Differentially Private Multi-task Learning

Sunil Kumar Gupta(✉), Santu Rana, and Svetha Venkatesh

Center for Pattern Recognition and Data Analytics, Deakin University, Geelong 3216, Australia
{sunil.gupta,santu.rana,svetha.venkatesh}@deakin.edu.au

Abstract. Privacy restrictions of sensitive data repositories imply that the data analysis is performed in isolation at each data source. A prime example is the isolated nature of building prognosis models from hospital data and the associated challenge of dealing with small number of samples in risk classes (e.g. suicide) while doing so. Pooling knowledge from other hospitals, through multi-task learning, can alleviate this problem. However, if knowledge is to be shared unrestricted, privacy is breached. Addressing this, we propose a novel multi-task learning method that preserves privacy of data under the strong guarantees of differential privacy. Further, we develop a novel attribute-wise noise addition scheme that significantly lifts the utility of the proposed method. We demonstrate the effectiveness of our method with a synthetic and two real datasets.

1 Introduction

Privacy matters. Tough legislations are in place to safeguard the privacy of individual data repositories, resulting in a large set of disconnected data islands with no means of connection between them. However, as data mining researchers we believe that knowledge is everywhere and the true potential of data will be unlocked when these disparate data islands are appropriately bridged - a pursuit remains unrealized in the presence of privacy restrictions.

Consider healthcare as an example. Modern healthcare facilities are equipped with Electronic Medical Records systems that capture detailed information about patients as they access hospital facilities. The value of such information is immense in creation of accurate prognosis models, central to efficient and appropriate delivery of care. However, we often encounter diseases that have *small number of samples in risk classes*, e.g. suicide is rare in populations. The prognosis model built in such situations may result in poor performance. Pooling knowledge across hospitals via *multi-task learning* framework can alleviate this problem, but privacy-protecting regulatory frameworks control access to sensitive data across hospital jurisdictions. Similar situations arise in other areas, *e.g.* building spam filters in a collaborative yet privacy-preserving manner etc. *Therefore, there is opportunity to develop privacy preserving multi-task learning models that provides strong guarantees on privacy protection.*

Early work on privacy preserving data analysis has used a variety of methods, e.g. query restriction [1], anonymization of information [2], secure multi-party function evaluation [3], data/output perturbations [4] etc. Of these, query

M. Chau et al. (Eds.): PAISI 2016, LNCS 9650, pp. 101–113, 2016.
DOI: 10.1007/978-3-319-31863-9_8

restriction provides limited utility [5], anonymization may reveal sensitive data in presence of auxiliary information [6] and secure multi-party function evaluation may not provide statistical guarantees. Recently, differential privacy has emerged as a framework for privacy preserving information disclosure with strong theoretical guarantees [7]. It ensures that the answer to a statistical query is not significantly different between two datasets that differ at most in one instance. A major strength of differential privacy is its ability to provide graded levels of privacy by specifying a leakage parameter ε, giving rise to the name ε-differential privacy. Differential privacy has been applied to many data mining areas [7,8]. The closest work to our problem of building prediction models is the output perturbation approach due to [7], who suggested that differentially private algorithms can be constructed by calibrating the standard deviation of a Laplacian noise according to the "sensitivity" of a function involved in the algorithm. This idea was followed by Chaudhuri et al. to build differentially private empirical risk minimization models [9]. However, applicability of these models is limited to only *single-task learning* scenarios.

Current methods leverage collective knowledge across prediction problems (or tasks) via multi-task learning [10,11], building prediction models where inter-task knowledge transfer is achieved via some form of joint modeling. Existing multi-task learning models, however, are not equipped to satisfy privacy requirements as they require unrestricted access to sensitive data [11] or derived statistics [10]. A recent state-of-the-art multi-task learning method is MTRL [12], which provides a flexible way of sharing knowledge across tasks by using a covariance matrix to model task relationships. As a result, it is able to exploit knowledge from tasks that have varying degree of relatedness - a crucial property when dealing with real world data.

Limited work exist on privacy preserving multi-task learning. Mathew and Obradovic [13] construct a distributed Id3-based decision tree for predicting hospitalization risk from multi-hospital data. Although no data is exchanged between the hospitals, leakage on privacy may occur due to exchanging unperturbed statistics. Pathak et al. [14] propose a differentially private multi-task learning via averaging classifiers from multiple sources using secure multi-party communication. This method has two drawbacks. (1) Averaging classifiers assumes that tasks are strongly correlated, and (2) the level of noise is calibrated with respect to the *smallest* source, resulting in *high* model perturbation. This seriously brings down the utility of the algorithm. *Therefore, the opportunity to develop a differentially private multi-task learning model* is open.

Taking this opportunity, we propose a novel multi-task learning that preserves privacy of individuals in participating sources, under the strong guarantee of differential privacy. The proposed model infuses privacy into the MTRL model [12]. This delivers strong privacy preserving property to a state-of-art multi-task learning model facilitating seamless sharing of knowledge without sharing data between the participating sources. In case of healthcare, this means that hospitals across the world can improve their prognosis models by leveraging their mutual knowledge and thus derive best practice to revolutionize healthcare.

Following sensitivity based approach of Dwork et al. [7], we derive the sensitivity for the proposed multi-task learning model and use it to calibrate the level of noise used for differential privacy. We provably show that the proposed scheme satisfies ε-differential privacy where it is possible for a source (e.g. hospital) to control its privacy requirements via its own ε parameter. Adding calibrated noise to task parameters helps us in securing privacy, however, it comes at the cost of reduced performance (or utility). Addressing this problem, we propose a novel scheme of attribute-wise noise addition that exploits the information content of attributes and *reduces* the overall noise. We demonstrate that this scheme can significantly lift the performance of the proposed technique. Using a synthetic dataset, we illustrate the behavior of the proposed models and validate their effectiveness using *two* real world problems: predicting cancer mortality and designing personalized spam filter.

2 Preliminaries

Differential Privacy. Differential privacy is a privacy preserving framework proposed by Dwork et al. [7]. This framework defines a notion of privacy for a learning algorithm $\mathscr{A}$. The algorithm $\mathscr{A}$ satisfies differential privacy if likelihood of its output for two datasets that differ at most by one instance are close. Due to this closeness, an adversary can not infer anything significant about the differing instance by using the algorithm output. The closeness of the likelihoods is characterized by a "leakage" parameter ε, giving rise to the name ε-differential privacy.

Definition 1 [7]: *An algorithm $\mathscr{A}$ is said to satisfy ε-differential privacy if for any two datasets $\mathscr{D}$ and $\mathscr{D}'$ that differ by at most one instance, and all $S \subseteq Range(\mathscr{A})$,*

$$\exp(-\varepsilon) \leq \frac{P(\mathscr{A}(\mathscr{D}) \in S)}{P(\mathscr{A}(\mathscr{D}') \in S)} \leq \exp(\varepsilon) \tag{1}$$

where $\mathscr{A}(\mathscr{D})$ and $\mathscr{A}(\mathscr{D}')$ are the outputs of $\mathscr{A}$ on datasets $\mathscr{D}$ and $\mathscr{D}'$ respectively.

Sensitivity. The sensitivity of a function f is the maximum change in its output due to any single data instance. A formal definition of sensitivity is provided below:

Definition 2 [7]: *The sensitivity of a function $f : D \rightarrow \mathbb{R}^M$ is defined as*

$$S(f) = \max_{\mathscr{D},\mathscr{D}'} ||f(\mathscr{D}) - f(\mathscr{D}')||$$

for all datasets $\mathscr{D}$ and $\mathscr{D}'$ that differ by at most one instance. Dwork et al. [7] showed that ε-differential privacy is satisfied by an algorithm if i.i.d. Laplacian noise with standard deviation $S(f)/\varepsilon$ is added in each co-ordinate of the output vector before its release.

Strong Convexity. The strong convexity property is used to derive our proposed model. We provide the definition of strong convexity in the following.

Definition 3. *A twice continuously differential function $C(f)$ is called strongly convex with parameter $\mu > 0$ iff the following inequality holds for all f in its domain*

$$\nabla^2 C(f) \succ \mu I, \tag{2}$$

where $\succ$ means that $\nabla^2 C(f) - \mu I$ is positive semi-definite.

3 Privacy Preserving MTL

Let us assume we have T_0 tasks, indexed as $t = 1, \ldots, T_0$. For the t-th task, we denote the training set as $\mathscr{D}_t = \{(\mathbf{x}_{ti}, y_{ti})\}_{i=1}^{N_t}$ where $\mathbf{x}_{ti} \in \mathbb{R}^M$ is a M-dimensional feature vector and y_{ti} is the target, usually real-valued for regression and binary-valued for binary classification problems. Let β_t denote the weight vector for the task t, we also refer to this as *task parameter*. Collectively, we denote the data of t-th task by $\mathbf{X}_t = (\mathbf{x}_{t1}, \ldots, \mathbf{x}_{tN_t})^T$ and $\mathbf{y}_t = (y_{t1}, \ldots, y_{tN_t})^T$ and all the task parameters as $\beta = (\beta_1, \ldots, \beta_{T_0})$. When tasks differ in some of the features, a common feature list can be obtained via their union.

The multi-task learning literature is full of sophisticated models where the aim is to jointly model multiple tasks towards improved average prediction performance for all the tasks. In this paper, we use a multi-task learning model that learns relationship of tasks via a covariance matrix and uses it for joint modeling [12]. Although we have chosen this model to build a privacy preserving variant, one can use the technique described in this paper to many other multi-task learning models provided these models minimize a convex loss function. The results for non-convex models are more involved and out of scope of this paper.

The proposed multi-task learning model minimizes the following objective function

$$\min_{\beta,\Omega} \sum_t \frac{||\mathbf{X}_t\beta_t + b_t\mathbf{1} - \mathbf{y}_t||^2}{N_t} + \lambda_1 \mathrm{Tr}\left(\beta\beta^T\right) + \lambda_2 \mathrm{Tr}\left(\beta\Omega^{-1}\beta^T\right), \quad \text{s.t. } \Omega \succeq 0, \ \mathrm{tr}(\Omega)=1 \tag{3}$$

where b_t is the bias parameter of the t-th task and the notation $\mathbf{1}$ denotes a vector of all ones. *We refer to the above cost function as $C(\beta, \Omega)$*. Although, in this paper, we use the square loss, it is possible to extend this formulation for logistic loss. Similarly, extensions to multi-class classification is straight-forward. We take the above model and build its privacy preserving variant, which protects the data from being reversed engineered by an adversary from model parameters.

Since the cost function of 3 is jointly convex in β and Ω along with the constraints, we can find unique solution. Our approach is to optimize β for a fixed Ω and then optimize Ω given β. This leads to an iterative solution.

For square loss, task parameter β_t given Ω can be learnt in a closed form. This is done by setting the derivative of $C(\beta, \Omega)$ with respect to β_t to zero, leading to the following linear equation in β_t

Algorithm 1. The proposed Private-MTL

1: **Input**: Multi-task data $\{\mathbf{X}_t, \mathbf{y}_t\}_{t=1}^{T_0}$, parameters λ_1, λ_2, ε.
2: **Output**: Task parameters $\beta_{1:T_0}$ and matrix Ω.
3: **Initialization:** For initialization, learn task parameters using single task learning (STL), let us assume the task parameter for task t using STL is β_t, computed locally.
4: compute sensitivity for task t as $S_t = \frac{2}{N_t \Lambda_t}$ where $\Lambda_t = \lambda_1 + \lambda_2 \Omega^{-1}(t,t)$ and N_t is the number of instances in t-th task.
5: sample η_t from the density function: $p(\eta_t) \propto \exp\left(-\frac{\varepsilon}{S_t}||\eta_t||\right)$.
6: Add noise to task parameters as $\beta_t = \beta_t + \eta_t$.
7: **repeat**
8: update task relationship matrix as $\Omega = \frac{\left(\beta^T\beta\right)^{1/2}}{\mathrm{Tr}\left(\left(\beta^T\beta\right)^{1/2}\right)}$.
9: solve β_t given Ω and other noisy $\beta_{t'}$, $t' \neq t$ using (4).
10: set $\beta_t = \beta_t + \eta_t$ where η_t is sampled similar to step-5.
11: **until** convergence

$$\left[\left(\mathbf{X}_t^T\mathbf{X}_t\right)/N_t + \left(\lambda_1 + \lambda_2\Omega^{-1}(t,t)\right)\mathbf{I}\right]\beta_t = \left(\mathbf{X}_t^T\left(\mathbf{y}_t - b_t\mathbf{1}\right)\right)/N_t - \lambda_2 \sum_{t'\neq t} \Omega^{-1}(t',t)\,\beta_{t'}. \quad (4)$$

As seen from this equation, for a fixed task relatedness Ω, learning task parameter of t-th task, i.e. β_t requires data from only its own task. The knowledge from *other tasks* is brought through their task parameters, i.e. $\beta_{t'}$ where $t' \neq t$. To have a solution that preserves privacy according to the ε-differential privacy, we follow the sensitivity method suggested by [7]. We compute sensitivity (S_t) of our objective function and add a noise vector (η_t) calibrated using this sensitivity to the task parameters. This method can be shown to guarantee the privacy of data instances from all the tasks. Using this method, the noisy β_t is given as

$$\beta_t = \beta_t + \eta_t,\ p(\eta_t) \propto \exp\left(-\frac{\varepsilon}{S_t}||\eta_t||\right). \quad (5)$$

where we slightly abuse the notation of β_t using it to denote the task parameters both *before* and *after* adding noise.

For a fixed noisy β, the matrix Ω can be learnt by minimizing $\mathrm{Tr}\left(\beta\Omega^{-1}\beta^T\right)$ subject to constraints $\Omega \succeq 0, \mathrm{tr}(\Omega) = 1$. Zhang et al. [12] show that the closed form solution of this optimization problem is given as $\Omega = \frac{\left(\beta^T\beta\right)^{1/2}}{\mathrm{Tr}\left(\left(\beta^T\beta\right)^{1/2}\right)}$. We note that there is no need to add noise in Ω, as it is estimated from β, which is already noisy and privacy preserving. For all future references, we term this model as **Private-MTL**.

In spite of adding noise to β_t, the convergence of the optimization function in (3) is still guaranteed under the framework of stochastic optimization. The noisy perturbations (with mean zero) to β_t can be thought as updating β_t using a noisy gradient of the cost function, which is popular in stochastic optimization literature and known to converge [15]. Algorithm 1 provides a step-by-step summary of the proposed model.

Privacy Guarantees. We establish the conditions under which Algorithm 1 provide ε-differential privacy with respect to β. Proving differential privacy w.r.t. Ω is not necessary as by reverse engineering Ω, one can only reach to β, which is noisy and privacy preserving.

Theorem 1. *The task parameters $\beta_{1:T_0}$ learnt using Algorithm 1 preserves ε-differential privacy.*

Proof: The proof of the Theorem follows a similar sketch as the proof of Theorem 6 in Chaudhuri et al. [9]. Due to involvement of multi-task regularization, the sensitivity of the model, however, is different. Lemma 3 derives the sensitivity of our proposed model ($S_t = \frac{2L}{N_t \Lambda_t}$, where $\Lambda_t = \lambda_1 + \lambda_2 \Omega^{-1}(t,t)$ and loss function is assumed to be L-Lipschitz). This result, in combination with the sensitivity method of Dwork et al. [7] and output perturbation method of [9], establishes the theorem. For the sake of completeness, we provide a sketch of the proof here.

Let $\mathscr{D}$ and $\mathscr{D}'$ be any two datasets that differ only in n_t-th instance of task t. Further, let $\beta_t^{\mathscr{D}}$ and $\beta_t^{\mathscr{D}'}$ be the task parameters learnt using these datasets without any noise additions. Let β_t be the task parameter after noise addition. Then, for any β_t and dataset $\mathscr{D}$, we have $p(\beta_t|\mathscr{D}) \propto e^{-\frac{N_t \Lambda_t \varepsilon}{2L}\left(||\beta_t - \beta_t^{\mathscr{D}}||\right)}$, which leads to

$$\frac{p(\beta_t|\mathscr{D})}{p(\beta_t|\mathscr{D}')} = e^{-\frac{N_t \Lambda_t \varepsilon}{2L}\left(||\beta_t - \beta_t^{\mathscr{D}}|| - ||\beta_t - \beta_t^{\mathscr{D}'}||\right)} \tag{6}$$

where $p(\beta_t|\mathscr{D})$ and $p(\beta_t|\mathscr{D}')$ are the density function of the task parameter β_t given datasets $\mathscr{D}$ and $\mathscr{D}'$. Using triangular inequality and Lemma 3, we have

$$||\beta_t - \beta_t^{\mathscr{D}}|| - ||\beta_t - \beta_t^{\mathscr{D}'}|| \leq ||\beta_t^{\mathscr{D}'} - \beta_t^{\mathscr{D}}|| \leq \frac{2L}{N_t \Lambda_t}.$$

Plugging this result in (6), we have $\frac{p(\beta_t|\mathscr{D})}{p(\beta_t|\mathscr{D}')} \geq e^{-\varepsilon}$. Due to symmetry in choosing $\mathscr{D}$ and $\mathscr{D}'$, we also have $\frac{p(\beta_t|\mathscr{D})}{p(\beta_t|\mathscr{D}')} \leq e^{\varepsilon}$, guaranteeing ε-differential privacy.

Lemma 2. *The cost function of (3) for task t is Λ_t-strongly convex with $\Lambda_t = \lambda_1 + \lambda_2 \Omega^{-1}(t,t)$.*

Proof: We note that $C(\beta, \Omega)$ is doubly-differentiable. In this light, consider the strong convexity condition in (2). To prove the lemma, we need to show: $\nabla^2_{\beta_t} C(\beta, \Omega) \succ \left(\lambda_1 + \lambda_2 \Omega^{-1}(t,t)\right) \mathbf{I}$. The first derivative of the cost function in (3) is given as

$$\nabla_{\beta_t} C = \frac{\left(\mathbf{X}_t^T \mathbf{X}_t \beta_t - \mathbf{X}_t^T (\mathbf{y}_t - b_t \mathbf{1})\right)}{N_t} + \lambda_1 \beta_t + \lambda_2 \beta \Omega^{-1}(:,t). \tag{7}$$

Taking second derivative, we get the following result

$$\nabla^2_{\beta_t} C = \frac{1}{N_t} \mathbf{X}_t^T \mathbf{X}_t + \lambda_1 \mathbf{I} + \lambda_2 \Omega^{-1}(t,t)\, \mathbf{I}. \tag{8}$$

Clearly the matrix $\nabla^2_{\beta_t} C - \left(\lambda_1 + \lambda_2 \Omega^{-1}(t,t)\right) \mathbf{I}$ is positive semi-definite as, for any $\mathbf{v}$, $\mathbf{v}^T \left(\mathbf{X}_t^T \mathbf{X}_t\right) \mathbf{v} = \left(\mathbf{X}_t \mathbf{v}\right)^T \left(\mathbf{X}_t \mathbf{v}\right) \geq 0$. Therefore, the cost function $C(\beta, \Omega)$ for each β_t (i.e. task t) is Λ_t-strongly convex with $\Lambda_t = \lambda_1 + \lambda_2 \Omega^{-1}(t,t)$.

Lemma 3. Assuming bounded input ($||\mathbf{x}_{ti}|| \leq 1$) and L-Lipschitz assumption on the loss function, the sensitivity of $C(\beta, \Omega)$ for task t is at most $\frac{2L}{N_t \Lambda_t}$, where $\Lambda_t = \lambda_1 + \lambda_2 \Omega^{-1}(t,t)$.

Proof: To derive the sensitivity of $C(\beta, \Omega)$, consider two datasets $\mathscr{D}$ and $\mathscr{D}'$ that differ only in n_t-th instance of task t (denoted as $d_{t,n_t} = (\mathbf{x}_{t,n_t}, y_{t,n_t})$). Further let $G(\beta, \Omega) = C(\beta, \Omega)|_{\mathscr{D}'}$, $g(\beta) = C(\beta, \Omega)|_{\mathscr{D}} - C(\beta, \Omega)|_{\mathscr{D}'}$, $\beta_t^{\mathscr{D}} = \text{argmin}_{\beta_t} C(\beta, \Omega)|_{\mathscr{D}}$, and $\beta_t^{\mathscr{D}'} = \text{argmin}_{\beta_t} C(\beta, \Omega)|_{\mathscr{D}'}$. The sensitivity of β_t is given by $\max_{d_{t,n_t}} ||\beta_t^{\mathscr{D}} - \beta_t^{\mathscr{D}'}||$, which following a similar derivation as Lemma 7 and Corollary 8 in [9] can be shown to be at most $\frac{2L}{N_t \Lambda_t}$, where $\Lambda_t = \lambda_1 + \lambda_2 \Omega^{-1}(t,t)$.

Attribute-Wise Noise Addition. In above method, both ε and the sensitivity S_t are set identically for all attributes - noise is added isotropically. We aim to *reduce* the level of noise by exploiting attribute-specific properties. The main idea is that an attribute needs to be kept strictly private only when it is rich in information. For an attribute that does not carry much information, there is not much reason to make it private. For example, most patients in a cancer hospital will have chemotherapy, thus enforcing stringent privacy on 'whether someone has undergone chemotherapy or not' is unnecessary. For such attributes, we can relax the privacy constraint by setting parameter ε to a higher value. We do this by setting ε as a function of the attribute entropy ($\mathbb{H}_{ti}$). Entropy is a surrogate to capture the uniqueness of an attribute. In particular, we set the privacy level for the i-th attribute in task t as $\varepsilon_{ti} = \varepsilon_0 (1 + \kappa_f \exp(-\mathbb{H}_{ti}))$, where $\mathbb{H}_{ti}$ represents uncertainty of i-th attribute in task t and κ_f is an *"attribute-wise privacy scale parameter"*. The parameter κ_f decides the rate at which the privacy requirements are relaxed with decreasing attribute uncertainty. Depending on the level of entropy, ε_{ti} varies between $(\varepsilon_0, \varepsilon_0(1+\kappa_f)]$. For continuous valued feature differential privacy can be used. Using attribute-wise privacy parameter ε_{ti}, for task t, the perturbation in i-th element of task parameter is given as

$$\beta_{ti} = \beta_{ti} + \eta_{ti}, \; p(\eta_{ti}) \propto \exp\left(-\frac{\varepsilon_{ti}}{S_t}|\eta_{ti}|\right). \tag{9}$$

The proof on privacy guarantee is similar to the proof in Theorem 3.1. Only difference is that we are now using independent Laplacian noise with different parameters for each attribute instead of using *i.i.d.* noise. In Algorithm 1, the step-5, step-6 and step-10 are appropriately replaced by Eq. (9).

4 Experiments

4.1 Experimental Setup

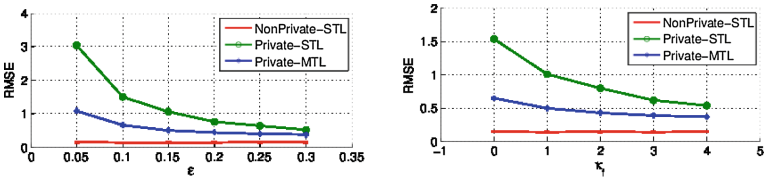
We experiment with a synthetic dataset and two real datasets.

We compare our Private-MTL with the following baselines: *(a) NonPrivate-STL:* In this algorithm, prediction weight vectors are learnt separately at each entity and released without privacy protection, *(b) NonPrivate-MTL:* In this

algorithm, prediction weight vectors are learnt using multi-task learning but without privacy restriction, *(c) Private-STL:* In this algorithm, weight vectors of all entities are learnt separately and noise is added to preserve privacy, and *(d) MDP-AC:* In this algorithm [14], weight vectors are shared via secure multiparty computation and averaged to obtain a global classifier. The noise is added corresponding to the smallest dataset. For all the above baselines and the proposed models, the regularization parameters are learnt using cross-validation.

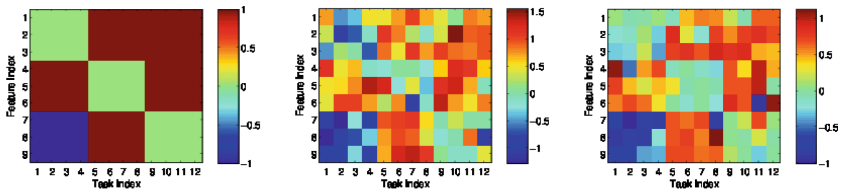
4.2 Experiments with Synthetic Dataset

We synthesize a multi-task learning dataset where tasks have various form of relationships: positive, negative and no relationship. Our aim is to show that our proposed model is able to estimate the task parameters (β_t) accurately even under privacy preserving restrictions. We create a total of 12 tasks with their task parameters defined in a 9-dim. space. We simulate three task groups putting the first 4 tasks in group-1, the next 4 tasks in group-2 and the last 4 tasks in group-3. We use different relatedness across task groups along different features. In particular, for the first 3 features, tasks in group-2 and group-3 are *positively* related, but *unrelated* to tasks in group-1. Similarly, for the next 3 features, tasks in group-1 and group-3 are *positively* related, but *unrelated* to tasks in group-2. Finally, for the last 3 features, tasks in group-1 and group-2 are



(a) (b)

Fig. 1. Experimental results on Synthetic dataset: (a) Root Mean Square Error (RMSE) as a function of privacy parameter (ε), and (b) RMSE as a function of attribute-wise privacy scale parameter (κ_f) at $\varepsilon_0 = 0.1$.



(a) (b) (c)

Fig. 2. Task parameters for Synthetic data experiments: (a) True, (b) Private-STL, and (c) Private-MTL. The task parameters shown are average of 50 run with $\varepsilon = 0.1$ and $\kappa_f = 1$; Between the task parameters obtained by Private-STL and Private-MTL, the latter resembles more to the true task parameters used for the synthesis of data.

negatively related, but unrelated to tasks in group-3. Overall relationship across tasks aggregated over all features becomes partial.

The i-th instance for t-th task, i.e. $\mathbf{x}_{ti}$ is generated from a 9-dim. multivariate normal distribution as $\mathbf{x}_{ti} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The outcome y_{ti} is generated as $y_{ti} = \beta_t^T \mathbf{x}_{ti} + e_{ti},\ e_{ti} \sim \mathcal{N}(0, 0.1)$, where e_{ti} is a random noise. We generate 100 instances per task.

We randomly divide the data from each task into a 70 % training set and a 30 % test set. Figure 1a shows the average predictive performance as a function of the privacy parameter ε for the proposed Private-MTL and the two STL baselines. As seen from the figure, NonPrivate-STL provides the lowest RMSE error. This is because there is no noise addition in the task parameters. Out of the two privacy preserving algorithms, Private-MTL performs better than the Private-STL. This benefit comes from the ability of the Private-MTL successfully leveraging from the knowledge of the other tasks. At the stronger privacy requirements (i.e. for lower values of ε), the performance benefit is even more pronounced with Private-MTL performing two times better than Private-STL at $\varepsilon = 0.05$. Figure 1b shows similar plots as a function of κ_f when the global privacy parameter is set at $\varepsilon_0 = 0.1$. This plot clearly shows significant improvement in performance due to attribute-wise anisotropic noise addition. As the data is synthesized, we know the true task parameters and compare that with the recovered task parameters by different algorithms. Figure (2a–c) provide depiction of both the true and the recovered task parameters. It is evident that the task parameters recovered by Private-MTL resemble more closely to the true task parameters than that of by Private-STL.

4.3 Experiments with Real Dataset

Cancer Dataset. The Cancer dataset is obtained from a large regional hospital in Australia[1]. This cohort consists of 4,200 cancer patients who visited the hospital during 2010-2012. The data contains a variety of information such as patient demographics, diagnosis records in terms of ICD-10 codes, and procedure codes in terms of ACHI system. Features are extracted following [16], resulting in 683 features. The task involves 1 year mortality prediction. To simulate a *multi-hospital scenario*, we randomly divide the whole cohort into 5 separate cohorts that are assumed to be coming from hospitals of different sizes: a large hospital (LH) with 3000 patients, a medium hospital (MH) with 600 patients, and 3 small hospitals (SH) with 200 patients each. The *unequal* division reflects the typical real-world setting where several medium to small hospitals work together with a large hospital in the nearby city.

Multi-user Spam Dataset. The spam dataset is obtained from the *ECML-PKDD challenge* held in 2006. We use the test dataset from the Task B challenge. The dataset contains 15 users, each having 400 labeled emails. The emails are supplied as a term-document matrix with a dictionary size of around 150,000.

[1] Ethics approval obtained through University and the hospital – 12/83.

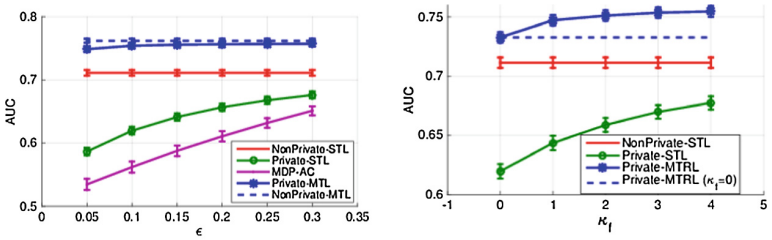
For each user spams constitute 50 % of the emails. The goal is to build a spam classifier for each user locally. Since the notion of spam emails is related across users, multi-task learning can be used to share classification knowledge across the users. However, emails being private in nature, we should aim to perform any knowledge sharing in a privacy preserving way.

Experimental Results

Cancer Dataset. Figure 3a shows the predictive performance averaged across all the hospitals as a function of privacy parameter ε on the Cancer dataset. Randomly selected 70 % data from each hospital is used for training and the rest for test. The experiment is performed 50 times and the average performance is reported along with respective standard error. As seen from the figure, the highest average Area Under the ROC Curve (AUC) over all the hospitals is achieved by the Private-MTL across the range of privacy parameter (ε) tested. Even at a stringent privacy requirement of $\varepsilon = 0.05$, Private-MTL achieves $\sim 5\,\%$ higher AUC than the NonPrivate-STL. This indicates that the hospitals benefit by collaboration than building tools independently. This is significant since collaborating privately introduces noise in the estimation, yet improvement in prediction is achieved. Private-STL may not be a suitable benchmark as we may assume that learning independently at each hospital does not require the use of privacy preserving algorithms. However, it is still a useful comparator to illustrate the absolute gain achieved just because of multi-task learning. At $\varepsilon = 0.05$, the improvement achieved by the multi-task learning is more than 30 % over the Private-STL. As expected, the performance by both the Private-STL and Private-MTL improves with increasing ε (decreasing privacy requirement), however, the benefits obtained by using multi-task learning remains significant. Further, the performance of Private-MTL almost reaches the performance of Non-Private MTL. MDP-AC performs lower but somewhat closer to the Private-STL. As the data is originally from a single source averaging the weight vectors had the potential to work. However, the amount of noise is computed based on the smallest sized hospital. This has resulted in a higher noise, leading to a poor performance.

In the Cancer dataset, we have one large, one medium and three small sized hospitals. Table 1 shows the performance by different algorithms at different hospitals at $\varepsilon = 0.1$. We compare the performance between Private-MTL over the NonPrivate-STL. We see that the gain in performance by the Private-MTL is more for medium and small sized hospitals (average gain = 4.5 %) than the large hospital (gain = 1.5 %). This shows the higher need of multi-task learning for medium/smaller sized hospitals.

Figure 3b shows the use of attribute-wise privacy for the cancer dataset. Figure 4 shows the histogram of attribute-wise entropy across all the 5 hospitals. It clearly shows that there are many features which have low entropy. Figure 3b shows the predictive performance as a function of the attribute-wise privacy scale parameter κ_f when the global privacy parameter is set at $\varepsilon_0 = 0.1$. Average



(a) (b)

Fig. 3. (a) Average AUC of prediction on Cancer dataset as a function of the privacy parameter ε, and (b) average AUC of prediction on Cancer dataset as a function of the attribute-wise privacy parameter κ_f when $\varepsilon_0 = 0.1$. For both average performance is reported over 50 random training/test splits (std. errors shown as error-bars).

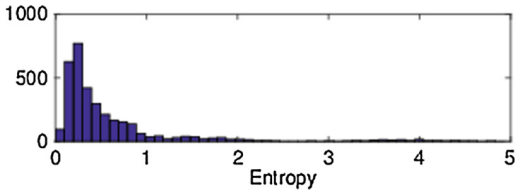


Fig. 4. Histogram of entropy across features over all the tasks (hospitals).

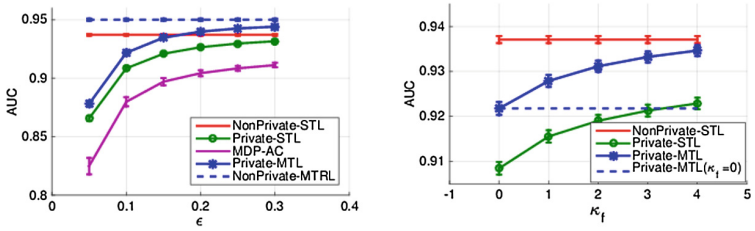
Table 1. Average AUC of prediction on Cancer dataset at different hospitals; a large hospital (LH) with 3000 patients, a medium sized hospital (MH) with 600 patients and three small hospitals (SH1-3) with 200 patients each at $\varepsilon = 0.1$. Performance is averaged over 50 random training/test splits. Standard errors are reported in parenthesis.

	AUC (std err)				
	LH	MH	SH1	SH2	SH3
NonPrivate-STL ($R1$)	0.800 (0.002)	0.683 (0.005)	0.750 (0.011)	0.713 (0.012)	0.612 (0.014)
Private-STL ($R2$)	0.791 (0.003)	0.632 (0.008)	0.576 (0.020)	0.573 (0.015)	0.527 (0.018)
Private-MTL ($R3$)	**0.815** (0.002)	**0.735** (0.005)	**0.787** (0.011)	**0.777** (0.010)	**0.657** (0.013)
$\Delta(R3 - R1)$	0.015	0.052	0.037	0.0.064	0.045

performance from 50 random splits of 70 % data for training and the rest for test is reported. The plot shows that AUC improves considerably with increasing κ_f. Private-MTL gains further 2.5 % in AUC at $\kappa_f = 4$.

Multi-user Spam Dataset. Figure 5a shows the average comparative predictive performance in terms of AUC on Spam dataset as a function of the privacy parameter ε. For each user, 70 % of the data is randomly selected for training and the rest for test. The average performance over 50 such splits are shown. As seen from the plot, the performance by all three privacy preserving algorithms are much worse than NonPrivate-STL at the stringent privacy requirement of $\varepsilon = 0.05$. However, they improved as the privacy requirement is lowered (ε

is increased). Private-MTL starts to become better at $\varepsilon \geq 0.2$, almost reaching up to NonPrivate-MTL. It implies that building spam filters collaboratively may perform better on average below a certain privacy restriction. This is a common phenomenon one will encounter while designing privacy preserving algorithms. Figure 5b shows the corresponding performance when $\varepsilon_0 = 0.1$ and the attribute-wise privacy parameter κ_f is varied. As expected, the performance improves when the attribute-wise anisotropic noise is introduced.



(a) (b)

Fig. 5. Average AUC achieved by various algorithms on the Spam dataset: (a) as a function of privacy parameter ε, and (b) as a function of attribute-wise privacy scale parameter κ_f when $\varepsilon_0 = 0.1$. The results are averaged over 50 random training-test splits. Standard errors are shown as error bars.

5 Conclusion

We propose a novel multi-task learning model that preserves privacy of individuals at participating tasks under differential privacy. To lift the model's utility, we develop a novel attribute-wise noise addition scheme that adds anisotropic noise calibrated according to uncertainty of the attributes leading to reduced noise. Comparing with the state-of-art baselines on two real world datasets we demonstrate the effectiveness of our approach. In future, we will continue exploring connections between multi-task learning and privacy by extending differentially private random forest [17] to multi-task learning or extending model-agnostic multi-task learning [18] to a privacy-preserving variant.

References

1. Chin, F.Y., Ozsoyoglu, G.: Auditing and inference control in statistical databases. IEEE Trans. Softw. Eng. **8**(6), 574–582 (1982)
2. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficientfull-domain k-anonymity. In: SIGMOD, pp. 49–60. ACM (2005)
3. Ben-David, A., Nisan, N., Pinkas, B.: Fairplaymp: a system for securemulti-party computation. In: ACM CCS, pp. 257–266. ACM (2008)
4. Traub, J.F., Yemini, Y., Woźniakowski, H.: The statistical security of a statistical database. TODS **9**(4), 672–679 (1984)

5. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: PODS, pp. 202–210. ACM (2003)
6. Ganta, S., Kasiviswanathan, S., Smith, A.: Composition attacks and auxiliary information in data privacy. In: SIGKDD, pp. 265–273. ACM (2008)
7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
8. Vaidya, J., Clifton, C.W., Zhu, Y.M.: Privacy Preserving Data Mining, vol. 19. Springer Science & Business Media, New York (2006)
9. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. J. Mach. Learn. Res. **12**, 1069–1109 (2011)
10. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Mach. Learn. **73**(3), 243–272 (2008)
11. Saha, B., Gupta, S., Phung, D., Venkatesh, S.: Multiple task transfer learning with small sample sizes. In: Knowledge and Information Systems, pp. 1–28 (2015)
12. Zhang, Y., Yeung, D.-Y.: A convex formulation for learning task relationships in multi-task learning. In: Uncertainty in Artificial Intelligence, pp. 733–442 (2010)
13. Mathew, G., Obradovic, Z.: Distributed privacy preserving decision support system for predicting hospitalization risk in hospitals with insufficient data. In: ICMLA, vol. 2, pp. 178–183 (2012)
14. Pathak, M., Rane, S., Raj, B.: Multiparty differential privacy via aggregation of locally trained classifiers. In: NIPS, pp. 1876–1884 (2010)
15. Spall, J.C.: Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control, vol. 65. Wiley, Hoboken (2005)
16. Tran, T., Luo, W., Phung, D., Gupta, S., Rana, S., Kennedy, R.L., Larkins, A., Venkatesh, S.: A framework for feature extraction from hospital medical data with applications in risk prediction. BMC Bioinform. **15**(1), 6596 (2014)
17. Rana, S., Gupta, S., Venkatesh, S.: Differentially-private random forest with high utility. In: ICDM, pp. 955–960. IEEE, Atlantic City (2015)
18. Gupta, S., Rana, S., Saha, B., Phung, D., Venkatesh, S.: A new transfer learning framework with application to model-agnostic multi-task learning. In: KAIS (2015)