanism. Then, Section 4 contains an extensive set of experimental

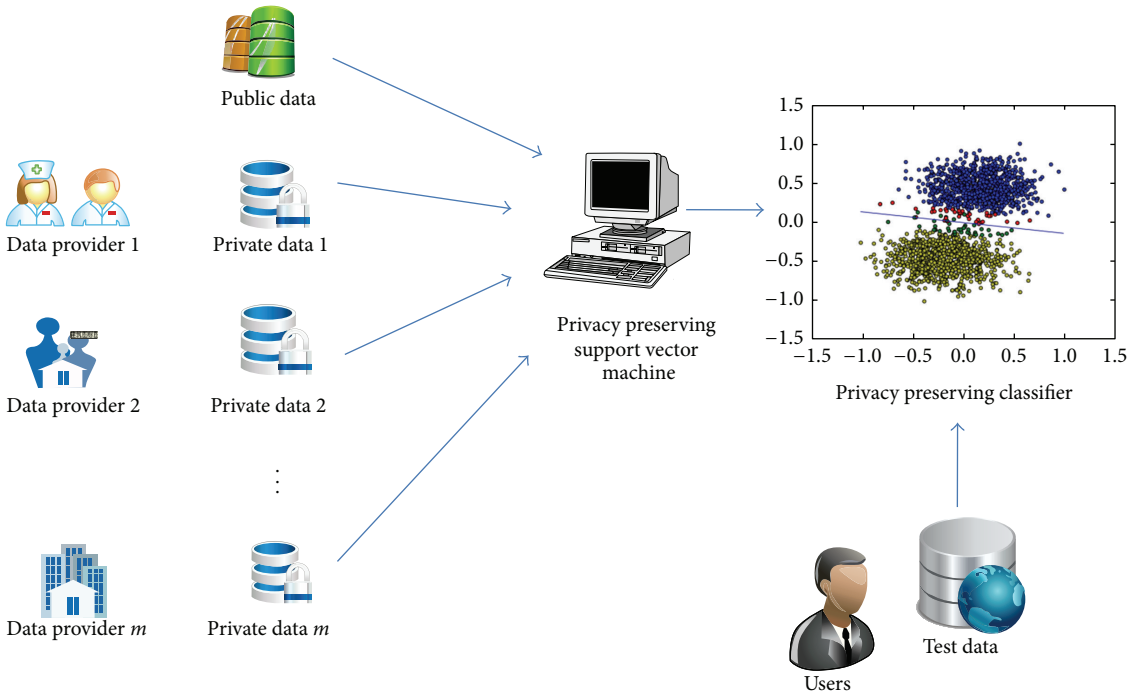


Figure 1: Biomedicine data sharing system. A small amount of public data and a large amount of private data are available for different data providers. A privacy preserving support vector machine can leverage both public and private data to maximize the classification accuracy under differential privacy. Then users can classify their test data via the released privacy preserving classifier.

evaluations. Finally, Section 5 concludes the paper with conclusions, limitations, and directions for future work.

2. Related Work

Rubinstein et al. [22] propose a private kernel SVM algorithm (shortened as PrivateSVM) which only works for a translation-invariant kernel $g(\Delta)$. The method approximates the original infinite feature space Ω of $g(\Delta)$ with a finite feature space $\widetilde{\Omega}$ using the Fourier transform $p(\omega)$ of $g(\Delta)$. Then add the noise to the weight parameters in the primal form based on the new space $\widetilde{\Omega}$. One weakness is that the parameters used to construct $\widetilde{\Omega}$ are randomly generated from $p(\omega)$ which degrades the approximation accuracy of $\widetilde{\Omega}$ to Ω. Another problem is that the utility bounds use the same regularization parameter value to compare the private and nonprivate classifiers. They take no consideration into the change of regularization parameter incurred by privacy constraints. Chaudhuri et al. [23] investigated a general mechanism, namely, DPERM, to produce private approximations of classifiers by regularized empirical risk minimization (ERM) with good perturbation error. Akin to PrivateSVM, DPERM requires that the underlying kernel is translation invariant. In this paper, we will compare our method to the PrivateSVM algorithm, since DPERM has comparable performance with PrivateSVM.

3. Preliminary

Consider an original dataset $D = \{(\mathbf{x}_i, y_i) \mid i \in Z^+, 1 \leq i \leq n\}$ that contains a small portion of public data D_{public} and a large part of private data D_{private}. Our goal is to release a differentially private support vector machine using both public and private data. In this section, we first introduce the definition of differential privacy; then, we give a brief overview of SVM and RBF kernel.

3.1. Differential Privacy. Differential privacy has emerged as one of the strongest privacy definitions for statistical data release. It guarantees that if an adversary knows complete information of all the tuples in D except one, the output of a differentially private randomized algorithm should not give the adversary too much additional information about the remaining tuples. We say that datasets D and D' differ in only one tuple if we can obtain D' by removing or adding only one tuple from D. A formal definition of differential privacy is given as follows.

Definition 1 (ϵ-differential privacy [18]). Let $\mathscr{A}$ be a randomized algorithm over two datasets D and D' differing in only one tuple, and let $\mathcal{O}$ be any arbitrary set of possible outputs f $\mathscr{A}$. Algorithm $\mathscr{A}$ satisfies ϵ-differential privacy if and only if the following holds:

$$\Pr\left[\mathscr{A}(D) \in \mathcal{O}\right] \leq e^{\epsilon} \Pr\left[\mathscr{A}\left(D'\right) \in \mathcal{O}\right]. \quad (1)$$

Intuitively, differential privacy ensures that the released output distribution of $\mathscr{A}$ remains nearly the same whether or not an individual tuple is in the dataset.

A common mechanism to achieve differential privacy is the Laplace mechanism [18] that adds a small amount of independent noise to the output of a numeric function f to

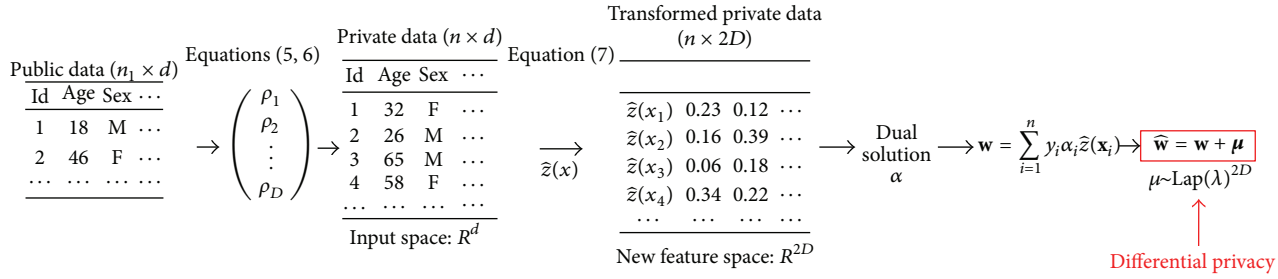


Figure 2: Detailed framework of our hybrid SVM.

fulfill ϵ-differential privacy of releasing f, where the noise is drawn from *Laplace distribution* with a probability density function $\Pr[\eta = x] = (1/2b)e^{-|x|/b}$. A Laplace noise has a variance $2b^2$ with a magnitude of b. The magnitude b of the noise depends on the concept of *sensitivity* which is defined as follows.

Definition 2 (sensitivity [18]). Let f denote a numeric function, and the sensitivity of f is defined as the maximal L_1-norm distance between the outputs of f over the two datasets D and D' which differ in only one tuple. Formally,

$$\Delta_f = \max_{D,D'} \left\| f(D) - f\left(D'\right) \right\|_1. \tag{2}$$

With the concept of sensitivity, the noise follows a zero-mean Laplace distribution with the magnitude $b = \Delta_f/\epsilon$. To fulfill ϵ-differential privacy for a numeric function f over D, it is sufficient to publish $f(D) + X$, where X is drawn from $\text{Lap}(\Delta_f/\epsilon)$.

3.2. Review of SVM and RBF Kernel. SVM is one of the most popular supervised binary classification methods that takes a sample and a predetermined kernel function as input, and outputs a predicted class label for this sample. Consider training data $D = \{(\mathbf{x}_i, y_i) \mid i \in Z^+, 1 \le i \le n\}$, where $\mathbf{x}_i \in R^d$ denotes the training input points, $y_i \in \{1, -1\}$ are the training class labels, and n is the size of training data. Here, d is the dimension of input data and "+1" and "−1" are class labels. A SVM maximizes the geometric margin between two classes of data and minimizes the error from misclassified data points. The primal form of a soft-margin SVM can be written as

$$\min_{\mathbf{w} \in R^F} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} l\left(y_i, f_{\mathbf{w}}\left(\mathbf{x}_i\right)\right), \tag{3}$$

where $\mathbf{w}$ is the normal vector to the hyperplane separating two classes of data, $l(y, \widehat{y})$ is a loss function convex in $\widehat{y}$, C is a regularization parameter that weighs smoothness and errors (i.e., large for fewer errors, smaller for increased smoothness), and $f_{\mathbf{w}}(\mathbf{x}_i) = \langle \phi(x_i), \mathbf{w} \rangle$, where $\phi(\mathbf{x}) : R^d \to R^F$ is a function mapping training data point from their input space R^d to a new F-dimensional feature space R^F (F may be infinite). Sometimes we map the training data from their input space to another high-dimensional feature space in order to classify nonlinearly separable data. When F is large or infinite, the innerproducts in feature space R^F may be computed efficiently by an explicit representation of the kernel function $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. For example, $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$ is a linear kernel function for a linear SVM, and $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/\sigma^2)$ is a RBF kernel function, which is translation invariant.

In this paper, we use a RBF kernel function. Our method can be applied to any translation invariant kernel SVM. With the hinge loss $l(y_i, f_{\mathbf{w}}(\mathbf{x}_i)) = \max(0, 1 - y_i f_{\mathbf{w}}(\mathbf{x}_i))$, we can obtain a dual form SVM written as

$$\begin{aligned} &\max_{\boldsymbol{\alpha} \in R^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ &\text{s.t.} \quad 0 \le \alpha_i \le C, \quad \forall i \in 1, \ldots, n, \end{aligned} \tag{4}$$

where $\alpha_i \in \boldsymbol{\alpha}$, $i \in (1, n)$ is a persample parameter and $w_j \in \mathbf{w}$, $j \in (1, d)$ is a perfeature weight parameter. The weight vector $\mathbf{w}$ can be converted from sample weight vector $\boldsymbol{\alpha}$ via $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$ in the linear SVM.

4. Privacy Preserving Hybrid SVM

In this section, we first introduce a framework overview and then the technical details of our hybrid SVM method. We assume that all data samples follow the same distribution. Here, we assume that all original data from different data sets follow some unknown joint multivariate distribution and all data tuples are samples from this distribution.

4.1. The General Framework. Figure 2 illustrates the general framework of hybrid SVM. Algorithm 1 presents the hybrid SVM algorithm. First, we use the small amount of public data and (5) and (6) to compute the parameter $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D)^T$, $\boldsymbol{\rho}_i \in R^d$ in the mapping function of the approximation form to the RBF kernel. Second, with $\boldsymbol{\rho}$, we transform the private data from the original sample space to the new $2D$-dimensional feature space via the mapping function $\widehat{z}(x)$ in (7). Then we can compute the parameter $\boldsymbol{\alpha}$ in the dual space with the transformed private data and $\mathbf{w}$ in the primal space via the linear relationship between $\boldsymbol{\alpha}$ and $\mathbf{w}$ in the linear SVM. Finally, draw $\boldsymbol{\mu}$ from $\text{Lap}(\lambda)^{2D}$ where $\lambda = 2^{2.5} C \sqrt{D}/n\epsilon$ and return $\widehat{\mathbf{w}} = \mathbf{w} + \boldsymbol{\mu}$ and $\boldsymbol{\rho}$. Then users can transform their test data to the new $2D$-dimensional feature space with $\boldsymbol{\rho}$ and classify the transformed data with $\widehat{\mathbf{w}}$. Here the computation

Input: Public data D_{public}, private data D_{private}, the dimensionality D of $\boldsymbol{\rho}$, a regularization parameter C, and privacy budget ϵ;
Output: Differentially private SVM;
(1) Use the public data to compute $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D)^T$ via (5), (6);
(2) Transform each record of the private data to new $2D$-dimensional data via the mapping function $\widehat{z}(x)$ defined by (7);
(3) Compute the parameter $\boldsymbol{\alpha}$ in the dual space with the transformed private data, and $\mathbf{w}$ in the primal space via $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \widehat{z}(\mathbf{x}_i)$;
(4) Draw $\boldsymbol{\mu}$ from $\text{Lap}(\lambda)^{2D}$, $\lambda = 2^{2.5} C\sqrt{D}/n\epsilon$, then return $\widehat{\mathbf{w}} = \mathbf{w} + \boldsymbol{\mu}$ and $\boldsymbol{\rho}$.

Algorithm 1: Hybrid SVM algorithm.

of parameter $\boldsymbol{\rho}$ has no privacy risk because it is retrieved directly from public data. More details about hybrid SVM will be given in the successive subsections.

Privacy Properties. We present the following theorem showing the privacy property of Algorithm 1.

Theorem 3. *Algorithm 1 guarantees ϵ-differential privacy.*

Proof. For step 1, no private data is used, and hence step 1 does not impact the privacy guarantee. Due to Corollary 15 in [22] and the fact that the hinge-loss is convex and 1-Lipschitz in $\widehat{y}$, the sensitivity of $\mathbf{w}$ over a pair of neighbouring datasets is $\Delta_{\mathbf{w}} = 2^{2.5} C\sqrt{D}/n$. Then the scale parameter λ in step 4 is set to $\lambda = \Delta_{\mathbf{w}}/\epsilon = 2^{2.5} C\sqrt{D}/n\epsilon$ due to the Laplace mechanism introduced in Section 3.1. Therefore, Algorithm 1 preserves ϵ-differential privacy which completes the proof. □

4.2. The Computation of $\boldsymbol{\rho}$. Rahimi and Recht [24] approximate a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ induced by an infinite dimensional feature mapping with a random RKHS $\widehat{\mathcal{H}}$ induced by a random finite-dimensional mapping z. The random finite-dimensional RKHS $\widehat{\mathcal{H}}$ can be constructed by drawing D i.i.d. vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ from the Fourier transform of a positive-definite translation-invariant kernel function $k(x, y)$, such as the RBF kernel function. Then we can obtain an approximation form $z(x)^T z(y)$ of $k(x, y)$ using the real-valued mapping function $z(x) : R^d \rightarrow R^D$ defined by the following equation:

$$z(x) = \sqrt{\frac{2}{D}} \left[\cos\left(\boldsymbol{\rho}_1^T \mathbf{x} + b_1\right) \cdots \cos\left(\boldsymbol{\rho}_D^T \mathbf{x} + b_D\right)\right]^T, \quad (5)$$

where $b_1, \ldots, b_D$ are i.i.d. samples drawn from a uniform distribution $U[0, 2\pi]$. $z(x) : R^d \rightarrow R^D$ maps the data from its original d-dimensional input space to the new D-dimensional feature space. Their approach is based on the fact that the kernel function of a continuous positive-definite translation-invariant kernel is the Fourier transform of a nonnegative measure. The uniform convergence property of the approximation form $z(x)^T z(y)$ to the kernel function $k(x, y)$ has also been proved in [24]. In our context, the kernel function $k(x, y)$ refers to the RBF kernel function.

In our problem setting, since a small amount of public data can be considered as x in $z(x)$ and only the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ are needed to construct the random finite-dimensional RKHS $\widehat{\mathcal{H}}$, we can compute the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ with an optimization function defined as follows:

$$\min_{\boldsymbol{\rho} \in R^{D \times d}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{2}{D} z(x_i)^T z\left(x_j\right) - k\left(x_i, x_j\right) \right|. \quad (6)$$

Since (6) is an unconstrained nonlinear optimization function, we solve it using L-BFGS (the full name is Limited-memory Broyden Fletcher Goldfarb Shanno) algorithm.

Thus, we can obtain a more accurate approximation form $z(x)^T z(y)$ of the kernel function $k(x, y)$ by deploying the public data to compute the $\boldsymbol{\rho}$, than randomly sampling $\boldsymbol{\rho}$ from the fourier transform of the kernel function $k(x, y)$ as shown in [25]. To guarantee differential privacy, we need only consider the data-dependent weight parameter $\mathbf{w}$. Fortunately we can employ the differentially private linear SVM approach in [25] to compute $\mathbf{w}$ after transforming all private data to a new $2D$-dimensional feature space using the mapping $\widehat{z}(x) : R^d \rightarrow R^{2D}$ defined in (7) with the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ as follows:

$$\widehat{z}(x) = \frac{1}{\sqrt{D}} \left[\cos\left(\boldsymbol{\rho}_1^T x\right), \sin\left(\boldsymbol{\rho}_1^T x\right), \ldots, \cos\left(\boldsymbol{\rho}_D^T x\right), \sin\left(\boldsymbol{\rho}_D^T x\right)\right]^T. \quad (7)$$

4.3. The Computation of $\widehat{\mathbf{w}}$. With the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ to approximate the RBF kernel function, we can convert RBF kernel SVM in the d-dimensional input space into the linear SVM in a new $2D$-dimensional feature space with (7), then use the privacy preserving linear SVM algorithm in [25]. The general idea of this algorithm is that with the transformed $2D$-dimensional private data, we first compute the parameter $\boldsymbol{\alpha}$ in the dual space and then $\mathbf{w}$ in the primal space using $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \widehat{z}(\mathbf{x}_i)$; then we draw $\boldsymbol{\mu}$ from $\text{Lap}(\lambda)^{2D}$, where $\lambda = 2^{2.5} C\sqrt{D}/n\epsilon$ and compute noisy $\widehat{\mathbf{w}}$ with $\widehat{\mathbf{w}} = \mathbf{w} + \boldsymbol{\mu}$.

5. Experiments

In this section, we experimentally evaluate our hybrid SVM and compare it with one state-of-the-art method, called

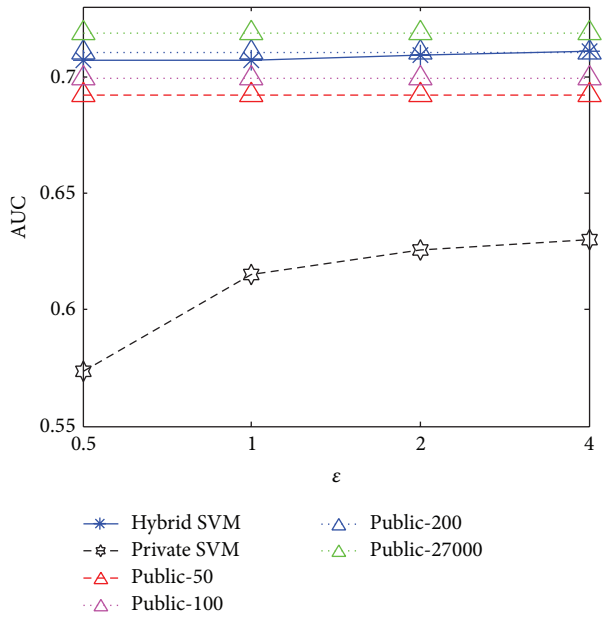


Figure 3: AUC versus privacy budget for US.

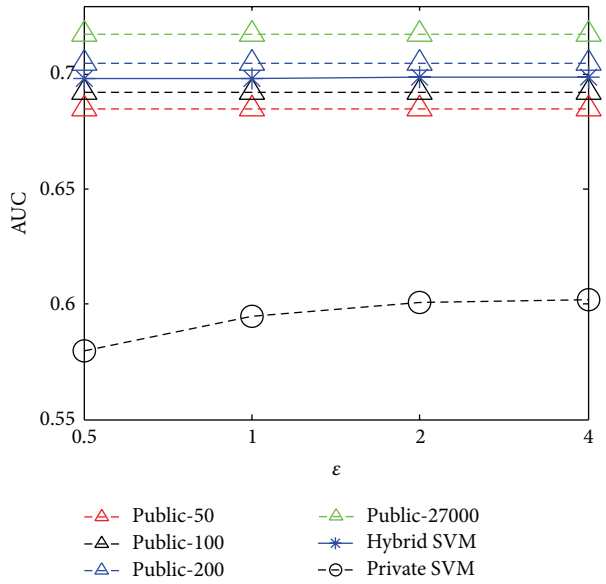


Figure 4: AUC versus privacy budget for Brazil.

private SVM and on baseline method. We evaluate the utility of the trained SVM classifier using the AUC metric. Hybrid SVM and private SVM are implemented in MATLAB R2010b, and all experiments were performed on a PC with 3.2 GHz CPU and 8 G RAM.

5.1. Experiment Setup

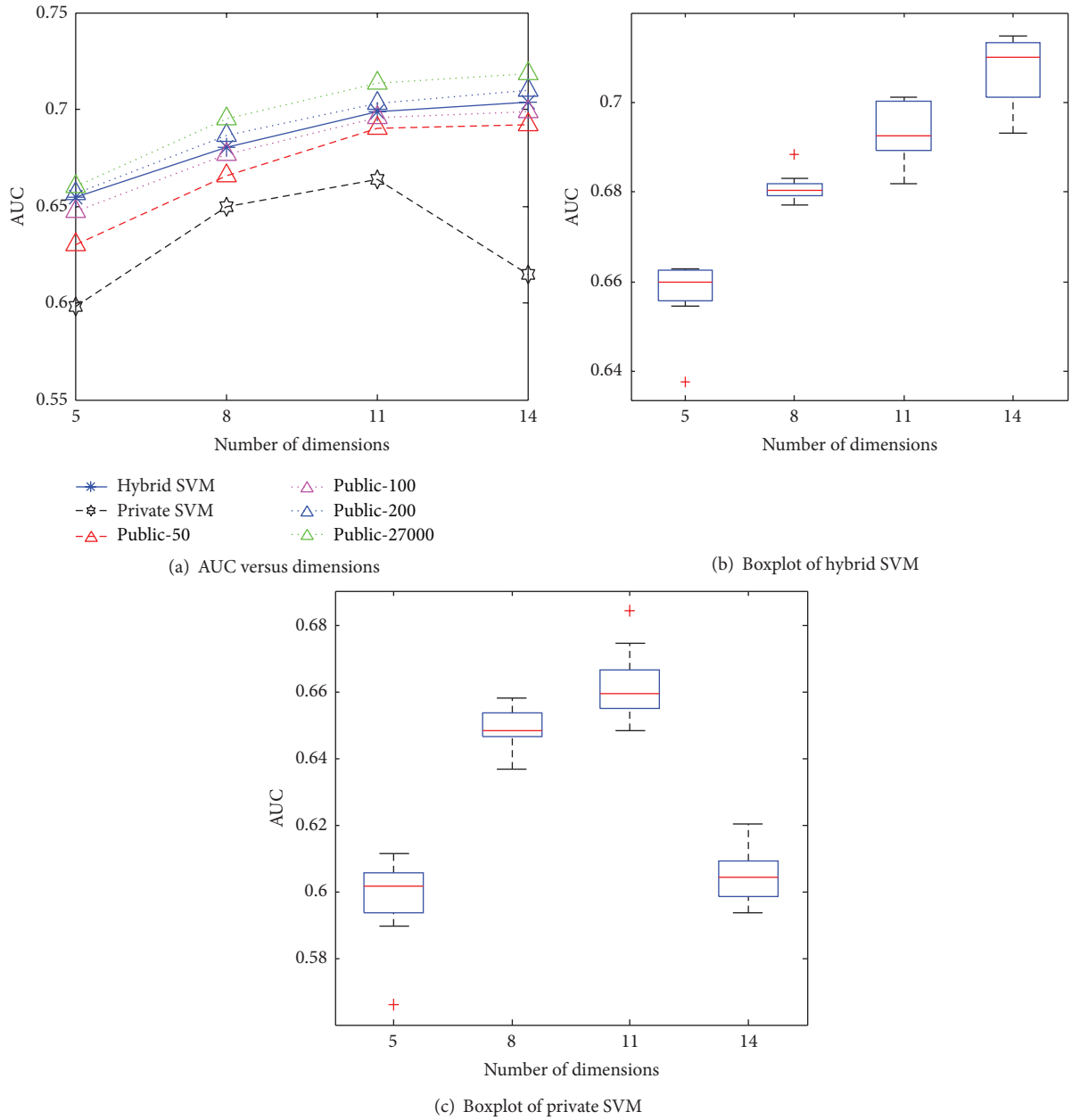
Datasets. We used two open source datasets from the Integrated Public Use Microdata Series (Minnesota Population Center, Integrated public use microdata series—international: Version 5.0., 2009, https://international.ipums.org), the US and Brazil census datasets with 370,000 and 190,000 records collected in the US and Brazil, respectively. One motivation for using these public datasets is that it bears similar attributes (e.g., demographic features) as some medical records, but it is publicly available for testing and comparisons. From each dataset, we selected 40,000 records, with 10,000 records serving as the public data pool. There were 13 attributes in both datasets, namely, *age, gender, marital status, education, disability, nationality, working hours per week, number of years residing in the current location, ownership of dwelling, family size, number of children, number of automobiles*, and *annual income*. Among these attributes, *marital status* is the only categorical attribute containing more than 2 values, that is, *single*, *married*, and *divorced/widowed*. Because SVMs do not handle categorical features by default, we transformed *marital status* into two binary attributes, *is single* and *is married* (an individual divorced or widowed would have false on both of these attributes). With this transformation, our two datasets had 14 dimensions. For each dataset, we randomly extract a subset of original data as a public data pool, from which public data is sampled uniformly, and use the remaining 30000 tuples as the private data.

Table 1: Experiment parameters.

Parameter	Default value
Number of records in the public data used by hybrid SVM	20
Number of records in the private training dataset	27000
Number of records in the test dataset	3000
Number of dimensions	14
Privacy budget ϵ	1.0

Comparison. We experimentally compared the performance of our hybrid SVM against two approaches, namely, public data baseline and private SVM [25]. The public data baseline is a RBF kernel SVM that uses only public data. In our experiment figures, we use "Public—#" to denote the public data baseline method with # as the size of public data. The private SVM is a state-of-the-art differentially private RBF kernel SVM that uses private data only. The parameters in all methods are set to optimal values.

Metrics. We used the other attributes to predict the value of *annual income* by converting *annual income* into a binary attribute: values higher than a predefined threshold were mapped to 1, and otherwise to −1. Here, we set the predefined threshold as the median value of *annual income*. The classification accuracy was measured by the AUC (the area under an ROC curve) [26]. The boxplot was used to measure the stability of our method and private SVM. The boxplots of "Public—50," "Public—100," and "Public—200," are qualitatively similar to our hybrid SVM; hence, we do not report boxplots of these baseline methods. We performed 10-fold cross-validation 10 times for each algorithm and reported the average results. We varied three different parameters: the privacy budget ϵ, the dataset dimensionality, and the



(a) AUC versus dimensions

(b) Boxplot of hybrid SVM

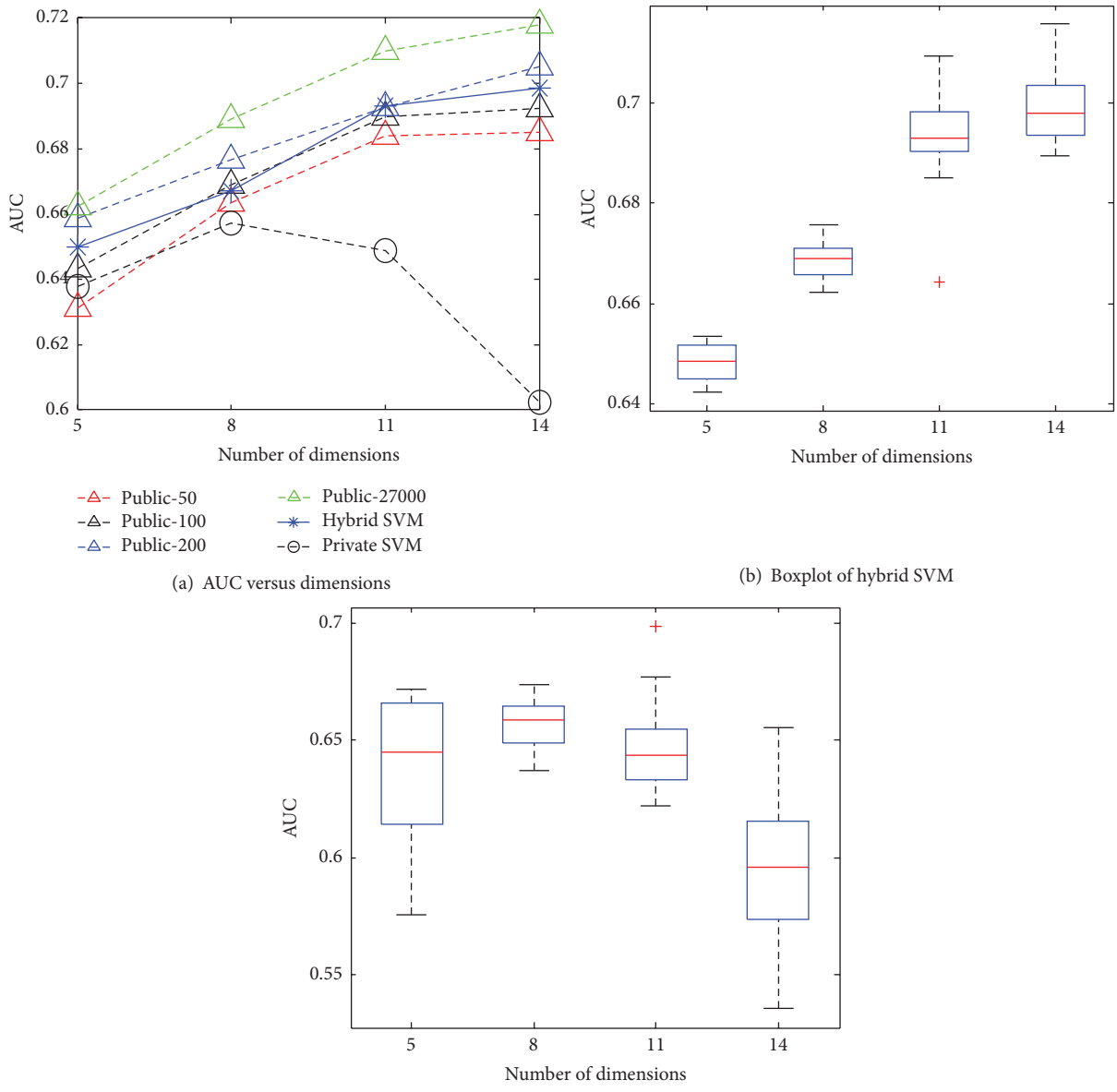
(c) Boxplot of private SVM

FIGURE 5: AUC versus dimensions for US.

data cardinality (i.e., the size of training data). To vary the data cardinality parameter, we randomly generate subsets of records in the training records set, with the sampling rate varying from 0.1 to 1. For various data dimensionalities with the range being 5, 8, 11, and 14, we select three attribute subsets in the US and Brazil datasets for classification. The first five dimensions include: *age*, *gender*, *education*, *family size*, and *annual income*. The second eight dimensions contain the previous five attributes, and additionally *nativity*, *owner of dwelling*, and *number of automobiles*. The third eleven dimensions consist of all the attributes in the second 8 dimensions and *is single*, *is married*, and *number of children*. Table 1 summarizes the parameters and their default values in the experiments.

5.2. AUC versus Privacy Budget. Figures 3 and 4 illustrate the AUCs of each method under various privacy budgets from 0.5 to 4, where "Public—#" means the public data baseline methods with various sizes of public data. Observe that our hybrid SVM outperforms the private SVM and performs better than the public data baseline defined by the public data. The AUC of our method remains stable under all privacy budgets and is significantly close to the public data baseline that uses the complete private data set as public data.

5.3. AUC versus Dataset Dimensionality. Figures 5 and 6 present the AUCs of each algorithm as a function of the dataset dimensionality for the US and Brazil datasets. With



(a) AUC versus dimensions

(b) Boxplot of hybrid SVM

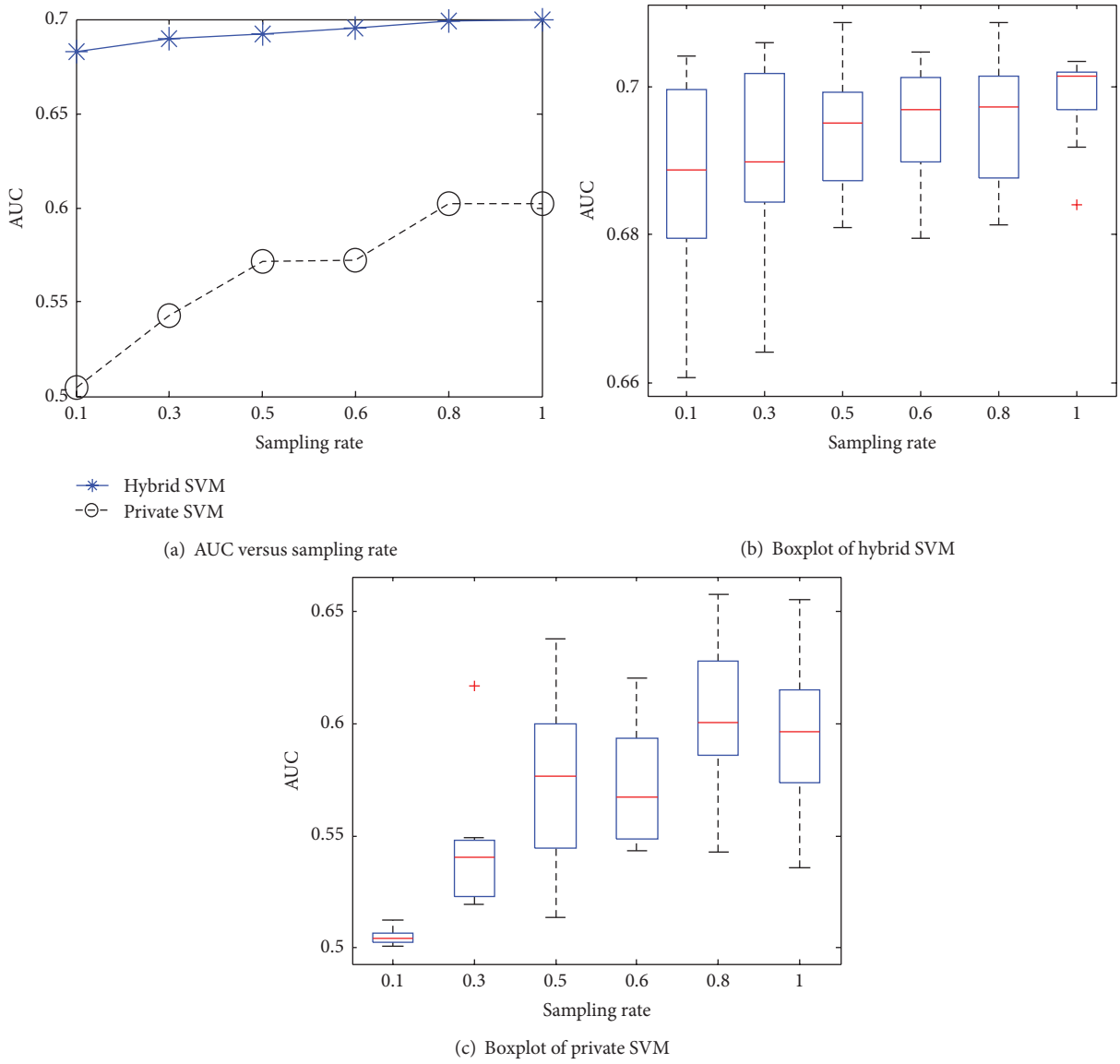
(c) Boxplot of private SVM

Figure 6: AUC versus dimensions for Brazil.

a higher number of dimensions, the AUCs of the hybrid SVM and of the SVM that uses the public data (baseline) increase. This is reasonable because the training data size with the default value being 27,000 is much larger than the number of data dimensions which are at most 14. When the number of dimensions grows, the performance improves. In contrast, the performance of the private SVM degrades in 14 dimensions with poor boxplots because more noise is introduced with higher dimensions.

5.4. AUC versus Data Cardinality. Figures 7 and 8 investigate the relationship between the sampling rate and AUC of hybrid and private SVMs. From the figures, our method consistently outperformed the private SVM at different sampling rates. It is worth mentioning that AUCs of the hybrid SVM are large even at small sampling rates and tend to stabilize when the size of training data grows (i.e., large sampling rate). The boxplots reflect that the private SVM has larger variance than the hybrid SVM, because private SVM selects the values of $\boldsymbol{\rho}$ randomly from the Fourier transform of RBF kernel. In contrast, hybrid SVM computes $\boldsymbol{\rho}$ via the public data. This helps improve the accuracy of $\boldsymbol{\rho}$ and leads to less variance.

5.5. Computation Time. Finally, Figure 9 shows the time cost of our proposed algorithm with varying dimensions and different sampling rates. We only report the results for the US dataset; the results for the Brazil dataset are greatly similar. One can notice that the dimensionality, rather than the sampling rate, determines the computational cost of the hybrid SVM. The overhead of the hybrid SVM is



(a) AUC versus sampling rate

(b) Boxplot of hybrid SVM

(c) Boxplot of private SVM

Figure 7: AUC versus sampling rate for US.

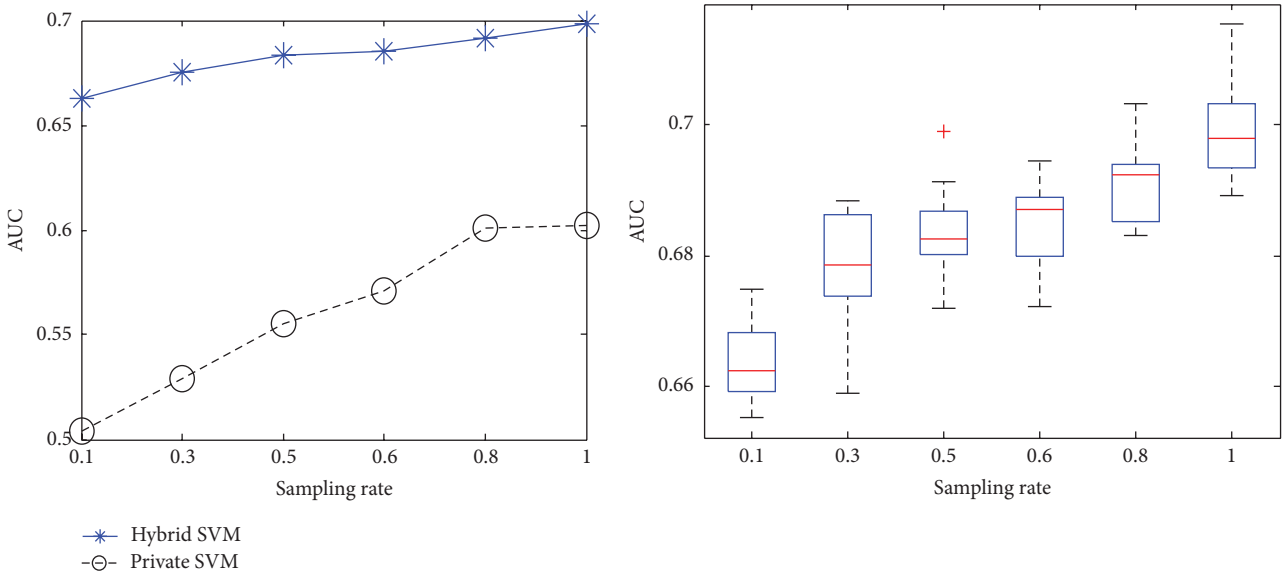
from computing $\boldsymbol{\rho}$ with the public data, since a nonlinear optimization equation needs to be solved. As the other private SVM methods, our hybrid SVM is intended for off-line use, and hence the time is generally acceptable for even 14 dimensional datasets.

6. Discussion and Conclusion

We proposed and developed a RBF kernel SVM using a small amount of public data and a large amount of private data to preserve differential privacy with improved utility. In this algorithm, we use public data to compute the parameters in an approximation form of the RBF kernel function and then train private classifiers with linear SVM after converting all private data into a new feature space defined by the approximation form. A limitation of our approach is that we used the L-BFGS method [27], which is not very efficient, to find the optimal solution. Because the objective function in (6) is not a convex function, our model is computationally intensive in order to calculate the local optimal values, especially when the size of the public data set is large. We will develop more efficient methods and test the model on clinical records in future work. Another limitation is that we assume all original data from different data sets follow some unknown joint multivariate distribution. Our assumption might now always be true in practice, and calibration is necessary for future investigation. That is, in the presence of distributional difference, we will leverage transfer learning to build the global model.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

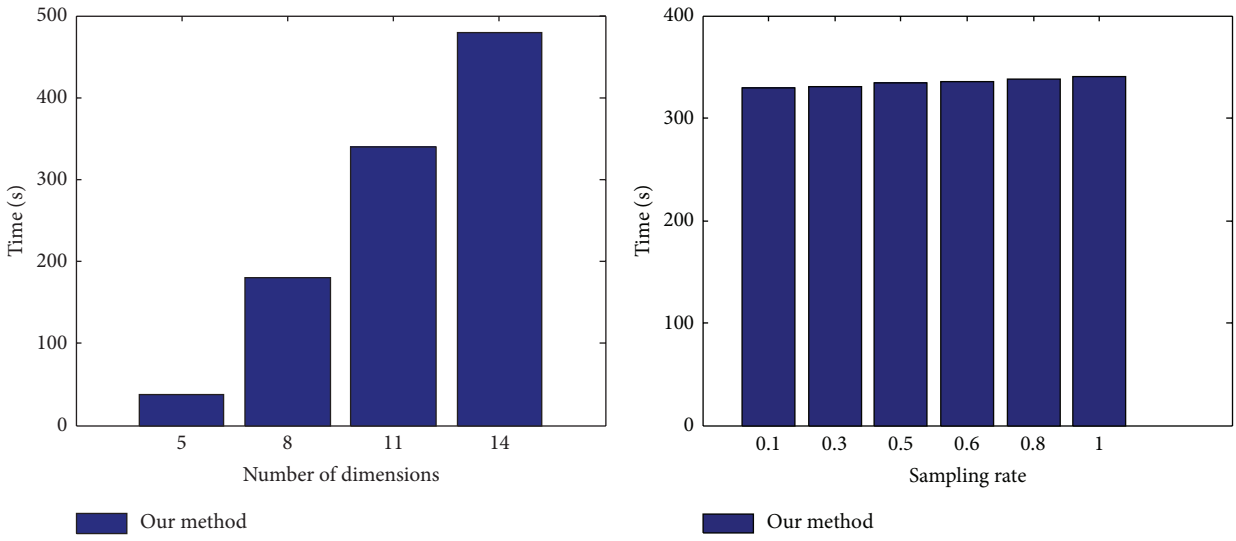


(a) AUC versus sampling rate

(b) Boxplot of hybrid SVM

(c) Boxplot of private SVM

Figure 8: AUC versus sampling rate for Brazil.



(a)

(b)

Figure 9: Time versus dimensions and sampling rate.

Acknowledgments

Lucila Ohno-Machado and Xiaoqian Jiang are partially supported by NLM (R00LM011392) and iDASH (NIH Grant U54HL108460).

References

[1] L. Ohno-Machado, V. Bafna, A. A. Boxwala et al., "iDASH: integrating data for analysis, anonymization, and sharing," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 196–201, 2012.

[2] L. Ohno-Machado, "To share or not to share: that is not the question," *Science Translational Medicine*, vol. 4, no. 165, Article ID 165cm15, 2012.

[3] N. Homer, S. Szelinger, M. Redman et al., "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, Article ID e1000167, 2008.

[4] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[5] D. McGraw, "Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 29–34, 2013.

[6] L. Ohno-Machado, C. Farcas, J. Kim, S. Wang, and X. Jiang, "Genomes in the cloud: balancing privacy rights and the public good," in *AMIA Clinical Research Informatics Summit*, 2013.

[7] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: a survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, article 14, 2010.

[8] X. Jiang, A. D. Sarwate, and L. Ohno-Machado, "Privacy technology to support data sharing for comparative effectiveness research: a systematic review," *Medical Care*, vol. 51, no. 8, pp. S58–S65, 2013.

[9] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[10] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*, pp. 338–340, 2nd edition, 2011.

[11] C. Dwork, "Differential privacy: a survey of results," in *Theory and Applications of Models of Computation—TAMC*, pp. 1–19, 2008.

[12] N. Mohammed, X. Jiang, R. Chen, B. C. M. Fung, and L. Ohno-Machado, "Privacy-preserving heterogeneous health data sharing," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 462–469, 2013.

[13] J. Gardner, L. Xiong, Y. Xiao et al., "Share: system design and case studies for statistical health information release," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 109–116, 2013.

[14] H. Li, L. Xiong, L. Zhang, and X. Jiang, "DPSynthesizer: differentially private data synthesizer for privacy preserving data sharing," in *Proceedings of the 40th International Conference on Very Large Data Bases (VLDB '14)*, Hang Zhou, China, 2014.

[15] S. A. Vinterbo, A. D. Sarwate, and A. A. Boxwala, "Protecting count queries in study design," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 750–757, 2012.

[16] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Information Sciences*, vol. 231, pp. 83–97, 2013.

[17] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 493–501, August 2011.

[18] C. Dwork, F. McSherry, K. Nissim, and A. Smith, ". Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference—TCC*, pp. 265–284, 2006.

[19] H. Li, L. Xiong, and X. Jiang, "Differentially private synthesization of multi-dimensional data using copula functions," in *Proceedings of the 17th International Conference on Extending Database Technology (EDBT '14)*, pp. 475–486, Athens, Greece, 2014.

[20] X. Jiang, Z. Ji, S. Wang, N. Mohammed, S. Cheng, and L. Ohno-Machado, "Differential-private data publishing through component analysis," *Transactions on Data Privacy*, vol. 6, no. 1, pp. 19–34, 2013.

[21] Z. Ji, X. Jiang, S. Wang, L. Xiong, and L. Ohno-Machado, "Differentially private distributed logistic regression using public and private biomedical datasets," in *Proceedings of the 3rd Annual Translational Bioinformatics Conference (TBC '13)*, Seoul, Korea, 2013.

[22] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: privacy-preserving mechanisms for svm learning," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, pp. 65–100, 2009.

[23] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, pp. 1069–1109, 2011.

[24] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS '07)*, Vancouver, Canada, December 2007.

[25] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: privacy preserving mechanisms for SVM learning," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, pp. 65–100, 2012.

[26] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 404–415, 2005.

[27] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming B*, vol. 45, no. 3, pp. 503–528, 1989.