Easing the burden of setting privacy preferences: a machine learning approach

Toru Nakamura[1], Shinsaku Kiyomoto[1], Welderufael B. Tesfay[2], and Jetzabel Serna[2]

[1] KDDI R&D Laboratories Inc., Saitama, Japan
{tr-nakamura, kiyomoto}@kddilabs.jp
[2] Chair of Mobile Business and Multilateral Security, Goethe University Frankfurt
Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany
{welderufael.tesfay, jetzabel.serna}@m-chair.de

Abstract. Setting appropriate privacy preferences is both a difficult and cumbersome task for users. In this paper, we propose a solution to address users' privacy concerns by easing the burden of manually configuring appropriate privacy settings at the time of their registration into a new system or service. To achieve this, we implemented a machine learning approach that provides users personalized privacy-by-default settings. In particular, the proposed approach combines prediction and clustering techniques, for modeling and guessing the privacy profiles associated to users' privacy preferences. This approach takes into consideration the combinations of service providers, types of personal data and usage purposes. Based on a minimal number of questions that users answer at the registration phase, it predicts their privacy preferences and sets an optimal default privacy setting. We evaluated our approach with a data set resulting from a questionnaire administered to 10,000 participants. Results show that with a limited user input of 5 answers the system is able to predict the personalised privacy settings with an accuracy of 85%.

Keywords: Privacy; privacy-by-default; privacy policy; privacy preferences;

1 Introduction

Default privacy settings play a major role in restricting or revealing personally identifiable information of online service users. On the one hand, highly restrictive privacy settings limit the information sharing utilities of online services, while on the other hand less restrictive privacy settings can significantly damage the privacy of users. The best case scenario is to have a personalised privacy and utility optimal preference setting that meets the user's particular needs. The challenge is that service providers do not provide privacy optimal and tailored preference settings by default, and most users are not capable of establishing such settings by themselves. The extent to which users are capable of setting their preferences depends on their skill level and understanding of the setting [1]. According to [2], typical preferences, e.g., those set by social network sites such as Facebook on behalf of users, meet the expectations of users only 37% times. Moreover, authors in [3] stated that users exhibit a privacy paradox behaviour, in that, despite their increasing privacy concerns most of them are reluctant to take further steps and alter the default settings set by the service providers that do not take individual preferences into account. Furthermore, not having properly and optimally set privacy preferences greatly increases the privacy concerns of end users. In particular, the new direction of commercial services such as O2O (Online-to-Offline), are attended by a series of privacy concerns that have become a serious issue, mainly due to the expansion of service collaborations [4, 5]. In this regard, situations such as being diverted to services users were previously totally unaware of having a relationship with, have resulted in even more privacy concerns among users. An example of this is Internet ads. Studies conducted by [6, 7], have suggested that Internet ads, which are personalised through the use of private data, may be responsible for leaking users' private information. As a result, privacy is an increasingly important aspect that might hinder users' willingness to publish personal data. Therefore, to properly address users' privacy concerns, they need to be aware of what data are being collected and for what purposes. To accomplish this aim, access control mechanisms based on users' privacy preferences are a key function for providing personal data without creating anxiety in users. However, it is difficult to manually configure appropriate privacy settings where the combinations of service providers, types of personal data, and the purposes to which personal data are put, become huge.

Hence, it is important to simplify this task of setting privacy-preserving default preferences by providing tailoring mechanisms that will address individual privacy concerns, and provide personalised privacy settings to users.

In this paper, we propose an intelligent mechanism for automatic generation of personalised privacy settings. It aims to provide optimised privacy preference settings by default to support users' online interactions, while minimising individual's privacy risks. To this aim, our proposed approach consists of delivering a minimal set of questions to each user at the time of registration to a new service, and from the users' answers predict the personalised default privacy settings for each user. We consider a set of 80 different parameters associated with different types of data for 16 different utilisation purposes. First, we formulated a questionnaire that allowed us to find out the privacy concerns of users, and their acceptability of providing personal data for different purposes. The questionnaire was carried out in the form of web survey with approximately 10,000 participants. Second, we propose a guessing scheme based on machine learning. The basic scheme implements *SVM* (*Support Vector Machine*). In this scheme we first generate the SVM models for a full set of settings by considering only a few answers for the privacy settings. Finally, in order to improve the overall performance, we propose an extension of the basic scheme by using SVM combined with clustering algorithms.

The rest of the paper is organised as follows, Section 2 gives an overview of privacy policy management. Section 3 describes the main methodology used in this research work. Section 4 introduces the proposed approach, which is evaluated in Section 5. Section 6 discusses the advantages and limitation of this approach. Section 7 provides an overview of related work in the area of privacy preferences while Section 8 draws the main conclusions and points out future directions of research.

2 Privacy policy management

In this section we discuss the different dimensions of privacy policy settings and management tools.

Privacy policy management has become the common approach adopted by online service providers in order to specify, communicate and enforce privacy rights of online users. In this model, each online service provider delivers a privacy policy associated to each of its online services, and, users are required to read and accept the privacy policy right before starting to use the corresponding service. If a user does not agree with the privacy policy of the service, the user simply cannot use the service. Furthermore, because it is presumable that users would need to check a large number of privacy policies, it becomes a tedious task that most users find difficult to understand. An experimental study conducted by Acquisti and Grossklags [8], demonstrated that, when confirming privacy policies, users lack knowledge about technological and legal forms of privacy protection. Their observations suggested that several difficulties obstruct individuals in their attempts to protect their own private information, even those concerned about and motivated to protect their privacy. These findings were reinforced by authors in [9] who also supported the presumption that users are not familiar with technical and legal terms related to privacy. Moreover, it was suggested that users' knowledge about privacy threats and technologies that help to protect their privacy is inadequate [10]. Furthermore, Solove also suggested that, even though, privacy law has been relying too heavily upon the privacy self-management model [11], this model simply could not achieve its objectives, and stated that, it has been pushed beyond its limits.

In this regard, the Platform for Privacy Preferences Project (P3P) [12, 13] was designed to enable online services to express their privacy policies in a standard format. In this way, privacy policies could be retrieved automatically and interpreted easily by user agents. The user agent modules will then enable users to be informed of site practices and to try to automate the decision-making process. In this direction, the Privacy Bird [14, 15] was designed to automatically retrieve the P3P policies of a web site. Other approaches to describe privacy policies were also introduced in [13, 18, 19]. Backes *et al.* presented a comparison of enterprise privacy policies using formal abstract syntax and semantics to express the policy contents [17], while Tondel and Nyre [20] proposed a similarity metric for comparing machine-readable privacy policies. Furthermore, a privacy policy checker for online services was introduced by authors in [21]. The checker compared the user privacy policy with the provider privacy policy and then automatically determined whether the service could be used. However, according to authors in [22] this type of approaches resulted in inadequate user acceptance for real world scenarios.

Worth to note that interpreting a privacy policy is just the first step, afterwards, users need to manually configure a set of privacy settings designed to match a given privacy policy. Furthermore, even though some browsers have a privacy module that tries to match privacy preferences to privacy policies, in practice, it has not been widely adopted by online services [16]. That is, mainly due to its complex policy definitions and because the module is to be implemented only on web browsers. Thus, until recently, many research works have focused on studying privacy policy specification, while fewer studies have dedicated efforts to simplify the task of setting privacy preferences, which is the main focus of our research work.

3 Methodology

This section introduces the methodology used for data collection and provides and insight of the distribution of participants and their privacy preferences.

3.1 Data Collection

In this study, we first have developed a questionnaire that allowed us to learn about users' willingness to share different types of personal data, considering different services and utilisation purposes, and consequently allowed us to map those preferences to the user privacy preference setup. For this purpose, we first identified different kinds utilisation purposes (Table 1) and personal data (Table 2), as defined in P3P [12].

Table 1. Utilization purposes

No.	Data purpose
A	Providing the service
B	System administration
C	Marketing
D	Behaviour analysis
E	Recommendation

We published the questionnaire as an online survey and collected the answers from 10,000 participants recruited by a research service company. While the main goal of the questionnaire was to identify users' privacy preferences, we also raise privacy awareness by delivering information about the potential benefits and risks of providing access to certain data for each service.

Table 2. Kinds of personal data

No.	Data type
1	Addresses and telephone numbers
2	Email addresses
3	Service accounts
4	Purchase records
5	Bank accounts
6	Device information (e.g., IP addresses, OS)
7	Browsing histories
8	Logs on a search engine
9	Personal info (e.g., age, gender, etc.)
10	Contents of email, blog, twitter etc.
11	Session information (e.g., Cookies)
12	Social information (e.g., religion, volunteer records)
13	Medical information
14	Hobby
15	Location information
16	Official ID (e.g., national IDs or license numbers)

3.2 Descriptive Results

The distribution of the participants was uniform (see Table 3), and each participant answered an 80 item questionnaire corresponding to the 80 combinations resulting from online services, types of personal data and utilisation purposes, each on a Likert scale of 1 to 6 ("1" for strongly disagree, and "6" for strongly agree.). Figure 1 summarizes the distribution of the results grouped by digital nativity[3] of users. As it can be observed, the percentage of participants decreases with the increasing acceptance of providing personal data, however, apparently the digital nativity of participants had no influence.

Finally, we used the collected data as an input for our proposed guessing schemes (Section 4). Furthermore, in order to simplify our models, we merged the obtained results into the following three classes on a scale from 0 to 2, i.e., i) 1 & 2 into scale 0; ii) 3 & 4 into scale 1; and, iii) 5 & 6 into scale 2.

4 Approach

This section introduces our initial approach, which, considers two guessing schemes, both implementing SVM as a basis. We selected SVM because it is considered a powerful learning system, although mainly for binary-class problems [38]. Nevertheless, we consider that SVMs can also efficiently perform non-linear classification by implicitly mapping their inputs into high-dimensional feature spaces through a nonlinear mapping chosen a priori. Therefore, for the purpose of our experiments, we used a multilabel and multiclass SVM approach.

[3] Individuals born after 1980, raised in a digital, media-saturated world - Prensky 2001

Table 3. Distribution of participants

Gender	Age	ratio (%)
Male	20s	10.0
Male	30s	10.0
Male	40s	10.0
Male	50s	10.0
Male	Over 60	10.0
Female	20s	10.0
Female	30s	10.0
Female	40s	10.0
Female	50s	10.0
Female	Over 60	10.0

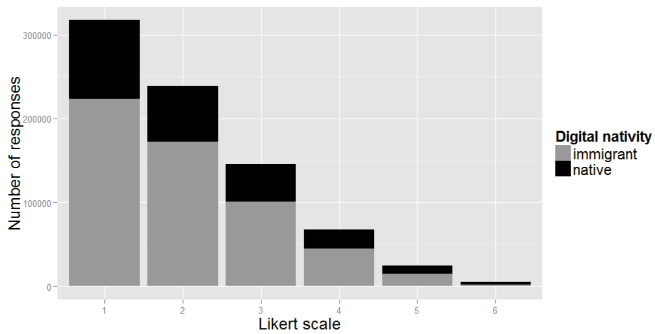

Fig. 1. Distribution of responses according to the willingness of sharing personal data and digital nativity of participants

4.1 Overview of the architecture

The proposed approach consists of *a predictor generator* and *a privacy setting prediction engine*, and a *privacy settings database*. The predictor generator, generates a question set, by selecting a minimum (optimal) number of relevant questions, which, are associated to the online service, data type and utilisation purpose from the database. The prediction engine, also generates the corresponding predictor from the modeling of existing privacy settings. The optimal question set is provided to the user, and, once the user provides the answers to the delivered question set, his/her responses are used by the privacy setting prediction engine, which, generates the predicted settings for the user. Once the personalised settings have been generated, they are communicated to the user. The high level view of the system is shown in Figure. 2

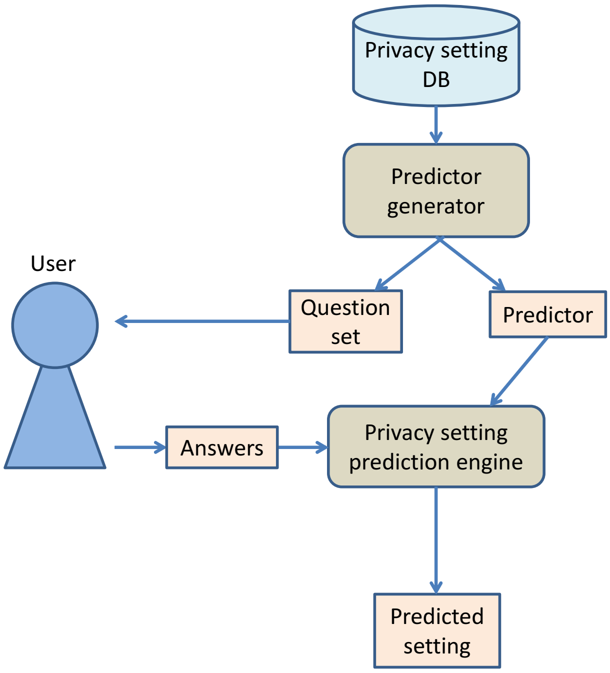

Fig. 2. High level view of the proposed system

4.2 Experimental approach

In order to demonstrate the applicability of the proposed system, we implemented a proof of concept of both the predictor generator and the privacy setting prediction engine. We evaluated them in terms of accuracy using the collected data, i.e., results of the questionnaire introduced in the previous section. In particular, the items of the questionnaire corresponded to the privacy settings in our proposed approach. Collected data was split in training data and test data. Concretely, the training data corresponded to the privacy setting database of our proof of concept. In the evaluation scheme, we first fixed a question set. Next, we regarded the values of the answers of the fixed questions as the feature vector, and we generated the optimal prediction model using the training data with our predictor generator. Afterwards, this step (previous evaluations) had been repeated for all the candidates of question sets. As a result, we obtained the question set that achieved the best accuracy and its corresponding prediction model. Finally, we evaluated the accuracy for a test data by comparing each of the predicted values generated with the answers to the fixed questions and the prediction model with real values in the test data. An abstract view of the evaluation scheme is shown in Figure. 3.

Finally, our approach was designed taking into consideration two different schemes: the first based on the sole use of SVM; while the second scheme imple-

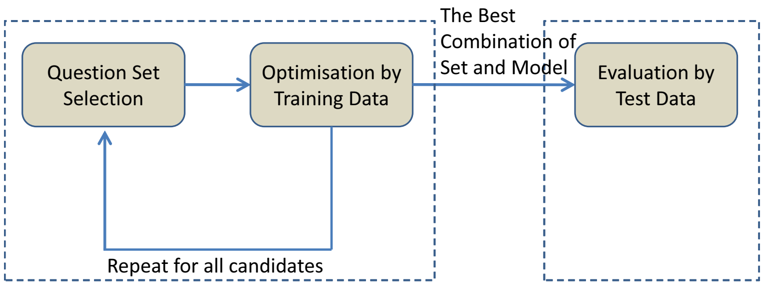

Fig. 3. Evaluation scheme

mented an additional layer that included clustering techniques. Both schemes, i.e., the SVM-based, and the combined scheme (SVM and clustering) consisted of two phases; the *learning phase* and *guessing phase*.

4.3 SVM-based Scheme

The learning and guessing phases performed by the SVM-based scheme are explained next.

[Learning Phase]

- We select n questions where $1 \leq n \leq Max$. Max equals the total number of questions and n equals the number of selected questions used for training the corresponding answers.
- Using the selected n questions, we generated the SVM privacy preference model. In this model, the class labels represent the acceptance level for each of the unselected $Max - n$ questions using a combination of answers for n as sample points in the training data.

[Guessing Phase]

- For each unknown point, i.e., a combination of answers to selected n questions, we use the SVM models generated in the learning phase for each unselected question and calculate the guessed values of the answers to those $Max - n$ unselected questions.

4.4 Combined Scheme

Similar to Section 4.3, the combined scheme consisted of two phases: the learning phase and guessing phase, the main steps of each phase are introduced next.

[Learning phase]

- We generate clusters from the training data with the corresponding clustering algorithm. Each cluster is assigned a cluster ID $i(1 \leq i \leq k)$, where k is the total number of clusters. A gravity point of a cluster is regarded as the representative values of the cluster.

– We select n questions, where $1 \leq n \leq Max$. Max equals the total number of questions and n equals the number of selected questions used for guessing the corresponding answers.
– We generate an SVM model in which the class label is mapped to the cluster ID by using as sample points, a combination of answers to selected n questions in the training data.

[Guessing Phase]

– For each unknown point (i.e., a combination of answers to selected n questions), we calculated the guessed values of a cluster ID to which the unknown point belongs. We regarded the representative values (i.e., the gravity point of the cluster) as the guessed values of answers to the $Max - n$ unselected questions.

5 Results

The proposed approach (Section 4) was implemented in a proof of concept and evaluated with real user data collected from the questionnaires. Hence, this section introduces our initial experimental results. We implemented the proposed scheme with R, and "e1071" package of SVM [39]. We evaluated each scheme by running the experiments 10 times. The data samples were chosen randomly, and were split into training data and testing data. Table. 4 shows the summary of parameters used in our experimental setup.

We performed two different experiments for each of the schemes. We first selected the top combinations, $TC = 15$ of n questions that achieved the highest accuracy considering 150 entries randomly selected; i.e., 100 entries for the training data, 50 entries for the testing data. We limited the experiment to 150 entries in order to decrease the running time when evaluating all possible combinations. We used the same top combinations, $TC = 15$ of n questions and evaluated the scheme using 10,000 entries (i.e., 9,000 for training data, and 1000 for testing data). Note that in the second experiment we cannot claim that the selected 15 combinations provide the highest accuracy.
The experiment's main steps for each of the schemes are explained in the following subsections.

Table 4. Experimental settings

Parameter	Value
Max	80
n	5
Top Combinations (TC)	$TC = 15$
Training Data (TRD)	$TRD = 100$, $TRD = 9000$
Test Data (TED)	$TED = 50$, $TED = 1000$

5.1 SVM-based Scheme

In what follows, we explain the procedures of evaluation of the model with the training data set.

- As shown in Table. 4, we first defined that n equals 5 as the number of selected questions, from a total number of $Max = 80$;
- We generated the corresponding SVM models in which the class labels were the acceptance level for each of the unselected $Max - n$ questions. We used as sample points a combination of answers for the selected n questions in the training data.
- For all 80 answers of each instance (participant) in the training data, we used the SVM models for each of the unselected $Max - n$ questions (i.e., 75), and n answers to selected n questions for each instance. Afterwards, we calculated the guessed values of the answers to the unselected questions.
- We calculated all the participants' guessed values of answers to unselected $Max-n$ questions by repeating Step 3 for all the participants in the training data.
- We compared the original values of answers to the 75 unselected questions in the training data with the guessed values of those calculated in Step 4. Finally, we regard the percentage of correctly guessed values as the accuracy of the proposed scheme.

The procedure of evaluation of the generated privacy by default preference model with the testing data is described as follows.

- We considered the SVM models generated in the learning phase.
- For all the 80 answers of a participant in the testing data, we calculated the guessed values of answers to the 75 unselected questions.
- We calculated all participants' guessed values of answers to the 75 unselected questions by repeating step 3 for each participant in the testing data.
- We compared the original values of the answers to the 75 unselected questions in the testing data with the guessed values of those calculated in step 4. We regard the percentage of correctly guessed values as the accuracy of the proposed scheme.

Table 5 shows the average of results obtained from 10 experiment runs considering the top 15 combinations (i.e., highest accuracy) of selected n questions. Each parameter of the SVM model was optimised by a grid search on the parameters C and γ. The results show a guessing accuracy of 83% for all top 15 combinations for 150 entries and 85% for 9 of the 15 top combinations.

5.2 Combined Scheme

The accuracy of the combined scheme was evaluated considering the guessed values of participants as the gravity points of the clusters to which participants belonged. The evaluation procedure consisted of the following steps.

Table 5. Results of SVM-scheme with optimization

Combination					Accuracy (TRD=100, TED=50)		Accuracy (TRD=9000, TED=1000)	
					TRD	TED	TRD	TED
A-8	B-12	C-16	D-14	E-11	0.894	0.83296	0.858903111	0.85662
B-7	C-12	D-6	D-14	D-15	0.88928	0.832106667	0.853968889	0.851904
B-12	B-15	D-5	D-8	E-6	0.88828	0.832293333	0.85102637	0.846982667
B-7	C-16	D-11	D-14	E-11	0.887986667	0.835893333	0.854038815	0.85178
B-4	B-15	D-14	E-6	E-11	0.887613333	0.832506667	0.852193333	0.849068
B-8	C-16	D-14	E-10	E-11	0.887186667	0.83728	0.854693481	0.852498667
A-8	B-12	D-6	D-14	E-11	0.884493333	0.83064	0.854496148	0.853093333
B-4	B-15	D-6	D-14	E-11	0.884226667	0.83424	0.852772296	0.85098
A-3	A-16	C-12	D-11	E-3	0.883733333	0.830426667	0.850421926	0.84796
B-7	B-12	D-14	D-15	E-6	0.883586667	0.83272	0.853168444	0.850312
B-7	C-14	D-10	D-16	E-11	0.88356	0.832106667	0.852408296	0.849949333
B-7	C-12	D-10	D-16	E-11	0.883373333	0.83552	0.851519259	0.848646667
A-2	B-7	D-14	D-16	E-11	0.8832	0.839066667	0.854657037	0.853193333
A-12	B-7	C-14	D-6	D-15	0.88316	0.8348	0.853704741	0.85178
A-12	B-8	C-16	E-10	E-11	0.882986667	0.832533333	0.852644741	0.849993333

- Using a clustering technique, we first generated clusters of participants, that corresponded to the combinations of answers of the $Max = 80$ questions. As a result, each participant was assigned a cluster ID.
- For each of the participants, we regarded the gravity point of his/her cluster as his/her guessed values for the Max answers.
- We compared the original values with the guessed values in the training data, and we regarded the percentage of the correctly guessed values as the accuracy of the selected clustering algorithm.

We run the experiments using K-means [40], Ward's method [41] and DB-Scan [42] as the selected clustering algorithm. For K-means and Ward's method, we evaluated them considering different number of clusters from 1 to 30. In the case of DB-Scan, we evaluated it considering different parameters *pts* from 2 to 6, and *eps* from 1 to 4. While K-means provided better accuracy (i.e., 77%) than Ward's method, for both the accuracy is increased by increasing the number of clusters; we evaluated the combination scheme with K-means using a total of 5 clusters. In the case of DB-Scan, it was difficult to directly compare it with K-means or Ward's method because in the DB-Scan algorithm the number of clusters cannot be decided in advance; however, in almost all cases, the accuracy of the DB-Scan algorithm was lower than K-means and Ward's method. Therefore, in the rest of the paper we focus only on K-means.

An overview of the main results for K-means, Ward's method and DB-Scan are shown in Table 6 and in Table 7 respectively.

The evaluation procedure of the combined scheme with training data is as follows.

- We generated clusters from training data using K-means. Each cluster was assigned a cluster ID $i (1 \leq i \leq 5)$.

Table 6. Accuracy of K-means and Ward

#Clusters	K-means	Ward's
1	68.01362	68.01
2	81.67737	80.11
3	82.44963	80
4	83.05238	82.07
5	83.51137	82.17
6	83.83588	82.17
7	84.4875	82.92
8	85.29425	83.16
9	84.98512	83.73
10	85.576	83.9
11	85.82725	84.18
12	86.26325	84.23
13	86.19075	84.47
14	86.46462	84.51
15	86.64112	84.74
16	86.9585	84.79
17	86.91762	84.84
18	86.8855	84.98
19	87.18925	85.2
20	86.96225	85.25
21	87.20975	85.23
22	87.20163	85.31
23	87.25513	85.5
24	87.44513	85.51
25	87.50288	85.67
26	87.41025	85.76
27	87.74637	85.94
28	87.6485	86.04
29	87.64587	86.11
30	87.79313	86.12

- We chose n equals 5 questions from a total number of $Max = 80$ questions.
- We generated an SVM model in which the class labels corresponded to the cluster ID by using a combination of answers to selected $n = 5$ questions in training data as sample points.
- For all the 80 answers of each participant in the training data, we calculated the guessed values of a cluster ID using the SVM model and the 5 answers of each participant to selected questions. We regarded the gravity point of the cluster as the guessed values of $Max - n$ i.e., 75 answers to unselected questions.
- We calculated all the participants' guessed values of answers to the 75 unselected questions by repeating step 3 for each participant in the training data.
- We compared the original values of answers to the 75 unselected questions in the training data with the guessed values of those calculated in step 4.

Table 7. DB Scan

pts	eps	#Clusters	Accuracy
2	1	76	0.767654
2	2	61	0.789983
2	3	44	0.709418
2	4	15	0.697803
3	1	41	0.764831
3	2	31	0.788213
3	3	17	0.707174
3	4	5	0.696771
4	1	34	0.762885
4	2	21	0.786394
4	3	11	0.802324
4	4	2	0.702045
5	1	28	0.761396
5	2	21	0.786629
5	4	2	0.702076
5	4	2	0.702076
6	1	19	0.759063
6	2	16	0.785908
6	3	7	0.802275
6	4	2	0.702083

We regarded the percentage of correctly guessed values as the accuracy of the proposed scheme.

The evaluation procedure of the combined scheme with testing data is as follows.

- We used the SVM model generated in the learning phase. The class label of the model was associated with the cluster ID by using a combination of answers to the 5 selected questions in the training data as sample points.
- For all the 80 answers of a participant in the testing data, we calculated the guessed values of a cluster ID for the participant with the SVM model and the 5 answers of the participant to selected questions. We regarded the gravity point of the cluster as the guessed values of the 75 answers to the unselected questions.
- We calculated all the participants' guessed values of answers to 75 unselected questions by repeating step 3 for all the participants in the testing data.
- We compared the original values of answers to the 75 unselected questions in the training data with the guessed values of those calculated in step 4. Afterwards, we considered the percentage of correctly guessed values as the accuracy of this scheme.

The result is shown in Table 8. "Cluster accuracy for training data" means the percentage of correctly guessed values for the cluster ID calculated in step 4 of the evaluation procedure for the training data.

Table 8. Accuracy of the combined scheme (TRD=100, TED=50)

Combination					Cluster accuracy - TRD	Accuracy - TRD	Accuracy - TED
A-11	A-15	B-4	C-2	D-6	0.744	0.8245	0.819975
A-12	B-7	B-8	D-11	E-9	0.76	0.83405	0.8238
B-6	B-7	D-7	E-10	E-11	0.752	0.83355	0.8188
A-10	B-4	D-4	E-6	E-8	0.724	0.822475	0.81155
A-10	B-4	D-6	D-9	E-6	0.73	0.82835	0.82105
A-10	B-4	D-6	D-9	E-7	0.736	0.8317125	0.820525
A-10	B-4	D-7	D-9	E-6	0.725	0.828875	0.821175
A-10	B-4	D-9	E-4	E-6	0.711	0.8275	0.8192
A-11	B-4	B-8	D-10	E-6	0.721	0.828625	0.822875
A-11	B-4	D-10	E-6	E-13	0.7	0.8228	0.8152
A-13	B-4	D-11	E-6	E-11	0.712	0.827275	0.820375
A-16	B-6	B-10	D-8	E-6	0.775	0.8337875	0.8232
B-4	B-10	D-4	D-13	E-7	0.761	0.8310375	0.819125
B-4	D-4	D-6	D-13	E-12	0.754	0.8316375	0.8213
B-4	D-6	D-9	E-4	E-7	0.705	0.8225	0.8181

The best accuracy achieved by the combined scheme was 82%. This accuracy was achieved using 8 of the top 15 combinations for 150 entries, and 12 of the top 15 combinations for 10,000 entries.

6 Discussion

The proposed default privacy preference setting guessing scheme based on SVM, and its extension, which included a combination of SVM with clustering techniques has achieved a reasonably high level of precision for guessing the default privacy setting with minimal user input. Specifically, we had 80 questionnaire items out of which only five were used to guess for the remaining 75 questions. These automated default settings not only relieve users of the burden of carrying out tiresome privacy setting tasks, but also relieve them from having to make information disclosure decisions later on.

Results show that the first scheme offers better accuracy (i.e., 85%) than the combined scheme (i.e., 82%). However, when compared to the combined scheme, the SVM only scheme performs more slowly due to the number of models that need to be created (i.e., 75). Thus, considering a minimum difference in accuracy (3%), one could decide to implement the combined scheme and have better performance, in particular considering that the additional time for clustering with K-means for 9,000 entries is minimal (i.e., 0.3 seconds) and therefore, could be neglected. To the best of our knowledge, this result demonstrates the first personalised privacy by default setting generated using SVM and clustering algorithms applicable to web services in general. Authors [43], introduced a user preference predicting approach for common preferences. Their study used similarity-based clustering to group users with similar interests achieving 80% of accuracy. Additionally, they introduced an error correcting procedure to boost

Table 9. Accuracy of Combination Scheme (#Training data = 9,000, #Test data = 1,000)

Combination					Cluster accuracy - TRD	Accuracy - TRD	Accuracy - TED
A-11	A-15	B-4	C-2	D-6	0.731411111	0.81693	0.81735875
A-12	B-7	B-8	D-11	E-9	0.748988889	0.82109125	0.82167
B-6	B-7	D-7	E-10	E-11	0.724666667	0.822432917	0.823305
A-10	B-4	D-4	E-6	E-8	0.744133333	0.820498889	0.8205675
A-10	B-4	D-6	D-9	E-6	0.746	0.81941375	0.81997875
A-10	B-4	D-6	D-9	E-7	0.763822222	0.823401111	0.8250475
A-10	B-4	D-7	D-9	E-6	0.759411111	0.822305694	0.82301125
A-10	B-4	D-9	E-4	E-6	0.751011111	0.819230278	0.8195725
A-11	B-4	B-8	D-10	E-6	0.743255556	0.820663889	0.820705
A-11	B-4	D-10	E-6	E-13	0.755888889	0.821184306	0.821355
A-13	B-4	D-11	E-6	E-11	0.743044444	0.821143889	0.82237
A-16	B-6	B-10	D-8	E-6	0.757722222	0.82313375	0.823545
B-4	B-10	D-4	D-13	E-7	0.7456	0.8230475	0.82392625
B-4	D-4	D-6	D-13	E-12	0.749477778	0.823683889	0.82439125
B-4	D-6	D-9	E-4	E-7	0.7408	0.823176528	0.8243375

the accuracy to 98%. However, the results from the error correcting procedure have been achieved using simulated data.

Even though our approach demonstrated the applicability of machine learning algorithms in privacy by default settings with a considerably high accuracy, it has some limitations that should be considered in future research. The guessing precision of the algorithms is dependent on the training and testing input data provided to it by the user-answered questionnaire items. However, the correctness and genuineness of the answers is dependent on the user providing rational and intentionally correct answers. In addition, the user study was carried out in Japan, and cultural attributes may influence the extent to which the results can be generalised and applied to other societies. Furthermore, we limited our study to 5 questions considering the top 15 combinations of 150 entries, therefore, additional research is needed in order to determine both the optimal number and best combination of questions that are sufficient to have an acceptable accuracy of prediction. In our future work, we plan to run more number of experiments with varying learning algorithms. Finally, the proposed approach only focused on default privacy preference settings and, not on the multi-dimensional privacy issues that users face when using Internet services and making data disclosure and non-disclosure decisions.

7 Related Work

With the advent of privacy violations and increased user privacy concerns, significant efforts have been put on privacy policy representation. However, approaches to end user privacy preference settings management are still limited. In this regard, Kolter and Pernul highlighted the importance of privacy preferences and proposed a user-friendly, P3P-based privacy preference generator [22] for service providers that included a configuration wizard and a privacy preference summary. In a similar form, the research approach proposed by Biswas [23] focused on privacy settings and consisted of an algorithm to detect the conflicts in privacy settings, specifically, between user preferences and application requirements in smart phone ecosystems.

Authors in [24] proposed Privacy Butler; a personal privacy manager to monitor a user's online presence based on a privacy policy. This concept focuses only on content related to user's online presence in a social network; and it monitors whether third parties have disclosed user's information without consent, this mechanisms verifies the content satisfactorily matches the privacy preference of the user; and, in case of a mismatch it attempts to modify or delete the corresponding content. Srivastava [25, 26] proposed a privacy settings recommender system also focused on online social network services.

Berendt *et al.* [27] emphasised the importance of automatic privacy preference generation and Sadah *et al.* [28] suggested that machine learning techniques have the power to generate more accurate preferences than users themselves and relieve them from the complex task of specifying their privacy preferences. This issue has been supported by Madejski *et al.* [29], whose study focused in online social networks and demonstrated that there exists a serious mismatch between intentions for privacy settings and real settings. Preference modelling for eliciting preferences was studied by Bufett and Fleming [30]. Mugan *et al.* [31] proposed a method for generating persona and suggestions intended to help users incrementally refine their privacy preferences over time. Fang *et al.* [32, 33] have proposed a privacy wizard for social networking sites. The purpose of the wizard is to automatically configure a users' privacy settings with minimal effort required by the user. The wizard is based on the underlying observation that real users conceive their privacy preferences based on an implicit structure. Thus, after asking the user a limited number of carefully chosen questions, it is usually possible to build a machine learning model that accurately predicts the users' privacy preferences. Although, similar work is presented, our approach is applicable to general online services, while theirs is limited in scope (i.e., used to restrict privacy of friends in social media, namely, Facebook). Moreover, their model works similar to an access control list where users put restrictions on their Facebook friends while ours sets the privacy preference of web services.

Furthermore, Lin *et al.* [34] applied hierarchical clustering techniques to analyse and understand users' mobile app privacy presences. The authors analysed mobile apps privacy behaviours using static analysis tools, and also crowdsourced users' mobile app privacy preferences using Amazon Mechanical Turk. While the results are interesting, their privacy preference clustering is more focused to mobile apps, i.e Android permission model. Guo and Chen [35] proposed an algorithm to optimise privacy configurations based on desired privacy level and utility preference of users, in this approach users are still required to set up a preference level. Contrary to this, Tondel *et al.* [36] proposed a conceptual architecture for learning privacy preferences based on the decisions that users make in their normal interactions on the web. Authors suggested that learning of privacy preferences has the potential to increase the accuracy of preferences without requiring users to have a high level of knowledge or willingness to invest time and effort in their privacy. Although interesting work, its design is based on the assumption that users are privacy conscious and are expected to be willing to take part in the preference generation by installing a user agent. Additionally, no practical implementation or experimentation has been provided.

Authors [37] designed a fine-grained privacy preference model using ontologies that enables users to set privacy preferences on their data. Even though their approach presents a light weight solution, the user will have to run the privacy preference ontology every time she wants to affect the setting. Further more, their approach is also dependent on the Web Access Control vocabulary.

8 Conclusions and future work

In this paper we introduced a machine learning approach in order to provide personalised default privacy settings. We argue that the complexity of setting privacy preferences is a burden that shouldn't be put on to users especially under the assumption that users are able to choose the best privacy setting for themselves. While this may be true in for some cases, it has been shown that ordinary online users fall far short of being able to do this. This calls for the need to help users with efficient and tailored privacy preference mechanisms. Therefore, in this study, we have designed and implemented a proof of concept based on machine learning in order to facilitate the privacy settings of users by asking them a minimum number of questions. The results show that machine learning algorithms have great potential to automate privacy preference setting with minimal input from users. Future work will include further enhancing the accuracy of the preference setting results. To this end, we plan to investigate techniques for finding the combination of questions that will maximise the accuracy of the prediction scheme. Furthermore, repeating the experiment with different user group and experimental setup could enrich the conclusions and generalisations drawn in this paper, therefore, in the research roadmap, we plan to collect data from European users as well. Finally, we plan to evaluate the proof of concept with real users and enable the system to learn from users' privacy preferences when they begin interacting with the associated service.

Acknowledgments. This paper is an extension of the paper published in the ICISSP2016 conference. Authors would like to thank anonymous reviewers for their insightful comments, which, allow us to improve the overall quality of the paper.

References

1. Eszter Hargittai et al. Facebook privacy settings: Who cares? *First Monday*, 15(8), 2010.

2. Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing facebook privacy settings: User expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 61–70, New York, NY, USA, 2011. ACM.
3. Andre Deuker. Addressing the privacy paradox by expanded privacy awareness the example of context-aware services. *Privacy and Identity Management for Life*, pages 275–283, 2010.
4. Anirban Basu, Jaideep Vaidya, and Hiroaki Kikuchi. Efficient privacy-preserving collaborative filtering based on the weighted slope one predictor. *Journal of Internet Services and Information Security (JISIS)*, 1(4):26–46, 11 2011.
5. Marcello Paolo Scipioni and Marc Langheinrich. Towards a new privacy-aware location sharing platform. *Journal of Internet Services and Information Security (JISIS)*, 1(4):47–59, 11 2011.
6. Saikat Guha, Bin Cheng, and Paul Francis. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, IMC '10, pages 81–87, 2010.
7. Aleksandra Korolova. Privacy violations using microtargeted ads: A case study. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 474–482, 2010.
8. A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *Security Privacy, IEEE*, 3(1):26 –33, 2005.
9. Irene Pollach. What's wrong with online privacy policies? *Commun. ACM*, 50(9):103–108, 2007.
10. Carlos Jensen, Colin Potts, and Christian Jensen. Privacy practices of internet users: self-reports versus observed behavior. *Int. J. Hum.-Comput. Stud.*, 63(1-2):203–227, 2005.
11. Daniel J. Solove. Privacy self-management and the consent paradox. *Harvard Law Review*, 126, 2013.
12. W3C. The platform for privacy preferences 1.0 (P3P1.0) specificati. In *Platform for Privacy Preferences (P3P) Project*, 2002.
13. L.F. Cranor. P3p: making privacy policies more useful. *Security Privacy, IEEE*, 1(6):50 – 55, 2003.
14. Lorrie Faith Cranor, Manjula Arjula, and Praveen Guduru. Use of a p3p user agent by early adopters. In *Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society*, WPES '02, pages 1–10, 2002.
15. Lorrie Faith Cranor, Praveen Guduru, and Manjula Arjula. User interfaces for privacy agents. *ACM Trans. Comput.-Hum. Interact.*, 13(2):135–178, 2006.
16. A. Pedersen. P3 - problems, progress, potential. *Privacy Laws & Business International Newsletter*, 2:20–21, 2003.
17. Michael Backes, Günter Karjoth, Walid Bagga, and Matthias Schunter. Efficient comparison of enterprise privacy policies. In *Proceedings of the 2004 ACM symposium on Applied computing*, SAC '04, pages 375–382, 2004.
18. A. Dehghantanha, N.I. Udzir, and R. Mahmod. Towards a pervasive formal privacy language. In *Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on*, pages 1085 –1091, 2010.
19. K. Bekara, Y. Ben Mustapha, and M. Laurent. Xpacml extensible privacy access control markup langua. In *Communications and Networking (ComNet), 2010 Second International Conference on*, pages 1 –5, 2010.
20. Inger Anne Tondel and Asmund Ahlmann Nyre. Towards a similarity metric for comparing machine-readable privacy policies. In *Open Problems in Network Security*, volume 7039 of *Lecture Notes in Computer Science*, pages 89–103, 2012.

21. G.O.M. Yee. An automatic privacy policy agreement checker for e-services. In *Availability, Reliability and Security, 2009. ARES '09. International Conference on*, pages 307 –315, 2009.
22. J. Kolter and G. Pernul. Generating user-understandable privacy preferences. In *Availability, Reliability and Security, 2009. ARES '09. International Conference on*, pages 299 –306, 2009.
23. D. Biswas. Privacy policies change management for smartphones. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 70 –75, 2012.
24. R. Wishart, D. Corapi, A. Madhavapeddy, and M. Sloman. Privacy butler: A personal privacy rights manager for online presence. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, pages 672 –677, 2010.
25. A. Srivastava and G. Geethakumari. A framework to customize privacy settings of online social network users. In *Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in*, pages 187–192, 2013.
26. A. Srivastava and G. Geethakumari. A privacy settings recommender system for online social networks. In *Recent Advances and Innovations in Engineering (ICRAIE), 2014*, pages 1–6, 2014.
27. Bettina Berendt, Oliver Günther, and Sarah Spiekermann. Privacy in e-commerce: Stated preferences vs. actual behavior. *Commun. ACM*, 48(4):101–106, 2005.
28. Norman Sadeh, Jason Hong, Lorrie Cranor, Ian Fette, Patrick Kelley, Madhu Prabaker, and Jinghai Rao. Understanding and capturing people's privacy policies in a mobile social networking application. *Personal Ubiquitous Comput.*, 13(6):401–412, 2009.
29. 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*, pages 340–345, 2012.
30. Scott Buffett and Michael W. Fleming. Applying a preference modeling structure to user privacy. In *Proceedings of the 1st International Workshop on Sustaining Privacy in Autonomous Collaborative Environments*, 2007.
31. Jonathan Mugan, Tarun Sharma, and Norman Sadeh. Understandable learning of privacy preferences through default personas and suggestions, 2011.
32. Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web*, pages 351–360. ACM, 2010.
33. Lujun Fang, Heedo Kim, Kristen LeFevre, and Aaron Tami. A privacy recommendation wizard for users of social networking sites. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 630–632. ACM, 2010.
34. Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. Modeling users? mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 199–212, 2014.
35. Shumin Guo and Keke Chen. Mining privacy settings to find optimal privacy-utility tradeoffs for social network services. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 656–665, 2012.
36. I.A. Tondel, A.A. Nyre, and K. Bernsmed. Learning privacy preferences. In *Availability, Reliability and Security (ARES), 2011 Sixth International Conference on*, pages 621–626, 2011.

37. Owen Sacco and Alexandre Passant. A privacy preference ontology (ppo) for linked data. In *LDOW*. Citeseer, 2011.
38. Steve R Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14, 1998.
39. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, and Chih-Chen Lin. Package 'e1071'. `https://cran.r-project.org/web/packages/e1071/e1071.pdf`, 2015.
40. James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.
41. Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
42. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.
43. M. Qin, S. Buffett, and W. Fleming. Predicting user preferences via similarity-based clustering. In *Canadian Conference on AI*, volume 5032 of *Lecture Notes in Computer Science*, pages 222–233. Springer, 2008.