Default Privacy Setting Prediction by Grouping User's Attributes and Settings Preferences

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

[1] KDDI Research, Inc., Saitama, Japan
{tr-nakamura, kiyomoto}@kddi-research.jp
[2] Chair of Mobile Business and Multilateral Security, Goethe University Frankfurt
Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany
{firstname.lastname}@m-chair.de

Abstract. While user-centric privacy settings are important to protect the privacy of users, often users have difficulty changing the default ones. This is partly due to lack of awareness and partly attributed to the tediousness and complexities involved in understanding and changing privacy settings. In previous works, we proposed a mechanism for helping users set their default privacy settings at the time of registration to Internet services, by providing personalised privacy-by-default settings. This paper evolves and evaluates our privacy setting prediction engine, by taking into consideration users' settings preferences and personal attributes (e.g. gender, age, and type of mobile phone). Results show that while models built on users' privacy preferences have improved the accuracy of our scheme; grouping users by attributes does not make an impact in the accuracy. As a result, services potentially using our prediction engine, could minimize the collection of user attributes and based the prediction only on users' privacy preferences.

Keywords: Privacy preference, Privacy setting, Machine learning

1 Introduction

Usage of personal data is increasing as it is believed to promote innovation. However, it also raises privacy concerns. In many cases, a service delivered to users is provided with embedded privacy functionality that can limit the sharing of personal data by the user in specific scenarios or given situations. For instance, Facebook provides the user with the privacy setting functionality that enable users to manage which other users can browse his/her posts, pictures, etc. Similarly, modern smartphones (e.g. Android and iPhone) provide users the possibility to control which applications can access different resources including personal or privacy related data. In future, such settings may be used not only for permitting to provide personal data but also for deciding some privacy level such as anonymization level. Generally speaking, personal data is anonymized in higher level, the usability becomes lower. So if starting with the most privacy-friendly pre-setting, the users may not be able to use high quality services unless

they manually change their settings. However, many users do not change the privacy settings, either because of the effort required or due to the lack of a proper understanding of privacy settings. Thus, to address this, general frameworks, such as PDS (Personal Data Store) [4] and PPM (Privacy Policy Manager) [14] have emerged, which provide the user with a generic privacy manager for various types of personal data and service providers.

When providing a privacy function, the default settings are very important because many users may not spend the time and effort to set their privacy preferences adequately. It is especially difficult to manually configure appropriate privacy settings as the combinations of service providers, types of personal data, and the applications for personal data have become so vast. 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 translate these concerns into personalized privacy settings to users.

In our initial efforts to overcome this, we proposed a conceptual design and a mechanism based on a Support Vector Machine (SVM) for the automatic generation of personalized privacy settings [17]. In our basic approach we have designed a questionnaire of 80 questions that considered the combination of 16 different data types shared for 5 different utilization purposes and services. The basic approach delivered a minimal set of (5) questions to each user at registration time, and from the user's answers, it predicted the default privacy settings for each user.

In this paper, we present a more advanced scheme and a prototype that improve the accuracy of the privacy setting prediction, based on the grouping of users' attributes and setting preferences. Thus, the contribution of this paper is twofold. First, we present an extension and improvement of previous work [17], which was focused on selecting optimal and minimal number of questions to predict the privacy settings. In this work, we further elaborate and give an in-depth analysis on the improvement mechanisms by considering user attributes and privacy preferences. Second, to showcase the applicability of the proposed models, we implemented a prototype of the prediction engine in R using SVM based models in order to predict user privacy settings.

The rest of the paper is organized as follows. Section 2 provides an overview of related work in the area of privacy preferences. Section 3 describes the main methodology and approach of the SVM-based prediction scheme proposed in [17] and the questionnaires designed and used to derived initial settings database. Section 4 describes the experimental evaluation for both user attributes and privacy preferences. Section 5 discusses the results of the evaluation. Section 6 draws the main conclusions and points out future directions for research.

2 Related Work

In privacy policy management the burden of checking on and maintaining privacy policies has been identified as a major issue. In one study, Madejski *et al.* [15] showed that a serious mismatch existed between intentions for privacy settings

and real settings in an online social network service. Users are commonly required to check the privacy policies of a given service offered by a service provider before starting to use the service. Thus, each service provider prepares a privacy policy for each service. Because it is frequently the case that users must check a large number of privacy policies, it becomes irksome and difficult to understand. Consequently, users are not able to determine or customise the privacy policies for themselves. Furthermore, if a user does not agree with the privacy policy of a service, the user simply cannot use the service.

In this regard, Solove suggested that the privacy self-management model cannot achieve its objectives, and it has been pushed beyond its limits, while privacy law has been relying too heavily upon the privacy self-management model [20]. Moreover, other studies such as the experimental study conducted by Acquisti and Grossklags [1] demonstrated users' lack of knowledge about technological and legal forms of privacy protection when confirming privacy policies. Their observations suggest that several difficulties obstruct individuals in their attempts to protect their own private information, even those concerned about and motivated to protect their privacy. This was reinforced by authors in [18] 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 [12]. In this regard, Guo and Chen [11] proposed an algorithm to optimise privacy configurations based on desired privacy level and utility preference of users.

Fang *et al.* [10, 9] have proposed a privacy wizard for social networking sites. The purpose of the wizard is to automatically configure a user's 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 user's preferences. This approach is very similar to ours. The difference is the target dataset. Fang *et al.* treated real data of Facebook, so the variety of the items was limited and the number of the participants is small. We treat more general data items and the number of the participants is larger because our approach does not focus on a specific service such as Facebook.

Some languages to describe privacy policies have been presented in [7, 8, 3]. Backes *et al.* examined some comparisons of enterprise privacy policies using formal abstract syntax and semantics to express the policy contents [2]. Tondel and Nyre [22] proposed a similarity metric for comparing machine-readable policies.

There is some existing research about learning privacy preferences. Berendt *et al.* [5] emphasised the importance of privacy preference generation and Sadah *et al.* [19] suggested that machine learning techniques have the power to generate more accurate preferences than users themselves in a mobile social networking application. Tondel *et al.* [21] proposed a conceptual architecture for learning privacy preferences based on the decisions a user makes in their normal interactions on the web. They suggested that learning of privacy preferences has the potential to increase the accuracy of preferences without requiring users to have

Table 1. Types 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 (age, gender, income)
10	Contents of email, blog, twitter etc.
11	Session information (e.g., Cookies)
12	Social Info. (e.g., religion, volunteer records)
13	Medical Info.
14	Hobby
15	Location Info.
16	Official ID (national IDs or license numbers)

Table 2. Usage purposes

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

a high level of knowledge or willingness to invest time and effort in their privacy. Kelley *et al.* [13] showed preferences for a mobile social network application. Preference modeling for eliciting preferences was studied by Bufett and Fleming [6]. Mugan *et al.* [16] proposed a method for generating persona and suggestions intended to help users incrementally refine their privacy preferences over time.

3 SVM Based Privacy Setting Prediction Scheme

This section introduces the SVM-scheme used as the basis of our approach, as well as the questionnaires designed in order to get the initial privacy settings database.

3.1 Design of Questionnaires

We designed a questionnaire survey focused on the acceptability from users to provide personal data, considering a combinations of 16 data types (cf. Table 1) for 5 utilization purposes (cf. Table 2). The data types and usage purposes were selected from the items defined in P3P [23]. In this work, we prioritized to make them close to P3P categories. We recognize that there are some misleading and uneasy to understand points, hence we will modify them next evaluation. Additionally, other attributes related to demographics and type of mobile device used were considered because they might have possibility to find any special features in the groups separated with them.

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

Table 4. Distribution of types of mobile phone

Mobile phone	ratio (%)
iPhone	23.5
Android	30.0
Others	1.71
Not smart phone	44.9

We collected responses from 10,000 Japanese participants and they answered our questionnaires by web-based system. As it is shown in Table 3 the distribution of the participants was uniform over all the categories. Each participant evaluated all 80 combinations of types of personal data and usage purposes on a Likert scale of 1 to 6 ("1" for strongly disagree, and "6" for strongly agree.). The distribution of mobile devices used by participants is shown in Table 4. Table 5 shows the distribution of the results. As can be observed from Table 5, the percentage decreases with the increasing acceptance of providing personal data. For the sake of simplicity, the obtained results were merged initially 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. In future, we also plan to perform experiments using a different merging approach. The differences between questions are shown in Figure 1.

Table 5. Distribution of result

Likert scale	1	2	3	4	5	6	Total
Number	317497	238826	145952	67629	24583	5513	800000
Ratio	0.3969	0.2985	0.1824	0.08454	0.03073	0.006891	1

3.2 Comparison Based on Attributes

The trend based on the attributes of participants is shown in Figure 2. Between genders, the trend for males is more positive than for females, that is, the ratios for answering "2" (means positive for providing personal data) and "1" (means neutral) for males are about 1% and 2% higher than those for females, respectively. Based on age, the most positive age group is in their 20s, while the most negative group is in their 40s. The ratios for those answering "2" and "1" in

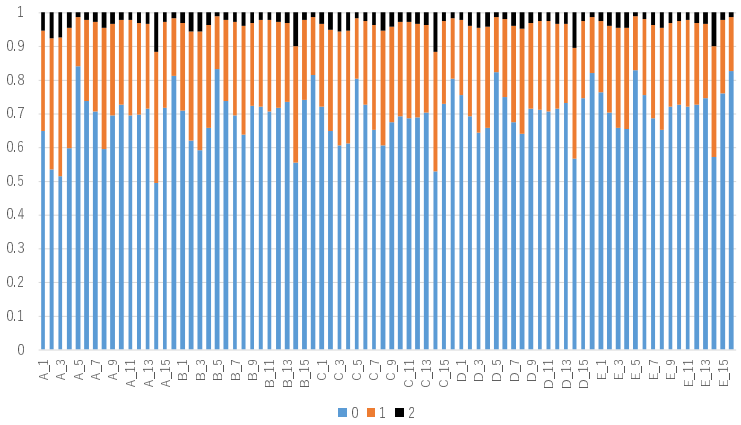

Fig. 1. Differences between questions

their 20s are about 3% and 6% higher than for those in their 40s, respectively. For the type of mobile phone, the ratio for answering "2" for iPhone users is about 1% higher than that for Android users, while the others are similar.

3.3 SVM-based Prediction Scheme

This paper considers as a basis only the first SVM-based scheme introduced in [17] and evaluates the change of accuracy when the dataset is either grouped by user attributes or grouped by user setting preferences. Thus, we used the same dataset detailed in Section 3.1 for the evaluation. A high level description of the prediction scheme is shown in Figure 3. The main procedure is as follows:

1. An existing user settings database is the input to a prediction model generator in order to generate an optimal question set and the prediction model.
2. A user is provided with the question set (5 questions).
3. The user's answers to the selected questions are then the input to the prediction model so that the privacy setting prediction engine generates the corresponding (personalized) prediction values.
4. The prediction values are then recommended to the user.

The abstract of the prediction-model-generating algorithm is shown in Figure 4. The prediction-model-generating algorithm is detailed below.

1. The existing user settings database is split into learning data and test data.
2. Questions are randomly selected for prediction.
3. SVM models are generated for the rest of the questions (75) in the learning data by using selected questions in the learning data as feature vectors.
4. The SVM models that were created in the previous step are evaluated using the test data.

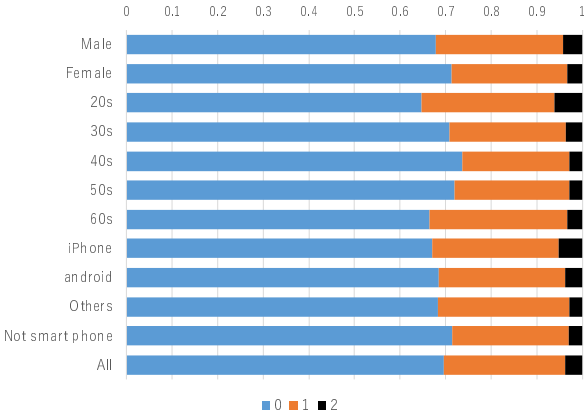

Fig. 2. Tendency on attributes

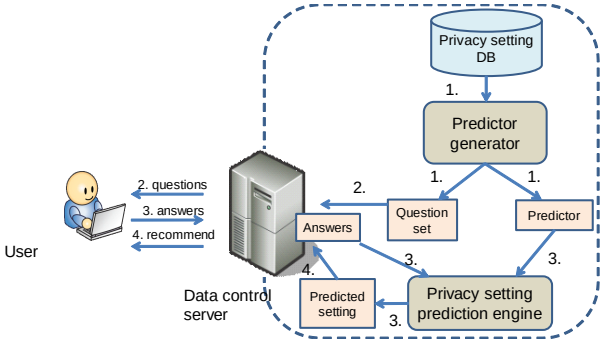

Fig. 3. The framework of our prediction scheme

5. The process is repeated to evaluate for an adequate number of combinations of questions, and the combination of questions achieving the highest accuracy as the selected questions is adopted.

4 Experimental Evaluation

Appropriate parameters need to be chosen such as the number of learning data, test data, items for prediction of answers, and combinations of items for evaluation in order to efficiently make experiments in various conditions. Generally, if a greater number of learning data items and combinations of items for evaluation are used for prediction, higher accuracy can be expected, but meanwhile, the processing time (especially critical for generating the SVM model) is also increasing.

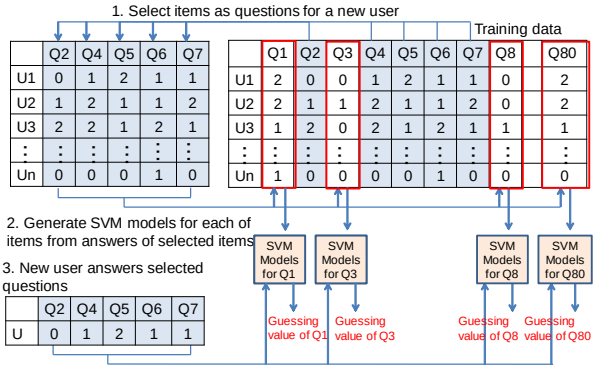

Fig. 4. The abstract of our prediction algorithm

A preliminary experiment was performed for choosing the appropriate values for these parameters. The experimental parameters are shown in Table 6. This experiment was performed using parallel processing with two machines. In this experiment, the parameters of SVM were not adjusted, and the default parameters such that $\gamma = 0.2$ and $cost = 1$ were always used.

Table 6. Experiment settings

OS	Windows8.1
Memory	8GB
CPU	intel core i7-4770 @ 3.40GHz
Language, Library	R, e1071(SVM), doSNOW(Multi core processing)

In order to discover an adequate number of samples of combinations of items and finding the most suitable combination for prediction of answers, the accuracy is evaluated by varying the number of samples of combinations from 1,000 to 10,000 in increments of 1,000 and fixing the number of learning data, test data, and items for prediction of answers at 100, 50, and 5, respectively. Learning data and test data were randomly chosen from the original dataset twice and called dataset A and dataset B. For each dataset, we randomly choose samples of combinations of items, evaluate all combinations, and find the best combination and its accuracy. After five evaluations, we regard the average of accuracy of the five evaluations as the accuracy of the dataset. The results show that 10,000 samples of combinations are sufficient because the maximum differences in accuracy in dataset A and B are only about 0.46% and 0.67%, Figure 5.

As a second step, in order to discover an adequate number of test data, the accuracy is evaluated by varying the number of test data from 500 to 5,000 in increments of 250 and fixing the number of learning data, items for prediction, samples of combinations of items at 100, 5, and 10,000, respectively. Learning

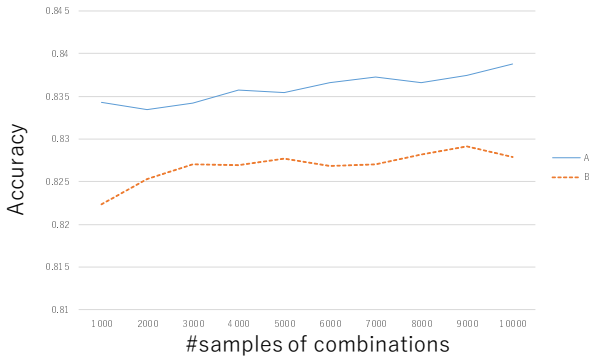

Fig. 5. Influence of the number of samples of combinations

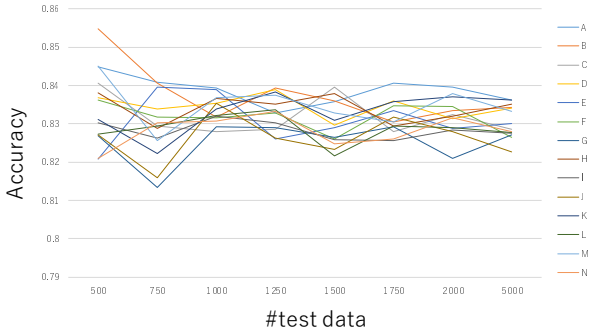

Fig. 6. Influence of the number of test data

data from the original dataset are randomly chosen 14 times, as samples of combinations of items, and called datasets A to N. For each dataset, we randomly choose test data from original dataset for ten times, evaluate all combinations of items, and find the best combination and its accuracy. After ten evaluations, we regard the average of accuracy of the ten evaluations as the accuracy of the dataset. The result is shown in Figure 6. The result shows that 1,000 test data are sufficient because the variance is about 0.00007 when the number of test data is 750, the variance is about 0.00001 when the number of test data is 1,000, and the variance does not decrease much with further increases of the number of test data above 1,000.

For learning data, the accuracy is evaluated by varying the number of learning data from 50 to 500 and fixing the number of test data, items for prediction of answers, samples of combinations of items at 1,000, 5, and 10,000, respectively. Test data are randomly chosen from the original dataset five times, as samples of combinations of items, and called datasets A to E. For each dataset, we randomly choose learning data from original dataset for ten times, evaluate all

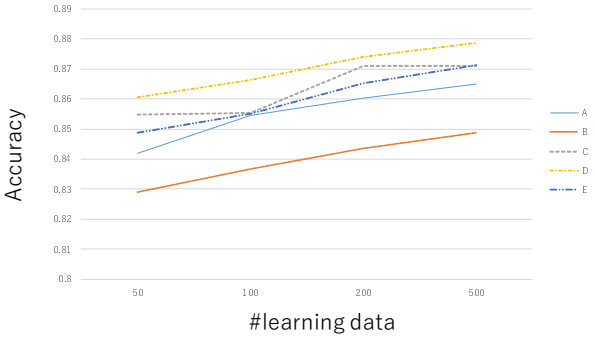

Fig. 7. Influence of the number of learning data

combinations of items, and find the best combination and its accuracy. After ten evaluations, we regard the average of accuracy of the ten evaluations as the accuracy of the dataset. The results show (Figure 7) that the accuracy linearly increases with the increase in size of learning data, hence the number of learning data is set to 100, considering the processing time for evaluation.

Finally, in order to discover an adequate number of items for prediction, the accuracy is evaluated by varying the number of items for prediction from 2 to 10 and fixing the number of learning data, test data, and samples of combinations of items at 100, 1,000, and 10,000, respectively. We randomly choose learning data and test data from original dataset for five times, evaluate all combinations of items, and find the best combination and its accuracy. After five evaluations, we regard the average of the accuracy of the five evaluations as the accuracy of the dataset. The results show (Figure 8) that the increase of accuracy is reduced when the number of items for prediction is greater than six, hence the number of items for prediction is set at five.

From the previous results, the parameters in this experiment are set as shown in Table 7. Note that the SVM parameters are not adjusted, and the default SVM parameters are used such that $\gamma = 0.2$ and $cost = 1$ both in this section and in Section 4.1 and 4.2.

Regarding the computation time, the process of selecting the best combination of items from 10,000 combinations, requires about 4,013 seconds with a single-core computation in the environment shown in Table 6. Using the same setup, the process of generating the prediction requires about 0.32 seconds. Note that, the process of choosing the best combination does not affect the user experience, thus, even with larger numbers it could be neglected; furthermore, the overall computation time could be reduced by using parallel computation.

4.1 Evaluation by Attributes Grouping

In this section, the original data set is grouped by the participants' attributes such as gender, age, and type of mobile phone. The accuracy is evaluated in order

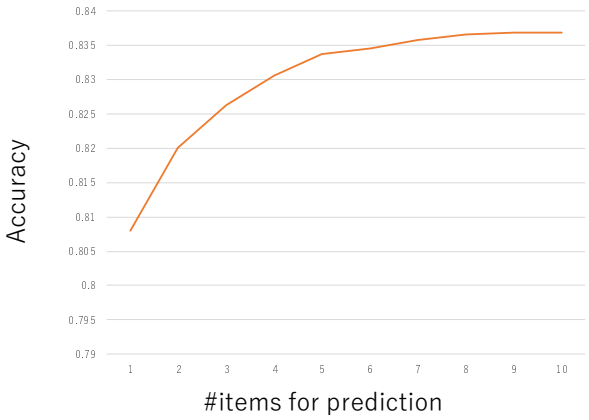

Fig. 8. Influence of the number of items for prediction

Table 7. Parameters in this experiment

# learning data	100
# test data	1000
# items for prediction	5
# samples of combinations	10000
γ (Parameter on SVM)	0.2
cost (Parameter on SVM)	1.0

to generate the prediction model from the grouped data set. The parameters used for the evaluations are the same as in Section 4. Note that the size of learning data or test data does not decrease even if the data set is divided into small subsets. Learning data and test data are randomly chosen from the grouped subset 10 times, as samples of combinations of items, and the average of the accuracy is evaluated in the 10 trials. The result is shown in Table 8. Note that on the type of mobile phone, the item "other smart phone" is omitted because the number is too small.

According to the results, in all the cases where the original data set is grouped by gender, age, and type of mobile phone, the total accuracy decreases compared to the original approach (data set not grouped), though there are some categories in which the accuracy increases.

4.2 Evaluation by Privacy Preferences

We selected the K-means algorithm, and used it to observe the participants' answer preferences. The number of clusters is varied between 1 and 10. For instance, the case where the number of clusters is 4 is shown in Figure 9.

The results show that there are two characteristic clusters: Cluster 1 and Cluster 4. The participants in Cluster 1 tend to answer "0" (means negative),

Table 8. Accuracy by grouping by attributes

Not grouping		Accuracy
	Total	**0.8415**
Gender	Male	0.8364
	Female	0.8348
	Total	**0.8356**
Age	20s	0.8073
	30s	0.8421
	40s	0.8519
	50s	0.8511
	Over 60	0.8243
	Total	**0.8353**
Type of mobile phone	iPhone	0.8248
	Android	0.8282
	Other smart phone	
	Not smart phone	0.8445
	Total	**0.8325**

and the participants in Cluster 4 tend to answer "1" (means neutral) for almost all the questions. It is easy to determine to which cluster a person belongs, e.g., Cluster 1, Cluster 4, or another cluster, because it is only necessary to ask his/her basic privacy attitude directly, for example, "Would you prefer that your personal data never be provided at all?". If accuracy is improved by grouping the original data set by clustering on the answer preferences, it may be possible to improve our scheme by adding only one question that may determine to which cluster a person belongs. Hence in the next subsection, the case is evaluated with the original data set divided into Cluster 1, Cluster 4, and the other clusters, and each prediction model is generated for each cluster.

4.3 Evaluation by Grouping of Clusters

The parameters used for the evaluations are the same as for Section 4 and 4.1. Learning data and test data are randomly chosen 10 times from the grouped subset, as samples of combinations of items, and the average of the accuracy is evaluated in the 10 trials. The case when applying the prediction model from the whole data set to each cluster is compared with the case when applying each prediction model from the data set grouped by each cluster to each cluster. The result is shown in Table 9.

Results in Table 9 show that the improvement in accuracy is less than 1% for Cluster 1 and Clusters 2+3, while the improvement for Cluster 4 is about 5% and the total improvement is about 1%.

5 Discussion

Results based on privacy preferences (Section 4.2) show that it is possible to improve the accuracy of the prediction scheme by grouping based on clustering

Table 9. Evaluation in grouping by clustering

Cluster	Using model from all data (Previous scheme [17])			Using models from divided data		
	Accuracy	Ratio	Accuracy × Ratio	Accuracy	Ratio	Accuracy × Ratio
1	0.9698	47.10%	0.456776	0.9738	47.10%	0.45866
2+3	0.7088	38.50%	0.272888	0.7126	38.50%	0.274351
4	0.7767	14.40%	0.111845	0.8237	14.40%	0.118613
		Total	**0.841509**		**Total**	**0.851624**

Table 10. Evaluation in the case dividing Cluster 2 and 3

Cluster	Using model from all data (Previous scheme [17])			Using models from divided data		
	Accuracy	Ratio	Accuracy × Ratio	Accuracy	Ratio	Accuracy × Ratio
1	0.9698	47.10%	0.4568	0.9738	47.10%	0.4587
2	0.7617	21.50%	0.1638	0.7855	21.50%	0.1689
3	0.6420	17.00%	0.1091	0.6768	17.00%	0.1151
4	0.7767	14.40%	0.1118	0.8237	14.40%	0.1186
		Total	**0.8415**		**Total**	**0.8612**

of the answer preferences and generating prediction models for each cluster. However, results based on users' attributes (Section 4.1) show no improvement, this may be, because there are less differences in the answer preferences tendency among the different categories of users. For instance, the answer preferences for those aged in their 20s and 40s show no significant difference, as it can be obeserve in Figure 10.

Regarding the results in Section 4.2, accuracy is improved for Cluster 4; however, no significant improvement is obtained for Cluster 1 and Clusters 2+3. The reason why the accuracy is not improved for Cluster 1 may be that sufficiently high accuracy was already achieved from using the prediction model generated from the whole data set because the ratio of answering "0" (i.e., negative) is very high (about 96.8%). The reason the accuracy is not improved for Clusters 2+3 may be that the prediction model is generated from mixed data with two clusters with different tendencies. Results of the additional evaluations, where Clusters 2+3 are split into Cluster 2 and Cluster 3 from the evaluation are shown in Table 10. These results show an improvement of accuracy of about 2.4% and 3.4% for Clusters 2 and 3, respectively. These results raise the possibility for improving the accuracy by subdividing the clusters even further based on the answer preferences.

6 Conclusions

In this paper, we proposed and evaluated the applicability of SVM-based models to predict default privacy settings of users at the time of registration to service providers. Furthermore, we evaluated the improvement in accuracy of a privacy

setting prediction scheme when the machine learning data sets were grouped based on users' attributes and setting preferences. First, we evaluated the case where the data sets were grouped by gender, age, and type of mobile phone; however, the accuracy was not improved. In terms of privacy protection, this result shows that the collection of additional user attributes could be minimized. We then evaluated our scheme by grouping privacy setting preferences using the K-means algorithm, from the results we could observe an improvement in accuracy. Future work will focus on enhancing the prediction accuracy, for instance by trying a different combination when merging the classes. We also plan to trial the model in real world scenarios; i.e. by integrating our prediction engine to an online service such as a social network site. We plan to analyze the behavior of users and collect their feedback regarding the usefulness and expected accuracy of the prediction engine. We also plan to execute some statistical tests on the significance of this improvement. Additionally, we would also like to investigate the impacts of the predicted settings with respect to the regulatory requirements, such as GDPR or the law of personal data protection in Japan, of service providers and the rights of users.

Acknowledgment

This research work has been supported by JST CREST Grant Number JPMJCR1404, Japan.

References

1. Acquisti, A., Grossklags, J.: Privacy and rationality in individual decision making. Security Privacy, IEEE 3(1), 26 –33 (2005)
2. Backes, M., Karjoth, G., Bagga, W., Schunter, M.: Efficient comparison of enterprise privacy policies. In: Proceedings of the 2004 ACM symposium on Applied computing. pp. 375–382. SAC '04 (2004)
3. Bekara, K., Ben Mustapha, Y., Laurent, M.: Xpacml extensible privacy access control markup langua. In: Communications and Networking (ComNet), 2010 Second International Conference on. pp. 1 –5 (2010)
4. Bell, G.: A personal digital store. Commun. ACM 44(1), 86–91 (January 2001), http://doi.acm.org/10.1145/357489.357513
5. Berendt, B., Günther, O., Spiekermann, S.: Privacy in e-commerce: Stated preferences vs. actual behavior. Commun. ACM 48(4), 101–106 (2005)
6. Buffett, S., Fleming, M.W.: Applying a preference modeling structure to user privacy. In: Proceedings of the 1st International Workshop on Sustaining Privacy in Autonomous Collaborative Environments (2007)
7. Cranor, L.: P3p: making privacy policies more useful. Security Privacy, IEEE 1(6), 50 – 55 (2003)
8. Dehghantanha, A., Udzir, N., Mahmod, R.: Towards a pervasive formal privacy language. In: Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on. pp. 1085 –1091 (2010)

9. Fang, L., Kim, H., LeFevre, K., Tami, A.: A privacy recommendation wizard for users of social networking sites. In: Proceedings of the 17th ACM conference on Computer and communications security. pp. 630–632. ACM (2010)
10. Fang, L., LeFevre, K.: Privacy wizards for social networking sites. In: Proceedings of the 19th international conference on World wide web. pp. 351–360. ACM (2010)
11. Guo, S., Chen, K.: 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). pp. 656–665 (2012)
12. Jensen, C., Potts, C., Jensen, C.: Privacy practices of internet users: self-reports versus observed behavior. Int. J. Hum.-Comput. Stud. 63(1-2), 203–227 (2005)
13. Kelley, P.G., Hankes Drielsma, P., Sadeh, N., Cranor, L.F.: User-controllable learning of security and privacy policies. In: Proc. of the 1st ACM workshop on Workshop on AISec. pp. 11–18. AISec '08 (2008)
14. Kiyomoto, S., Nakamura, T., Takasaki, H., Watanabe, R., Miyake, Y.: PPM: privacy policy manager for personalized services. In: Security Engineering and Intelligence Informatics - CD-ARES 2013 Workshops: MoCrySEn and SeCIHD, Regensburg, Germany, September 2-6, 2013. Proceedings. pp. 377–392 (2013)
15. Madejski, M., Johnson, M., Bellovin, S.: A study of privacy settings errors in an online social network. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on. pp. 340–345 (2012)
16. Mugan, J., Sharma, T., Sadeh, N.: Understandable learning of privacy preferences through default personas and suggestions (2011)
17. Nakamura, T., Kiyomoto, S., Tesfay, W.B., Serna, J.: Personalised privacy by default preferences - experiment and analysis. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy - Volume 1: ICISSP,. pp. 53–62 (2016)
18. Pollach, I.: What's wrong with online privacy policies? Commun. ACM 50(9), 103–108 (2007)
19. Sadeh, N., Hong, J., Cranor, L., Fette, I., Kelley, P., Prabaker, M., Rao, J.: Understanding and capturing people's privacy policies in a mobile social networking application. Personal Ubiquitous Comput. 13(6), 401–412 (2009)
20. Solove, D.J.: Privacy self-management and the consent paradox. Harvard Law Review 126 (2013)
21. Tondel, I., Nyre, A., Bernsmed, K.: Learning privacy preferences. In: Availability, Reliability and Security (ARES), 2011 Sixth International Conference on. pp. 621–626 (2011)
22. Tondel, I.A., Nyre, A.A.: Towards a similarity metric for comparing machine-readable privacy policies. In: Open Problems in Network Security. Lecture Notes in Computer Science, vol. 7039, pp. 89–103 (2012)
23. W3C: The platform for privacy preferences 1.0 (P3P1.0) specificati. In: Platform for Privacy Preferences (P3P) Project (2002)

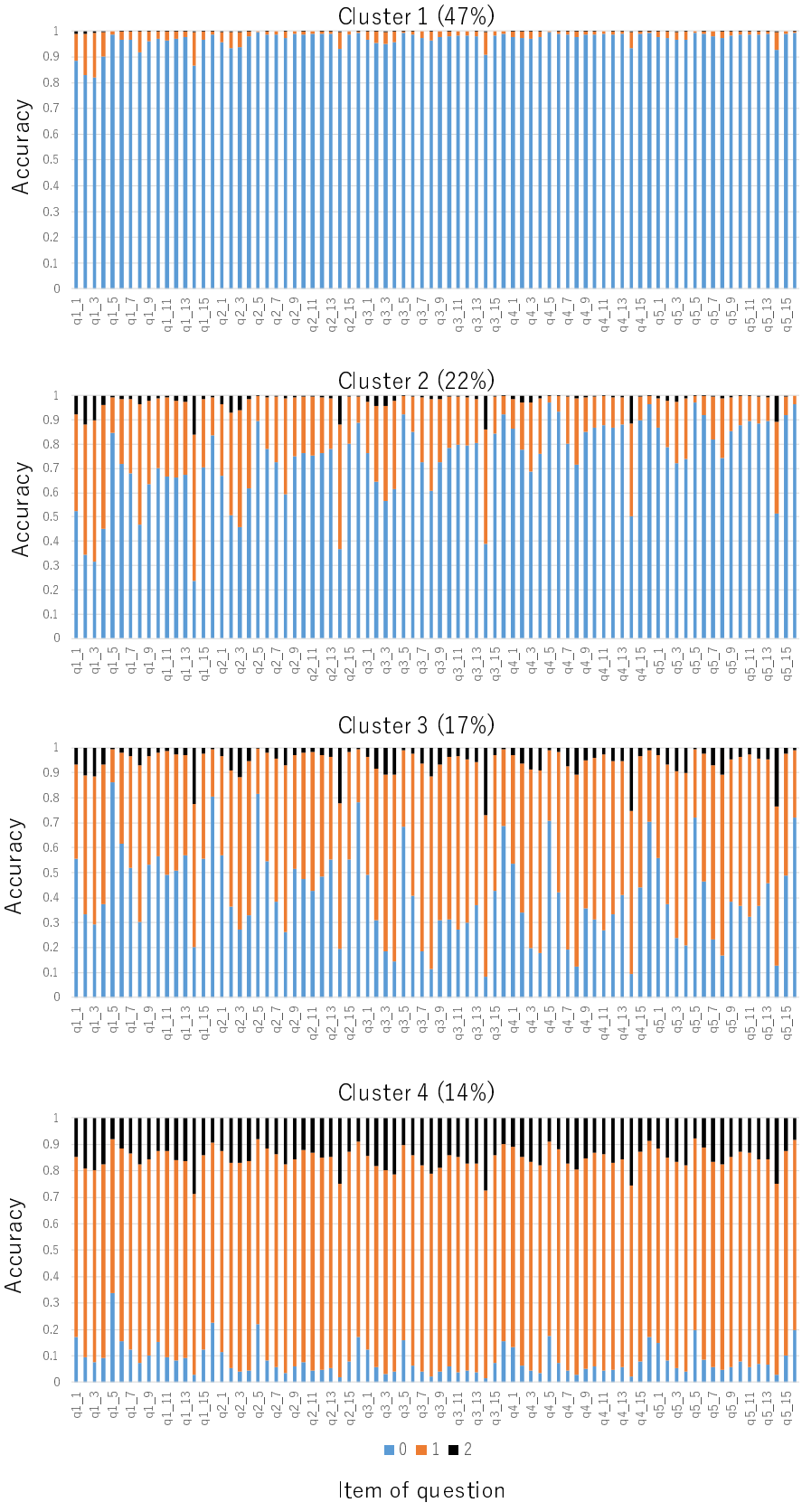

Fig. 9. Tendency of each cluster

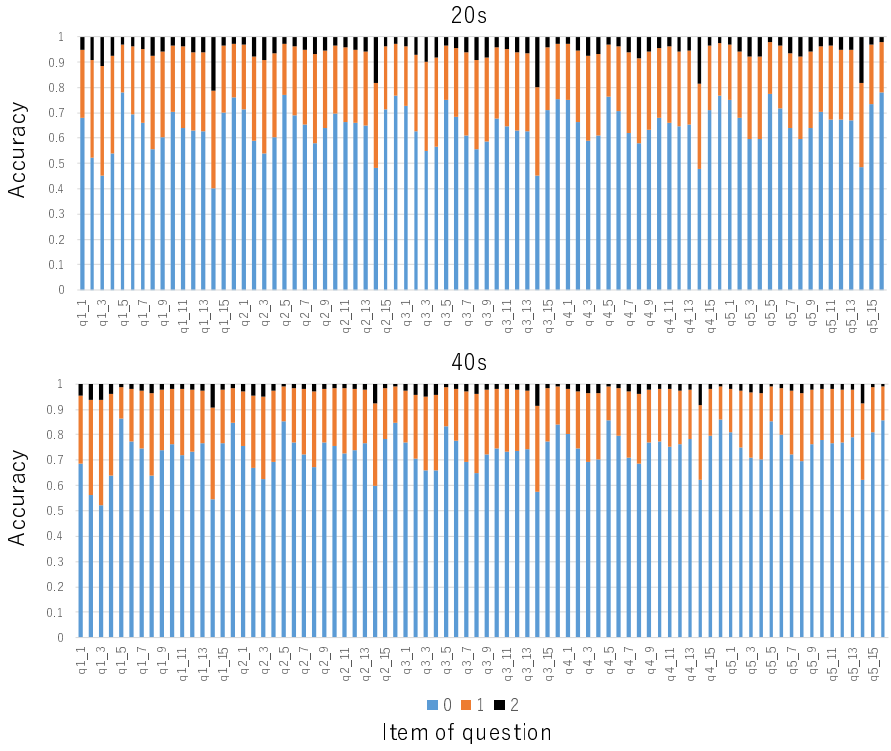

Fig. 10. Tendencies of 20s and 40s