CP-PDS: Contextual Privacy-aware Framework for Personal Data Storage

Bikash Chandra Singh, Md Sipon Miah, Tapan Kumar Godder, and Md Mahbubur Rahman

Department of Information and Communication Engineering, Islamic University, Kushtia-7003, Bangladesh
bikash070@gmail.com, sipon@ice.iu.ac.bd, tkice@iu.ac.bd, mrahman@ice.iu.ac.bd

Abstract

User's contextual information play a vital role when taking decision of whether to share personal data with third parties. When disclosing personal information from PDS with third parties, users may take into account several issues for scaling the privacy threat w.r.t consider the scaling of the acceptability of the services according to his/her requirement. More specifically, they want to release the right amount of data with third parties depending on his current context with a minimum amount of interaction by considering their benefits. Prior studies on PDS so far that evolve privacy preference mechanism has not considered user's contextual information. Moreover, prior research has shown that user privacy preferences may vary based on his/her contextual information. To address this issue, this research addresses to implement a contextual privacy-aware framework for PDS (CP-PDS) which exploits contextual information to build a learning classifier that can predict user privacy preferences under various contextual scenarios. Moreover, CP-PDS also consider the elements of access requests to build ensemble classifiers for learning user privacy preferences. From these two learning perspectives, CP-PDS computes the users privacy preferences decisions on newly arrived access requests. We have performed several experiments with a group of 125 evaluators to evaluate the effectiveness of the proposed approach.

Keywords— Privacy Preference, PDS, Contextual information

Manuscript received on 28 October 2018 and accepted for publication on 21 November 2018.

1 Introduction

IN recent years, we are witnessing that personal data are scattered in various online systems managed by service providers (e.g., online social media, hospitals, airlines, etc.). It happens due to individuals getting online services from these providers by providing their personal data. Despite many benefits users get from service providers, this may also cause serious privacy threats, as users are losing control on their data. To tackle this problem, the concept of Personal Data storage (PDS) [1-5] has been proposed. A PDS is a secure digital space which can gather and keep personal information under the control of the end user. This view is also enabled by recent developments in privacy legislation and, in particular, by the new EU General Data Protection Regulation (GDPR), whose article. 20 states the right to data portability, according to which "the data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format," thus making possible data collection into a PDS. Mainly, GDPR gives control to individuals over their personal data whether they want to share or not with third parties. However, as several studies [6], [7] have shown that average web users are not expert enough to define privacy preferences on PDS data, according to their privacy requirements, thus, a key issue is that of helping users to specify their privacy preferences on PDSs. More particularly, it is required to implement mechanisms which can facilitate individuals to ensure the enforcement of their privacy preferences properly.

To date, several studies on PDS have suggested to enforce privacy preferences that regulate the third parties access to PDS [3-5], [8]. For instances, the open PDS framework in [3] defines privacy enforcements as so third party applications can send code to be run against the data and the answer is sent back to them, rather than the raw data, whereas in [5] presents a PDS architecture that relies on secure portable tokens to enforce user preferences in PDS. However, some works are designed to limit access to personal information within an organization, by deploying privacy-aware access control [9] within that organization can make decisions and enforce access control policies, intercepting queries to data repositories and returning sanitized views (if any) on requested data.

However, such research work are mainly focused on the enforcement part only and do not consider the issue of helping users in protecting their PDS data. With this consideration, we have proposed privacy preference mechanisms in [10], [11]. In these works, we have proposed user's privacy preferences mechanisms so as to help users to set up their privacy preferences in PDS by exploiting the machine learning tools. More particularly, we have exploited semi-supervised active learning to learn user privacy preferences in PDS.

Unfortunately, research work mentioned earlier, do not consider users' contextual information to make privacy preference decisions. Consequently, most of these existing privacy preferences frameworks fail to consider all aspects of users' privacy preferences so as these approaches may not be appropriate to ensure users privacy properly under the circumstance of the users' contextual perspectives.

In contrast, contextual information can be used to design privacy preference framework that can define privacy preference according to the user's current situation and in this way, it can improve the overall usability. Indeed, user contextual data refers to any piece of data that can be used to state the present situation of the user. Literature shows that users prefer to set his privacy preferences taking into account the contextual data [12], [13]. For example, let us consider that a user may feel comfortable to take a decision, when he is in travel, regard on an access request seeking current location of the user in term of suggesting some nearby famous places relevant to the user preferences for visiting. However, the same user might not accept the same request when he is working with his colleagues in his office. From this example, it clearly states that user would like to set his privacy preferences based on his contextual data. Thus, it is required to develop privacy preference mechanism that can leverage contextual data with non-contextual (e.g., data access request elements) to learn user privacy preferences efficiently. In this paper, we propose a contextual based privacy preference mechanism for PDS. To figure out the latent correlations between the contextual data and user's opinion on the access request, we want to conduct an experimental analysis on user privacy preference based on users' opinions on contextual based access request. With this intention, we consider two aspects to learn user privacy preferences. In the first step, we exploit the context information of the user to train up the learning model about user privacy preferences. Afterward, the learning model can predict the decision automatically on the newly arrived contextual based access request according to user contextual information. In the second step, we consider the elements of access request to train up the learning model and predict decision on newly arrive contextual based access request. Then we integrate these two learning decisions to produce the final decision according to the privacy preference rules given in Table 3. According to the proposed rules, the final decision will be no when CP-PDS got no from contextual data, and in other cases it will be yes/maybe learned from the elements of the access request.

Moreover, in this paper, we also want to explore the mechanism to reduce the over-fitting problem occurs in ma-chine learning approaches. In general, over-fitting occurs when a learning model learns the noise/randomness along with the samples in the training dataset that negatively i mpacts the performance of the learning model on the upcoming new samples. To reduce over-fitting, the general approach is to vary the number of training dataset sequentially and check the accuracy on the testing dataset. The fact is that those combination of training dataset produce better accuracy on testing dataset will be used for further predication.

At this purpose, we proceed with a approach: first we select the total number of training dataset according to the history based active learning (cfr. Section 2). After that we check which are the most uncertain instances (having probability difference of the class labels are very close) in the total number of training dataset and select the top 20 uncertain training dataset for a learning model and check the accuracy on the testing dataset. By the same way, we then consider top 25 uncertain training dataset and check the accuracy on testing dataset. Like this way, we proceed on and select the best model.

Therefore, the main contributions of this paper are as follows:

- We propose a contextual based privacy preference framework in PDS using machine learning tools.
- We present a different approach to reduce the over-fitting problem for machine learning.
- Empirical studies based on the real dataset demonstrate the effectiveness of the proposed contextual privacy-aware framework in PDS.

The rest of this paper is organized as follows. Section 2 describes the background information that has been taken from [11]. Section 3 introduces the overall architecture of our proposed CP-PDS framework, whereas Section 4 illustrates the experimental results. Related work are discussed in Section 5. Finally, Section 6 concludes the paper.

2 Background

In this section we provide the basic knowledge on the framework proposed in [10], [11], called Privacy-aware Personal Data Storage (P-PDS). The building block idea of P-PDS is to exploit machine learning algorithms for learning privacy preferences/habits of PDS owners. At this purpose, after a training period by the PDS owner, P-PDS aims at building a classifier able to automatically decide if, based on PDS's owner preferences, access requests submitted by third parties to PDS have to be authorized or denied.

More precisely, in this scenario, we modeled an access request such to represent the most relevant information that let individuals take conscious decisions on whether they want to release their personal data to the requesting party. This, in addition to the requested data and the access purposes, the access request contains also the type of the requesting services (e.g., medical, social, bank services) and an indication of the benefits the user can achieve by releasing his/her data, represented in terms of temporal offers associated with the requiring service. More formally, an access request is defined as follows.

Definition 1. Access request [10].

An access request AR is a tuple (DC, s_t, d_0, p, o), where DC is the data consumer, that is, the third party requesting data to the PDS, s_t is the type of service provided by DC, d_0 is the requested data, whereas p is the access purpose. If the access is granted, DC will provide an additional benefit, called offer, modeled by o.

Having this definition, in [10] we mainly focused on identifying the machine learning algorithm that better fits in the P-PDS scope. At this purpose, in [10] we considered the semi-supervised learning methods, experimentally showing that these algorithms provide a better accuracy compared to supervised learning (i.e., SVM) and with a smaller number of training dataset. Moreover, we have tested different semi-supervised learning algorithms to see which one can provide better performance in PDS scenario. Experiments results shown that the best approach is the one exploiting an *ensemble* strategy.

The main characteristic of ensemble method is that it separately trains multiple classifiers, then to compute the final decision on a new instance (aka access requests) it aggregates the class probabilities returned by obtained classifiers. The work in [10] exploits ensemble learning by applying different classifiers for each distinct pair of fields of an access request so as to model each possible relationships among AR fields (see Table 2). More particularly, we can exploit 10 classifiers that are used to compute 10 different probabilities of having a new access request AR being label as authrized (i.e., class yes), denied (i.e., class no) or to be evaluated by user (i.e., maybe class). The final decision is taken as the class with greatest aggregate probability value.

Even though [10] represents an essential step towards a Privacy-aware PDS, one critical aspect of this solution might be represented by the usability of the system. Indeed, even if compared to manually setting privacy preference P-PDS it requires much less users effort (aka, labels in the initial training dataset), it still requires many interactions with PDS owners to collect a good training datasets. To cope with this issue, in [11], we leverage on active learning (AL) [14] to minimize user burden for getting the training dataset by, at the same time, achieving better accuracy in determining user privacy preferences.

In general, the main idea of AL is to properly select from the pool of unlabeled instances those that should be labeled by users, rather than randomly choosing them as done in semi-supervised approach. At this aim, AL first selects few instances for being labeled by user and exploit them to build a preliminary prediction model. Then, AL selects only those instances from the unlabeled pool for which it is highly uncertain how to label them according to the preliminary built model. This strategy is called uncertainty sampling [14].

However, in spite of having benefits with this approach in term of accuracy and usability, AL with uncertainty sampling does not consider the semantics of instances (aka, access requests' fields) and their relevance in the PDS owner's decision process. Let consider, as an example, two access requests AR1 and AR2 with identical values except of data consumer, assuming also that AR1 has been already labeled (e.g., yes) by PDS owner. In such case, when it arrives AR2, AL might not consider this to be labeled as its uncertainty value will be low since its similarity with AR1. However, in PDS scenario, a user might fully change his decision on an access request based on the requesting data consumer (aka its reputation). In order to keep into account this, in [11] we have revised the uncertainty sampling adopted in AL so as to increase the level of uncertainty for those access requests showing in some relevant fields (i.e., data consumer and service type) values that have been never labeled by PDS owner. This uncertainty adjustment is driven by the distance between the value of data consumer/service type of the new access request and the values of the corresponding elements in access requests already labeled by the PDS owner. At this purpose, this solution traces the history of labeled access requests, as such we call this model history-based active learning (HBAL).

In [11], we also improved the ensemble approach presented in [10]. Indeed, the traditional ensemble strategy aggregates the class probabilities to compute the final decision on access request. However, this solution does not consider the relevance of each classifier, which might have a key role when predicted classes label (i.e., yes, no, maybe) are in conflicts (e.g, some suggest to authorize other to deny the access request). However, the traditional ensemble approach does not make a distinction for conflicting classes since it does not take into account the semantics associated with each decision (aka class label). To overcome this limitation, in [11] we have proposed an alternative strategy for aggregating the class labels returned by classifiers. According to this approach, we have assigned a personalized weight to each single classifier, to reflect its relevance in the user opinion. We call this approach *personalized history-based active learning (PHBAL).*

3 Contextual Privacy-aware PDS (CP-PDS)

As we discussed in Section 2 that privacy preference framework, P-PDS proposed in [11] learns PDS owner's privacy preferences by exploiting only access request elements (e.g., data consumer, requested data, service type etc.). But this approach could not fully cover all aspects of user concern in term of ensuring privacy on their personal data as it did not consider user contextual information. More specifically, user might change his mind w.r.t privacy management based on his present situation (e.g., contextual data) when the access request arrives to PDS. For example, let suppose that PDS owner U receives an access request offering the service entertainment during his office hours (e.g., office hours refer user's contextual information). In this case, U always deny this access request. But in another context, say U is in home and passing free time then U may accept this access request. Therefore, based on the contextual information, U changes his mind completely regarding on access requests. Thus, if we do not consider contextual information to train up learning models then it will produce more prediction errors (e.g., false positives/false negatives). To reduce prediction errors, in this paper, we consider contextual information with access request elements to train up the learning models.

TABLE 1
CONTEXTUAL DATA

$Context$={Day of week, Time of the day, Place, Activity}	
Day of week	{Workweek days, Weekend days}
Time of the day	{Morning (6.00-11.59), Afternoon(12.00- 17.59), Evening(18.00- 23.59), Night (0.00-5.59)}
Place	{Home, Office/School, Outside}
Activity/Feelings	{Meeting, Working, Running, Studying, Traveling, Eating, Sleeping, Idle, Physical Exercise, Driving, Sick}

TABLE 2
RELATIONSHIPS AMONG ACCESS REQUEST FIELDS USED TO BUILD THE CLASSIFIERS

AR field	Relationships
Requested data	(d_0,DC), (d_0,p),(d_0, s_t), (d_0, o)
Intended Purpose	(p,DC),(p,s_t),(p,o)
Service type	(s_t,DC), (s_t,o)
Offer	(o,DC)

3.1 Privacy preference learning with contextual data and access request elements
By considering contextual data with access request elements, we can learn user privacy preferences with two possible approaches. In the first approach, we simply extend the framework presented in [11] with contextual data. More particularly, we can train the classifiers considering not only the access request elements but also contextual information. Precisely, we can exploit ensemble approach as done in [11] that having multiple classifiers built on access request elements complemented with *CTX* data. However, such approach could not fully reflect user's privacy preferences on access request according to his/her contextual information, since this approach merges contextual information *CTX* with different subset of access request elements to train up the multiple classifiers and then aggregate the probabilities produced by the multiple classifiers to provide the final decision. More particularly, since users decision completely change on access request AR based on his/her contextual information *CTX* thus, if we can learn user privacy preference merely from *CTX* data and AR elements separately, then we believe that it will produce more accurate prediction than the previous one. Therefore, we proceed on to learn PDS owner's privacy preference from two angles: 1) learn from AR elements, and 2) learn from contextual information *CTX*. Then we combine these two decisions for getting final decision on access requests. We believe that this approach might fully cover user's privacy management aptitude since it learns user's privacy preference from the views of context and AR separately.

To do so, we first ask PDS owner about contextual based access request *AR_CTX* for labeling and use these labeled *AR_CTX* to train up the learning models. More particularly, we use semi-supervised active learning to select *AR_CTX* comes to PDS for asking label by PDS owner so as to build good classifiers with less number of labeled training dataset (see Section 2). Then, we split this dataset *AR_CTX* into two subset, labeled AR and labeled *CTX* for learning user privacy preference from two different angles (cfr. Figure 1).

Learning with access request elements. In the this learning angle, we learn user privacy preferences by exploiting only elements of access requests AR, similarly to what as been done in previous work [10][11]. We refer to this as AR learning (cfr. Figure 1).

Learning with contextual data. In the second perspective, we learn user privacy preferences (aka a classifier) using contextual information *CTX*. We refer to this as contextual learning (cfr. Figure 1). Indeed, contextual information is a broad term that can be more formally represented as a tuple *CTX* of attributes, each one collecting a meaningful data to represent access request circumstances, like time of the request, current location of PDS's owner, etc. In this paper, for simplicity we assume that *CTX* consists of four attributes (see Table 1). Additional contextual data could be easily added as well, however this small set contains data able to well describe the actual user's circumstances. We exploit these attributes to train a classifier for contextual learning. More precisely, we build a classifier using the labeled *CTX* items (see in Figure 1). However, since we get PDS owner feedback on contextual access request, *AR_CTX* that shows whole scene to PDS owner about the access request thus, for contextual learning, a classifier being trained with only *CTX* items could not perfectly learn PDS owner's opinions on which type of access request *AR*, they want to authorize based on their contextual data. Let consider the previous example, where PDS owner U's contextual information just express U's location, time, and activities but only considering the contextual information could not express for which type of access request U be able to accept/deny. Thus, for training up a individual classifier, we consider some part of access request AR

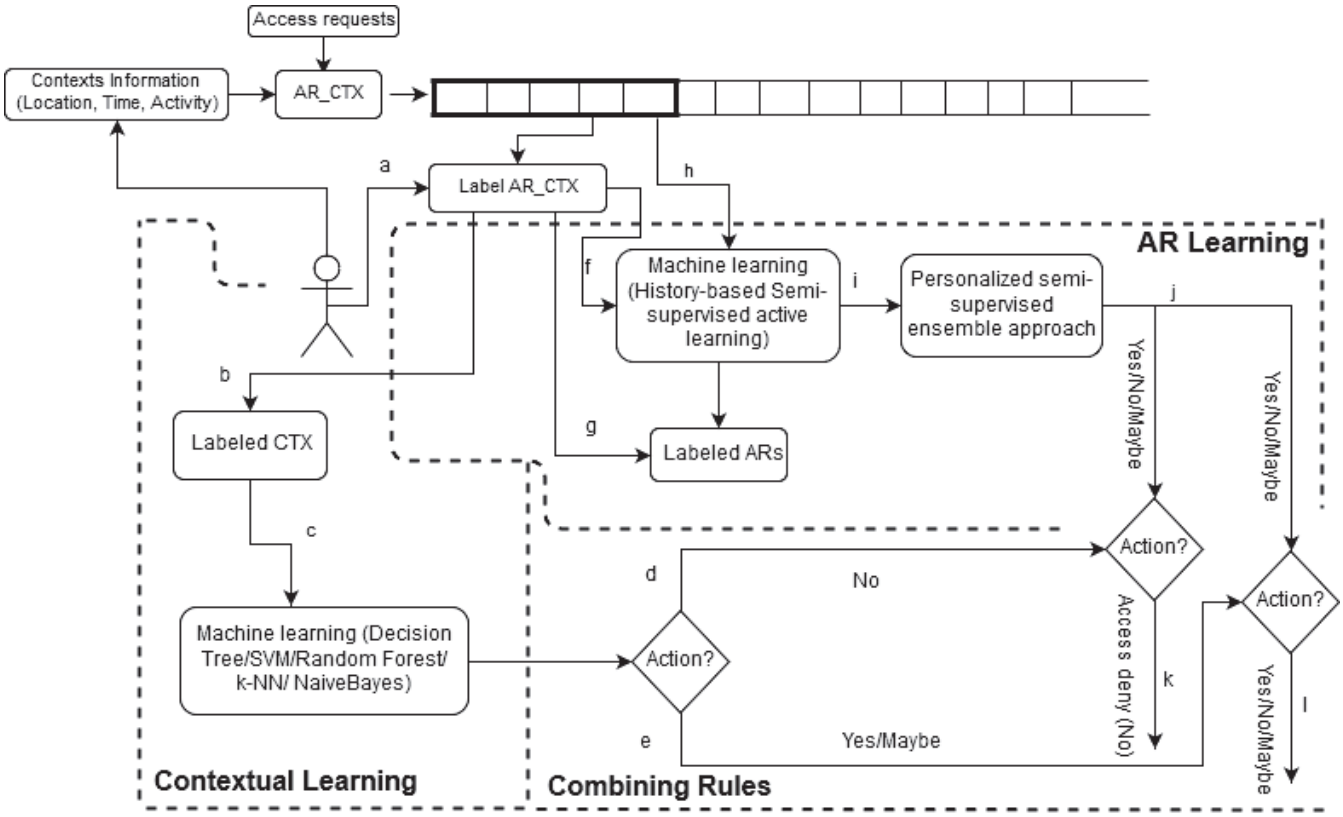

Figure 1: Contextual privacy-aware PDS (CP-PDS)

with CTX so as it can learn for which circumstance PDS owner authorize/deny which type of access requests. With this aim, among AR's fields we select two elements that are more relevant for the access decision, namely: the data being requested to access, and purpose of access with CTX to build two separate classifiers. With this consideration, we want to observe which classifier produce more accurate prediction. As it will be discussed in Section 4, experimental results show that classifiers that exploit CTX plus requested data produce better results. In both the cases, we exploit supervised machine learning to learn and predict user contextual privacy preferences. Supervised learning is a good choice for our setting, since we have a few number of contextual features (time, place, activity) plus an element of AR (requested data/purpose), therefore the combination of data values will not be large. This implies that we do not expect a huge number of labeled training data. Moreover, we exploit different supervised learning approaches to check which one can better work in our setting, namely: Decision Tree (DT), Random Forest (RF), NaiveBayes (NB) [15], Support Vector Machine (SVM) [16], and k-Nearest Neighbors (k-NN) (cfr. Section 4).

3.2 Final decision from AR and contextual learning

To take the final decision on a given access request, we combine the decisions (class labels) suggested by the AR and contextual classifiers. In particular, we believe that there might be some contexts that greatly impact users' decision, like in the previous example where users always denied entertainment services during office hours. To take this into account, we adopt a strategy that, when merging the decisions learned from the context with the one learned from access request, it gives preference to the decision of contextual learning in case it is a deny. As such, we first check the outcome of contextual learning. If the predicted decision is no, we assume that the access request will be automatically denied (see in combining rules part in Figure 1). In the other cases (yes/maybe), CP-PDS exploits the classifier built by the AR learning (i.e., PHBAL, see Section 2) to predict the class label (see in combining rules part in Figure 1). Based on this assumption, we define the combining rules depicted in Table 3.

Figure 1 presents the overall system's architecture. According to our assumption, the CP-PDS receives access re-quests complemented with contextual information AR_CTX. In the initial stage, CP-PDS asks PDS owners about the upcoming access request complemented with contextual information AR_CTX (Figure 1.a). As we described in Section 2 that we exploit the concept of active learning to select the contextual access requests (for good quality labeled training dataset) for labeling by PDS owners, thus in Figure 1, we exploit history based active learning (cfr. Section 2) to select the most uncertain contextual access requests to be labeled by users directly for generating good quality labeled training dataset

TABLE 3
COMBINING RULES

Combining rules		
Outcome from contextual learning	Outcome from AR learning	Final decision
No	Yes	No
No	No	No
No	Maybe	No
Yes	Yes	Yes
Yes	No	No
Yes	Maybe	Yes
Maybe	Yes	Yes
Maybe	No	No
Maybe	Maybe	Maybe

(Figure 1.f and Figure 1.h). Then CP-PDS split the labeled AR_CTX into two parts labeled AR and labeled CTX so as to learn user privacy preference from two different perspective separately. For AR learning, CP-PDS merely use labeled ARs to train classifiers to predict user's privacy preferences (Figure 1.g). Then CP-PDS exploits PHBAL to predict user decision on the upcoming new access request AR (Figure 1.j).

According to the second angle, as depicted in Figure 1(part contextual learning), CP-PDS exploits the users' contextual data labeled CTX. CP-PDS builds a classifier that learns user's contextual privacy preferences according to user opinions based on contextual data CTX on the access requests ARs that come to PDS for getting access. More precisely, as shown in Figure 1, CP-PDS exploits labeled CTX to get feedbacks whether they want to get online services or not based on their contextual data CTX (Figure 1.b) that adopt some fields like requested data fields/access purpose field in order to train classifier (e.g., SVM/Decision Tree etc.) for learning user's contextual privacy preferences (Figure 1.c).

To merge these two decisions for getting final decision in Figure 1(part combining rules), we follow the combining rules as shown in Table 3. According to the rules, if the decision from contextual learning is no (Figure 1.d) then CP-PDS will provide deny as final decision (Figure 1.k). On the other hand, if the decision is yes/maybe (Figure 1.e) then CP-PDS will check the decision of PHBAL (Figure 1.j) and provides this decision as the final decision (Figure 1.l) for the newly arrived access request AR.

In the following example, we consider contextual information with the purpose field to learn user's contextual privacy preference.

Example1.*Let us suppose that the following access request comes to CP-PDS: AR (DC=Slacker\ Radio, d0= {song type, singer name}, St=Music, p=Listening music, o=free). To predict the decision on this access request, CP-PDS first checks the decision from the contextual learning. Assume that the user contextual data is:*

Monday	Morning	Home	Studying

when this access request comes to PDS. Now, if (Day of week= Monday) ^ (Time of day =Morning (6.00-11.59)) ^ (Place=Home) ^ (Activity = Studying) ^ (Purpose = Listening music)) implies the action no and CP-PDS predicts the decision Yes/No/Maybe on the access request AR. Then the final decision will be deny even though the predicted decision on AR (DC=Slacker Radio,d0= {song type, singer name}, St=Music, p=Listening music, o=free) for accessing PDS is any one of Yes/No/Maybe.

4 Experiments

In this section, we illustrate the experiments we have performed to validate the proposed approach. More precisely, in Section 4.3, we test our two step approach by considering two different settings for the learning from contextual data (cfr. Section 3}), namely: 1) contextual data in conjunction with requested data field, and 2) contextual data in conjunction with purpose field, to check which of the two performs better. Moreover, we execute the validation test of our learning methods, with different approaches, that is, sequential and least probability. More precisely, validation test is required for reducing the over-fitting occurs when a learning model learns the noise/randomness along with the samples in the training dataset that negatively impacts the performance of the model on the upcoming new samples. With this aim, we vary the size of labeled training dataset for the learning models and predict the class label on the testing dataset to check which combination of training dataset can produce good accuracy.

TABLE 4
DATASETS USED IN THE EXPERIMENTS

Dataset	User's opinions on contextual access requests	User's opinions on non-contextual access requests	# Labeled access requests	# Users
DS-1	✓	✓	20+20	25
DS-2	✓		60	100

TABLE 5
CONFUSION MATRIX

	Predicted value: Yes	Predicted value: No	Predicted value: Maybe
Actual value: Yes	TP_{Yes}	$E_{Yes, No}$	$E_{Yes, Maybe}$
Actual value: No	$E_{No, Yes}$	TP_{No}	$E_{No, Maybe}$
Actual value: Maybe	$E_{Maybe, Yes}$	$E_{Maybe, No}$	TP_{Maybe}

Then, Section 4.5 presents a comparison of the proposed approach with the one presented in [11] with the aim of assessing the importance of considering contextual information. In Section 4.6, we illustrate the experiment that shows how much user's decision are impacted with contextual data. Finally, in Section 4.7, we present the results about the accuracy on the testing dataset to check whether it will be increased or decreased based user quality in terms of feedback on the training dataset.

4.1 Experimental setting

Datasets. We collected two datasets: one dataset, referred to in what follows as DS-1, contains users' feedback on both non-contextual and contextual access requests separately, and another dataset, named DS-2, that contains users' feedback on contextual access requests only, as shown in Table 4. We generate access requests containing realistic values for the data consumer, service type, requested data, purpose and offer fields. Moreover, we also consider contextual information, i.e., location, time, and activity. More precisely, we have considered: 55 different data consumer profiles; 18 different service types; 42 possible data fields; 21 purposes; offer values ranging from 0\% to 100%, and the following contextual data: 3 different locations (i.e., home, office, and outside), 7 days (e.g., Sunday, Monday. etc.), 4 time slots (i.e., morning, afternoon, evening, and night), and 13 different user's activities (e.g., meeting, driving, etc.). Based on these elements, we randomly generate access requests. Since, we use semi-supervised learning, we need both labeled and unlabeled access requests. Each dataset contains 317 access requests. For dataset DS-1 , we ask labels for 20 non-contextual access requests and 20 contextual access requests, whereas for dataset DS-2, 60 access requests are labeled, whereas the remaining ones are used as unlabeled data.

Evaluators. For access request labeling, we developed a web application, and we use a crowdsourcing platform for user engagement. We have recruited 125 participants from the Microworker crowdsourcing platform of different nationalities, ages, and educational levels. For dataset DS-1, we recruit 25 users to label 20 non-contextual and 20 contextual access requests. We exploit this dataset for checking the impact of contextual information in access decisions (see Sections 4.4, and 4.6). For dataset DS-2 we recruit 100 workers, and we used it for the experiments in Sections 4.3, 4.5, and 4.7.

To ensure good quality of the jobs submitted to Microworker, we have selected only workers with the best rating according to the Microworker platform. As further quality check, we measured the time each participant devoted to the labeling task and, if this is less than a reasonable time, we remove the participant. For dataset DS-2, we have presented 72 access requests to each participant. More particularly, 60 access requests have been used as labeled training dataset; 7 access requests are used for testing the performance of the proposed approach, whereas 5 access requests are used for checking the quality of the job execution (see Section 4.7 for more details). The same approach has been applied to dataset DS-1, except for the number of requested labels. More precisely, we have presented 52 access requests to each participant: 20 non-contextual access requests, 20 contextual access requests, 7 access requests for testing purpose, whereas 5 access requests are used for checking the quality of user feedbacks.

4.2 Evaluation metrics

In order to measure the effectiveness of the proposed approach, we use the traditional confusion matrix. Since we consider classes with 3 labels (yes, no, maybe), we exploit a 3X3 confusion matrix, presented in Table 5. More precisely, columns of the matrix represent the predicted value for a class, rows represent possible actual values, and an element

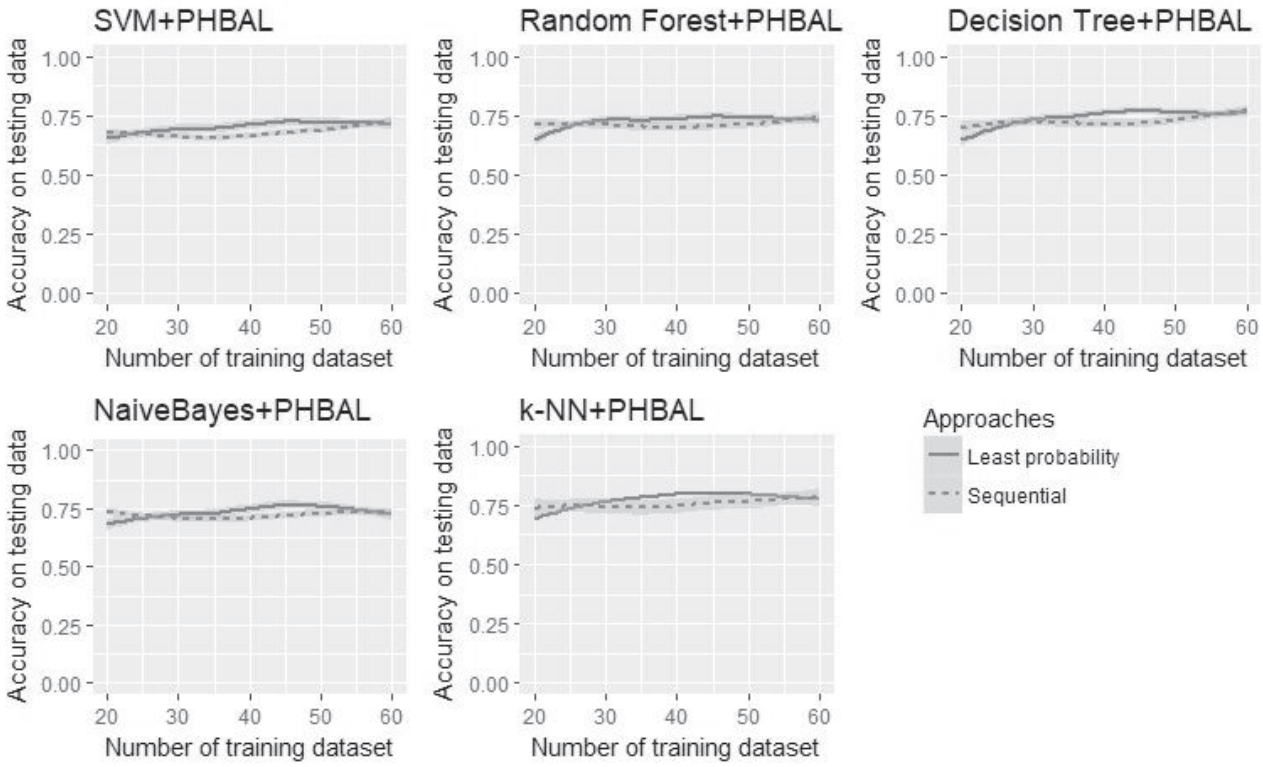

Figure 2: Accuracy on testing dataset using contextual based access requests (with requested data)

identified by row and column specifies the type of error, if any, in labeling an item whose real value is specified in the row with the label corresponding to the column. TP_Yes, TP_No, and TP_Maybe represent the true positive values, whereas the other notations represent error values. From the confusion matrix, we define the evaluation metric, accuracy as the ratio of total number of true positives (TPs) to total number of samples.

4.3 Context with purpose vs context with requested data

As we have explained in Section 3, we believe that purpose and context data are the more relevant field in conjunction with contextual data to learn user privacy preferences. Thus, in this section we report the experiments we have done to check which one provides better accuracy. With this aim, we run several experiments. Moreover, since we use machine learning approaches to learn user privacy preference decisions, thus we need to consider strategies to minimize the prediction errors. As we know, machine learning has pitfall such as over-fitting [17], [18]. To deal with this problem, we select samples by varying the size of the training dataset, to check which one produces better accuracy on testing dataset. Prior researches have defined different approaches to deal with this issue [19]-[21]. Therefore, we experiment the following approaches, the first is the sequential based approach, which is a traditional approach to reduce the over-fitting problem in machine learning [21], [22]. This approach sequentially considers a set of labeled datasets of increasing size, and check the accuracy on the testing dataset. However, we have also considered an alternative approach, that we called least probability based approach. With this approach, we want to select the good quality labeled training dataset for learning models. To do so, we select the labeled training dataset which has least probability distance among classes from the labeled training pool and increasing the size based on the probability distance, and check the accuracy on the testing dataset.

Figures 2 and 3 report the accuracy on the testing dataset when different supervised learning approaches (e.g., SVM, NaiveBayes etc.) learn user contextual privacy preferences by considering contextual data in conjunction with the requested data and the purpose field, respectively.

Figure 2 shows that k-NN+PHBAL outperforms the others with the least probability approach, when contextual information are considered in conjunction with the requested data field. More precisely, it provides around 81% accuracy on the testing dataset compared to other approaches with a training datasets of size 45, as shown in Table 6. In contrast, Figure 3 shows that DT+PHBAL, and RF+PHBAL outperform the other approaches with least probability. when the purpose field is considered in conjunction with contextual information. Moreover, it provides around 77% accuracy on the testing dataset, as shown in Table 7. These experiment shows that k-NN+PHBAL produces good accuracy on testing dataset when we consider contextual data in conjunction with the requested data field. Moreover, the experiments shows that least probability approach produces better accuracy on testing dataset than the sequential approach.

TABLE 6
ACCURACY ON TESTING DATASET BY LEARNING MODELS USING CONTEXT WITH REQUESTED DATA

Learning Models	Size of the training dataset achieving the highest accuracy	Accuracy on testing dataset
SVM+PHBAL	45	73.28%
RF+PHBAL	45	75.42%
DT+PHBAL	45	78.57%
NB+PHBAL	45	77.57%
kNN+PHBAL	45	81.00%

TABLE 7
ACCURACY ON TESTING DATASET BY LEARNING MODELS USING CONTEXT WITH PURPOSE

Learning Models	Size of training dataset achieving the highest accuracy	Accuracy on testing dataset
SVM+PHBAL	60	73.85%
RF+PHBAL	45	76.28%
DT+PHBAL	50	76.57%
NB+PHBAL	45	76.42%
kNN+PHBAL	50	74.28%

4.4 Non-contextual vs contextual based access requests

In these experiments, we compare the approach presented in this paper with the one proposed in [11], which does not leverage on contextual information. Figure 4 shows the results on the testing dataset. Moreover, in the first experiment as shown in Figure 4, we consider the contextual information with purpose field. Figure 4 shows that all learning approaches as such DT+PHBAL, RF+PHBAL etc. produce good accuracy on testing dataset than non-contextual based access request.

In the second experiment, we consider contextual information with requested data. The results are shown in Figure 5. From the figure it can be seen that all the considered learning approaches have better accuracy over the testing dataset than the analogous learning approaches over non-contextual based access request. Thus, the experiments confirm that when we consider contextual based access request then learning approach produce better accuracy.

4.5 Single vs two step learning process

In this experiment, we experiment an alternative approach, w.r.t the one proposed in this paper that, instead of applying a two step learning approach, feed a single step learning with context and non-contextual information together, instead of exploiting two step learning, as explained in Section 3.

As shown in Figure 7, the accuracy on the testing dataset is around 74% produced by PHBAL, whereas, from Figure 3, we can see that k-NN+PHBAL produces around 81% accuracy. Therefore, these experiments show that the 2 step process proposed in this paper is more effective in learning user privacy preferences.

4.6 Context impacts on users decisions

In this experiment, we investigate how many access control decisions users have changed due to the consideration of contextual information. The experimental results reported in Figures 4 and 5 have already shown that a better accuracy is achieved when we consider contextual based access requests. In this second set of experiments, we want to show how many access control decisions are driven by context data. For doing so, we first ask user opinions on non-contextual access requests and then we ask again the user feedback on the same access request with the associated contextual infomation. Figure 8 shows the results. More particularly, the experiment shows that 66.87% decisions are

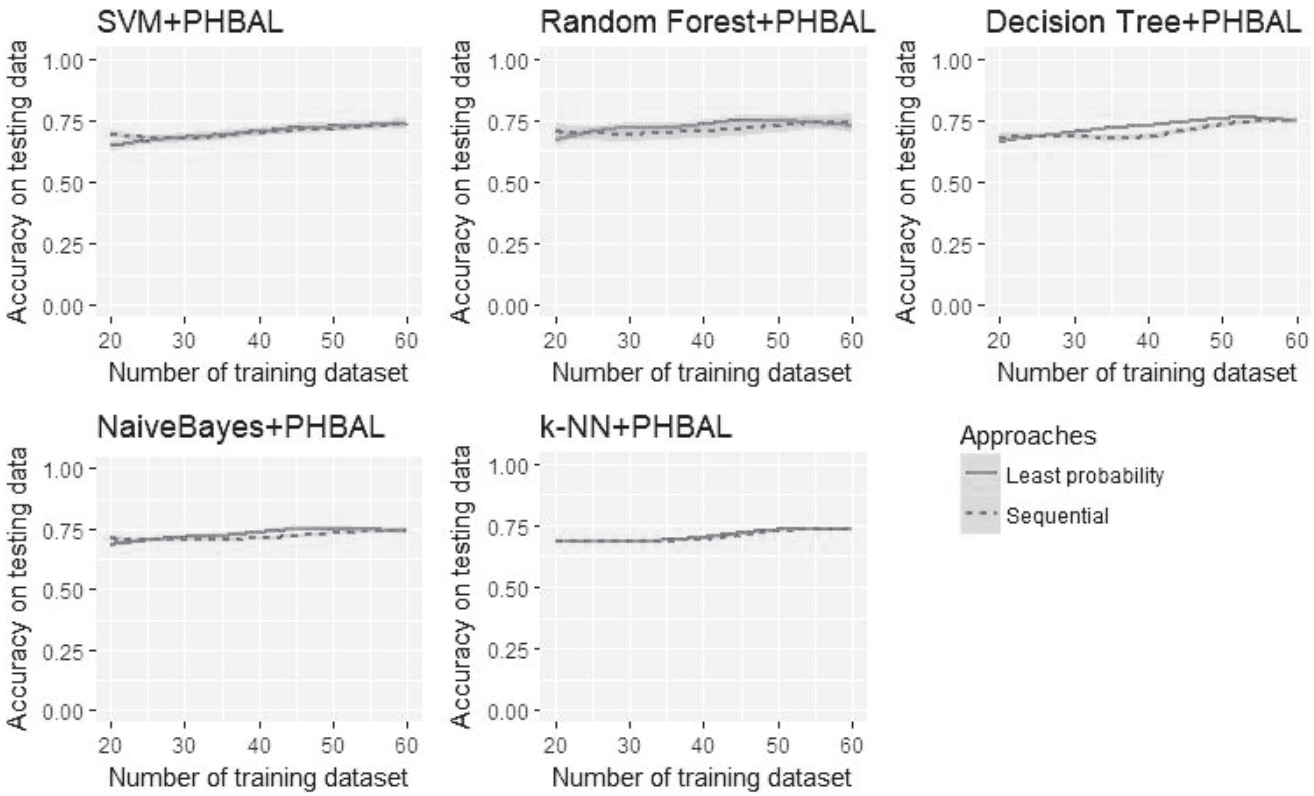

Figure 3 : Accuracy on testing dataset using contextual based access requests (with purpose)

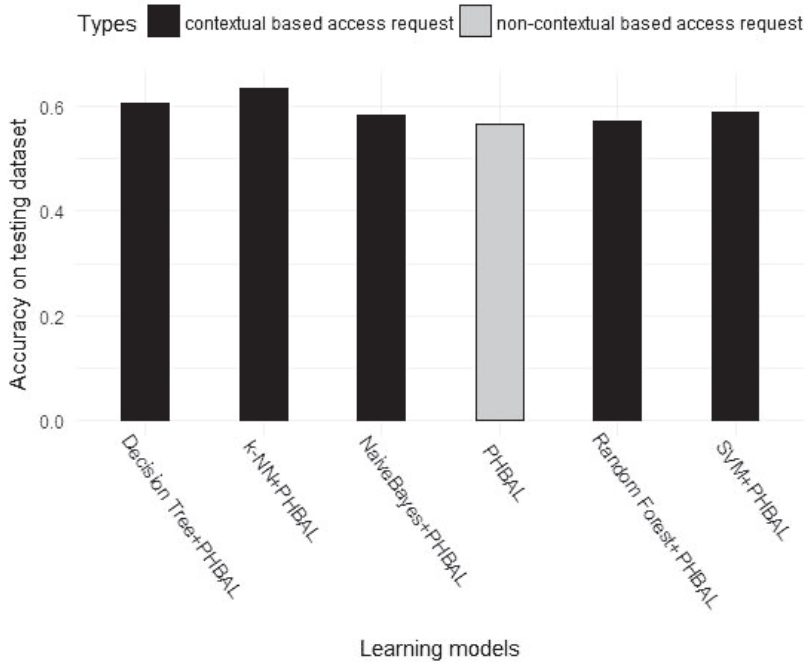

Figure 4 : Accuracy on testing dataset compared between non-contextual and contextual (with purpose) based access requests

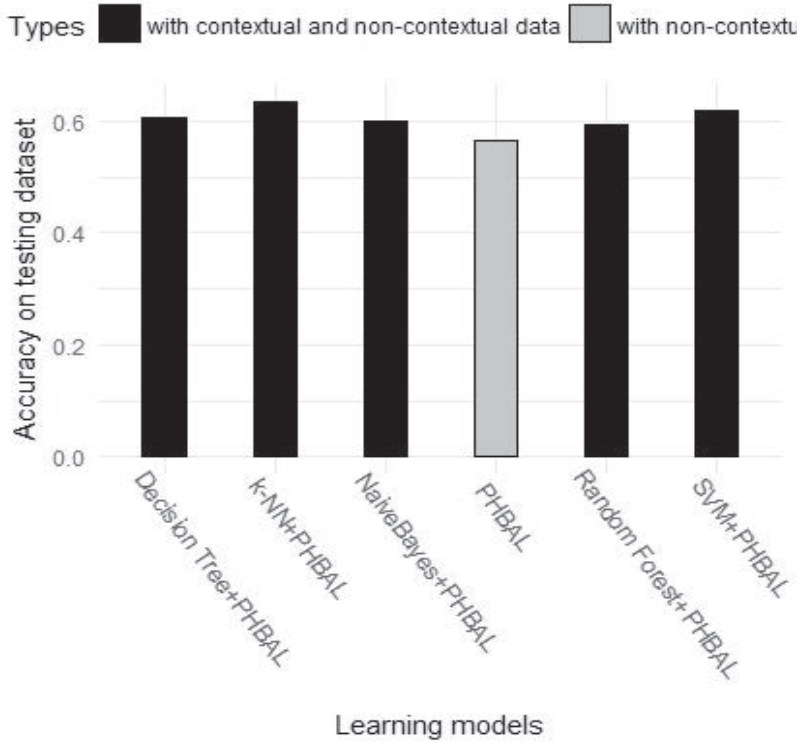

Figure 5 : Accuracy on testing dataset compared between non-contextual and contextual (with requested data) based access requests

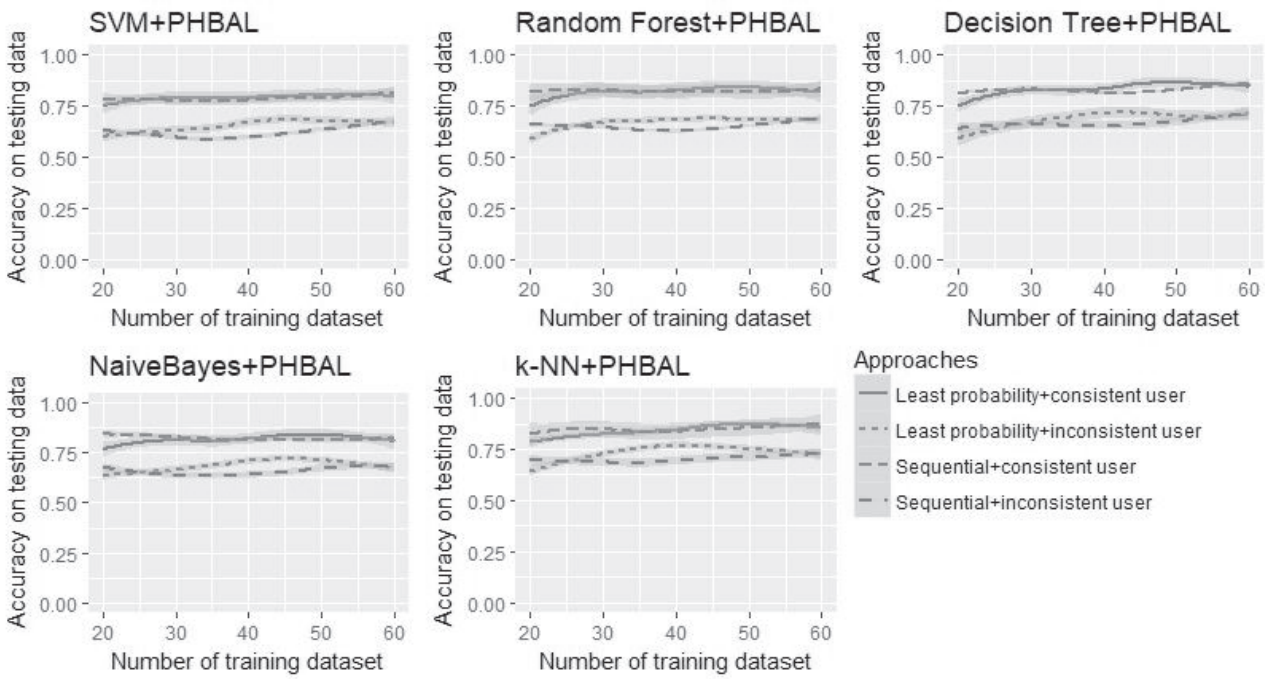

Figure 6 : Accuracy on testing dataset for consistent and inconsistent users (requested data field with contextual information)

the same, but, interestingly, 33.13% of the access request decisions are changed by users when we consider contextual information.

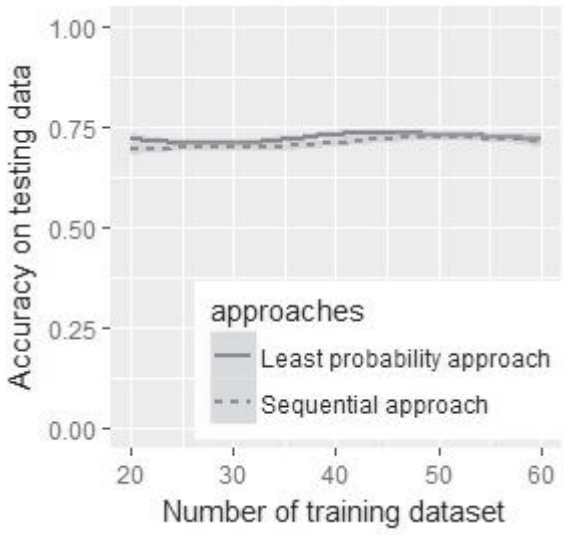

Figure 7: Accuracy on testing dataset

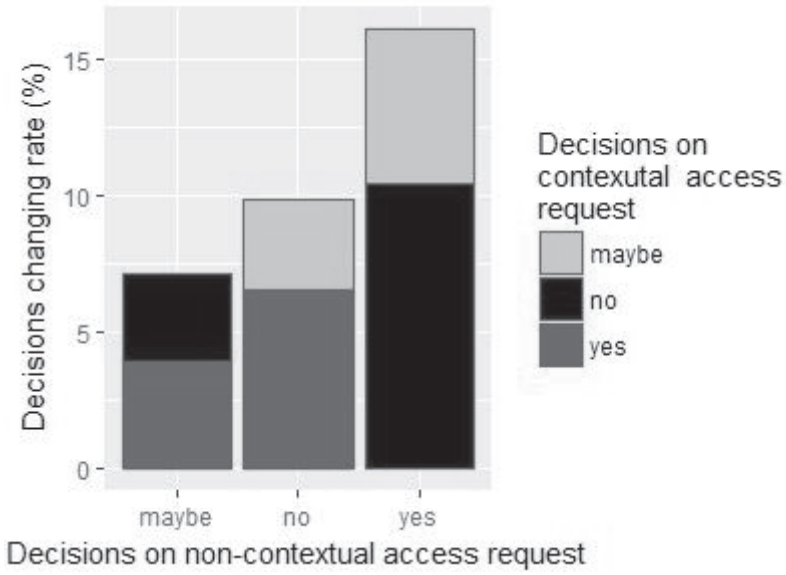

Figure 8: Users' decisions changing rate based on contextual based access requests

4.7 Participants quality

Clearly, the output of any machine learning approach depends on the quality of the user's input on the training dataset. Therefore, we are interested to investigate how a badly labeled training dataset impacts the final decision on the testing dataset. Thus, we set some strategies to identify consistent and inconsistent evaluators. First, three of the access requests are presented twice to evaluators, to check if they are always marked with the same label. Based on the assigned labels, we can judge whether the evaluator is consistent or not in his/her decisions, which gives us a measure of the quality of his/her jobs. Second, we have inserted two access requests in the first phase (e.g., among the first nine access requests) which contain a requested data field which is inconsistent with the requested access purpose and service. For example, we ask a label for an access request on these data {traveling date, traveling time, From (starting place), To (destination place), etc.,}, having a service purpose issuing a loan. We expect that, in case of an inconsistent access request, a participant that carefully reads the request will assign a deny label. Therefore, we consider a participant as consistent if he/she behaves correctly w.r.t. the above described checks. In our experiment, we show that 43% users have given feedback on access requests in consistent manner, whereas 57% users are inconsistent. Figure 6 shows the accuracy on testing dataset for consistent users and non-consistent users where it confirms kNN+PHBAL produces better result than other approachers as shown in Figure 2.

5 Related Work

Ensuring user privacy preferences in online activities is a challenging issue due to lack of efficient privacy preferences model [23]. The main reason behind this incident is that different domains have different purposes in term of ensuring privacy. However, in order to address the growing problem of spreading personal information on the Internet, researchers have begun to offer a variety of proposals and mitigate the problem by proposing an approach call PDS. De Montjoye et al [3] presented open PDS/Safe Answers mechanism that allows individuals to collect, and store their personal data in PDS as well as give access to their meta-data to third parties based on privacy policies. The framework defines a mechanism for returning to third parties only aggregated answers, based on their questions, instead of raw data. Although this framework never shares raw data, there is room for malicious applications to infer more information through a specific sequence of questions-answers, which can eventually breach user privacy. Nowadays, researchers also try to propose models for user-centric storage in the cloud domain based on the concept of PDS, where data are stored and controlled by users.

For instance, Oort [24] is a user-centric cloud storage system that allows users to select which applications can access to their own data and to whom their data can be shared with. Oort achieves this goal by considering global queries which find and combine the relevant data fields to share with relevant users. Sieve [25] allows user to upload encrypted data to a single cloud storage which is not trusted. It utilizes key-homomorphic scheme to provide cryptographically enforced access control. Amber [26] has proposed an architecture that decouples users' data from application. Moreover users can choose applications that facilitate with global queries to find their data but it does not mention either how the global queries work or how the application providers interact with. Mortier et al. [27] have proposed a trusted platform called Databox, which can manage personal data by a fine grained access control mechanism but do not focus on policy learning. [40] presents an architecture for privacy-preserving personalized services and provides mechanisms for managing privacy for the users and plays the role for checking private data flow. Recently, [28][41] proposed a Block chain-based Personal Data Store (BC-PDS) framework which leverage on BlockChain to secure the storage of personal data. The fact is that these approaches do not consider user's contextual data in term of implementing privacy preference mechanisms in user-centric storage system. Therefore, these approaches could not capture all the aspects of users concern regarding privacy completely. To fill up this gap, we implement privacy preference approach based on user's contextual data with access request elements that comes to PDS.

However, researchers already implemented contextual based privacy preferences in smart-phone environments. For instance in [29]-[32] proposed mechanisms to predict permission decisions at runtime that relies on user's contextual information in mobile platforms, whereas in [33]-[35] proposed user's location sharing privacy preferences by considering contextual information. L. Yuan et al. [36] presented a privacy-aware model for photo sharing based on machine learning that exploits contextual information. T. liang et al. [37] developed a learning approach that recommends context-aware app by utilizing a tensor-based framework so as to effectively integrate user's preferences, app category information and multi-view features. In [38], authors presented a privacy preference model for helping users to manage their privacy in context-aware systems in term of sharing location on the basis of the general user population using crowd-sourcing architecture.

Bilogrevic et al. [39] presented a privacy preference framework that (semi-)automatically predicts sharing decision, based on personal and contextual features. The authors only try to focus on general information sharing with nearby people such as location.

However, in this paper, we do not consider only the contextual features but also all aspects of personal information as such health-care data, location data, e commerce data etc. to implement user's contextual based privacy preferences in PDS. To do so, we exploit machine learning tools to assist user's to protect their privacy preferences on personal data from unauthorized in a semi-automated fashion based on user's contextual information.

6 Conclusion

In this paper, we propose a contextual privacy-aware framework for PDS (CP-PDS) for helping users to manage their privacy preference in PDS. More particularly, the proposed CP-PDS that decides in a semi-automated fashion whether or not to authorize access requests based on user's contextual information and preferences. To further improve the performance of CP-PDS, we are interested to investigate with more user's contextual data for better understanding of user's feelings in term of authorizing personal data from PDS. Furthermore, we plan to extent the usability of CP-PDS in broader space as such in IoT scenario or cloud computing services.

References

[1] M. Vescovi, C. Perentis, C. Leonardi, B. Lepri, and C. Moiso, "My data store: toward user awareness and control on personal data," in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. ACM, 2014, pp. 179–182.

[2] C. Moiso, F. Antonelli, and M. Vescovi, "How do i manage my personal data?-a telco perspective." in DATA, 2012, pp. 123–128.

[3] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland, "openpds: Protecting the privacy of metadata through safeanswers," PloS one, vol. 9, no. 7, p. e98790, 2014.

[4] R. Want, T. Pering, G. Danneels, M. Kumar, M. Sundar, and J. Light, "The personal server: Changing the way we think about ubiquitous

computing," Ubicomp 2002: Ubiquitous Computing, pp. 223–230, 2002.
[5] T. Allard, N. Anciaux, L. Bouganim, Y. Guo, L. Le Folgoc, B. Nguyen, P. Pucheral, I. Ray, I. Ray, and S. Yin, "Secure personal data servers: a vision paper," Proceedings of the VLDB Endowment, vol. 3, no. 1-2, pp. 25–35, 2010.
[6] D. A. Albertini, B. Carminati, and E. Ferrari, "Privacy settings recommender for online social network," in Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on. IEEE, 2016, pp. 514–521.
[7] M. Madejski, M. Johnson, and S. M. Bellovin, "A study of privacy settings errors in an online social network," in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on. IEEE, 2012, pp. 340–345.
[8] B. C. Singh, B. Carminati, and E. Ferrari, "A risk-benefit driven architecture for personal data release," in Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on. IEEE, 2016, pp.40–49.
[9] M. C. Mont and R. Thyne, "A systemic approach to automate privacy policy enforcement in enterprises," in International Workshop on Privacy Enhancing Technologies. Springer, 2006, pp. 118–134.
[10] B. C. Singh, B. Carminati, and E. Ferrari, "Learning privacy habits of pds owners," in Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on. IEEE, 2017, pp. 151–161.
[11] B. C. Singh, B. Carminati, and E. Ferrari, "Privacy-aware personal data storage (p-pds): Protecting user privacy from external applications," 2018.
[12] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer, L. Cranor, and N. Sadeh, "Privacy expectations and preferences in an iot world," in Symposium on Usable Privacy and Security (SOUPS), 2017.
[13] S. Chakraborty, C. Shen, K. R. Raghavan, Y. Shoukry, M. Millar, and M. B. Srivastava, "ipshield: A framework for enforcing context-aware privacy." in NSDI, 2014, pp. 143–156.
[14] B. Settles, "Active learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6, no. 1, pp. 1–114, 2012.
[15] J. Zhang, D.-K. Kang, A. Silvescu, and V. Honavar, "Learning accurate and concise naïve bayes classifiers from attribute value taxonomies and data," Knowledge and Information Systems, vol. 9, no. 2, pp. 157–179,2006.
[16] D.-C. Li and C.-W. Liu, "Extending attribute information for small data set classification," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 452–464, 2012.
[17] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," Journal of Machine Learning Research, vol. 11, no. Jul, pp. 2079–2107, 2010.
[18] S. Whiteson, B. Tanner, M. E. Taylor, and P. Stone, "Protecting against evaluation overfitting in empirical reinforcement learning," in Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011, IEEE Symposium on. IEEE, 2011, pp. 120–127.
[19] F. Liebgott and B. Yang, "Active learning with cross-dataset validation in event-based non-intrusive load monitoring," in Signal Processing Conference (EUSIPCO), 2017 25th European. IEEE, 2017, pp. 296–300.
[20] A. Ali, R. Caruana, and A. Kapoor, "Active learning with model selection." in AAAI, 2014, pp. 1673–1679.
[21] M. Sugiyama and N. Rubens, "A batch ensemble approach to active learning with model selection," Neural Networks, vol. 21, no. 9, pp. 1278–1286, 2008.
[22] D. J. MacKay, "Bayesian interpolation," Neural computation, vol. 4, no. 3, pp. 415–447, 1992.
[23] H. Nissenbaum, "A contextual approach to privacy online," Daedalus, vol. 140, no. 4, pp. 32–48, 2011.
[24] T. Chajed, J. Gjengset, M. F. Kaashoek, J. Mickens, R. Morris, and N. Zeldovich, "Oort: User-centric cloud storage with global queries," 2016.
[25] F. Wang, J. Mickens, N. Zeldovich, and V. Vaikuntanathan, "Sieve: Cryptographically enforced access control for user data in untrusted clouds." in NSDI, 2016, pp. 611–626.
[26] T. Chajed, J. Gjengset, J. Van Den Hooff, M. F. Kaashoek, J. Mickens, R. Morris, and N. Zeldovich, "Amber: Decoupling user data from web applications." in HotOS, vol. 15, 2015, pp. 1–6.
[27] G. C. Caw R. Mortier, J. Zhao, J. Crowcroft, Q. Li, L. Wang, H. Haddadi, Y. Amar, A. Crabtree, J. Colley, T. Lodge et al., "Personal data management with the databox: whats inside the box?" 2016.
[28] Z. Yan, G. Gan, and K. Riad, "Bc-pds: Protecting privacy and selfsovereignty through blockchains for openpds," in Service-Oriented System Engineering (SOSE), 2017 IEEE Symposium on. IEEE, 2017, pp. 138–144.
[29] K. Olejnik, I. Dacosta, J. S. Machado, K. Huguenin, M. E. Khan, and J.-P. Hubaux, "Smarper: Context-aware and automatic runtimepermissions for mobile devices," in Security and Privacy (SP), 2017, IEEE Symposium on. IEEE, 2017, pp. 1058–1076.
[30] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov, "Android permissions remystified: A field study on contextual integrity." in USENIX Security Symposium, 2015, pp. 499–514.
[31] L. Tsai, P. Wijesekera, J. Reardon, I. Reyes, S. Egelman, D. Wagner, N. Good, and J.-W. Chen, "Turtle guard: Helping android users apply contextual privacy preferences," in Symposium on Usable Privacy and Security (SOUPS), 2017.
[32] P. Wijesekera, J. Reardon, I. Reyes, L. Tsai, J.-W. Chen, N. Good, D. Wagner, K. Beznosov, and S. Egelman, "Contextualizing privacy decisions for better prediction (and protection)," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018, p. 268.
[33] E. Toch, J. Cranshaw, P. H. Drielsma, J. Y. Tsai, P. G. Kelley, J. Springfield, L. Cranor, J. Hong, and N. Sadeh, "Empirical models of privacy in location sharing," in Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM, 2010, pp. 129–138.
[34] J. Xie, B. P. Knijnenburg, and H. Jin, "Location sharing privacy preference: analysis and personalized recommendation," in Proceedings of the 19th international conference on Intelligent User Interfaces. ACM, 2014, pp. 189–198.
[35] Y. Zhao, "Recommending privacy preferences in location-sharing services," Ph.D. dissertation, University of St Andrews, 2017
[36] L. Yuan, J. Theytaz, and T. Ebrahimi, "Context-dependent privacyaware photo sharing based on machine learning," in IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, 2017, pp. 93–107.
[37] T. Liang, L. He, C.-T. Lu, L. Chen, P. S. Yu, and J.Wu, "A broad learning approach for context-aware mobile application recommendation," arXiv preprint arXiv:1709.03621, 2017.
[38] E. Toch, "Crowdsourcing privacy preferences in context-aware applications," Personal and ubiquitous computing, vol. 18, no. 1, pp. 129–141, 2014.
[39] I. Bilogrevic, K. Huguenin, B. Agir, M. Jadliwala, M. Gazaki, and J.-P. Hubaux, "A machine-learning based approach to privacy-aware information-sharing in mobile social networks," Pervasive and Mobile Computing, vol. 25, pp. 125–142, 2016.
[40] M. S. Rahman, A. Basu, T. Nakamura, H. Takasaki, and S. Kiyomoto, , "PPM: Privacy Policy Manager for Home Energy Management, " Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications (JoWUA), vol. 9, no. 2, pp. 42–56, 2018.
[41] B. G. Rohan "Personal data vault management system and secured data access to service provider using blockchain technology", (Doctoral dissertation, University of Massachusetts Lowell), 2018.

Bikash Chandra Singh successfully completed BSc and MSc degree from the department of Information & Communication Engineering at Islamic University, Bangladesh. Currently, he is pursuing the PhD degree in Computer Science from University of Insubria, Italy. Moreover, he is a faculty member at Islamic University, Bangladesh. He has published a number of research papers in the field of data privacy and security. His research interests are related to big data analysis, data privacy \& security, machine learning, cloud computing, and Internet of things (IoT).

Md Sipon Miah received his BSc (Hon's), and MSc in Information and Communication Engineering (ICE) from the Islamic University (IU), Kushtia-7003, Bangladesh, in 2006 and 2007, respectively. Since 2010, he has been with the Department of Information and Communication Engineering (ICE), in the Islamic University (IU), Kushtia-7003, Bangladesh. He is currently an Associate Professor in the same Department. Sipon is currently pursuing a Structured Ph.D. in computer science in the Department of Information Technology (IT), National University of Ireland Galway (NUIG), Galway, Ireland. In 2016 Sipon was awarded the prestigious Hardiman Scholarship. His research interests include Spectrum Sensing, Energy Harvesting, MU-MIMO based Cognitive Radio Networks and Massive MIMO based Cognitive Radio Networks. biography appears here. Degrees achieved followed by current employment are listed, plus any major academic achievements.

Tapan Kumar Godder received his BSc (Hon's), and MSc in Information and Communication Engineering (ICE) from the Islamic University (IU), Kushtia-7003, Bangladesh the Bachelor's, Master's and MPhil degree in Applied Physics & Electronics from Rajshahi University, Rajshahi in 1994,1995 and 2007, respectively. He is currently full Professor in the department of ICE, Islamic University, Kushtia-7003, Bangladesh. He has published several papers in international and national journals. His areas of interest include internetworking, AI & mobile communication.

M. Mahbubur Rahman received his BSc (Hon's) and MSc in Physics from Rajshahi University, in 1983 and 1994 respectively and PhD degree in Computer Science & Engineering in 1997 from Rajshahi University. He is courrently a Professor in the department of ICE, Islamic University, Kushtia, Bangladesh. He has a good number of papers published in international and national journals. His research interest includes internetworking, AI & mobile communication.