A Privacy Protection System in Context-aware Environment

The Privacy Controller Module

Tahani Hussain
System and Software Development
Kuwait Institute for Scientific Research
State of Kuwait
thussain@kisr.edu.kw

Ranya Alawadhi
Information Science
Kuwait University
State of Kuwait
ranya.alawadhi@ku.edu.kw

ABSTRACT

As context-aware applications are becoming increasingly popular, there are also mounting demands for privacy protection systems. In our work, we propose a context-aware privacy protection system that consists of three modules and aims to recognize the user privacy behavior, classify the context-aware applications and recommend a set of protection action scenarios for the user privacy profile settings. Each module is a challenging problem that needs to be addressed using supervised and unsupervised Machine Learning (ML) algorithms. Part 1 of our work, this paper, consists of deploying hybrid techniques to handle the privacy controller module tasks. Logistic Regression (LR) learning algorithm is integrated with Statistical Method (SM) to recognize user privacy complex activities. The potential of the proposed system is demonstrated using a large-scale real-world dataset provided by institutes from Kuwait, the United States and Belgium. The system demonstration shows promising results with an accuracy of 97.9%.

CCS CONCEPTS

•Security and privacy → Human and societal aspects of security and privacy → Privacy protections •Security and privacy → Software and application security → Social network security and privacy

KEYWORDS

Context-aware, Privacy, Protection, Machine Learning, Classification, Behavior Recognition, Intelligent System.

iiWAS '20, November 30-December 2, 2020, Chiang Mai, Thailand
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8922-8/20/11…$15.00
https://doi.org/10.1145/3428757.3429090

1 Introduction

The term context-aware is defined as “any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves” [1]. Context-aware applications, on the other hand, described as the ability of the application in extracting and using context information to adapt its functionality to the current context in order to provide services that are appropriate to the particular user, place, time, event, etc [2].

As context-aware applications are becoming increasingly popular, there are also mounting demands for flexible and adaptable services. Many of these applications have met the desired quality of service for users due to the substantial efforts of the developers in designing and implementing applications that recognize the user context-aware behavior and recommend the most effective services for users. Typically, a massive collection, organization, processing and sharing of context are required to enhance the quality of services provided by the developers. Often, such contexts contained highly sensitive personal information such as current location, date of birth, or even medical history and financial records. Consequently, user’s privacy concerns and privacy protection system demands have massively increased. In 2018 a survey on the consumer data privacy [3] found that 94% of the responders are worried about their privacy with 57% of them reported that they are even more concerned about their privacy than they were before. Needless to say, the vital issue in a privacy protection system is how to conserve the application’s quality of service while providing an acceptable level of privacy protection. Moreover, user’s privacy behavior diverges based on their personality and sensitivity toward privacy. Due to these and many more reasons, addressing privacy protection issues in context-aware environment is considered both challenging and complex problem.

In previously related work [4], we proposed a privacy-aware method to address the privacy-related concerns on behalf of the users in context-aware environments through automated decision-making processes by monitoring their privacy behavior and personal data usage. The architecture of the method comprises three modules: privacy preferences manager, service classifier and privacy controller. As an extension, we focused on the implementation of the modules utilizing Machine Learning (ML) classification and recognition algorithms supported with statistical methods. In this paper, Part 1 of our work, we present the implementation of Privacy Controller Module (PCM) based on Logistic Regression (LR) learning algorithm supported with Statistical Method (SM) to handle the privacy complex activities recognition to maintain the user’s privacy preferences and sensitivity level to all their context data. These preferences can be changed either by the user or by the hybrid technique based on user privacy behavior with the related services. Needless to say, that the user’s privacy preferences are continuously updated either when new context is requested by the service provider or when a new service is needed and added by the user. Of course, to have an effective privacy protection system, the module must accurately recognize user privacy protection preferences. Thus, the efficiency of the proposed method is assessed using a large-scale real-world dataset provided by institutes from Kuwait, the United Sates and Belgium.

The rest of the paper is structured as follows: Section 2 presents related works; Section 3 presents the architecture of the proposed system; Section 4 shows the real-world dataset used in the experiment; Section 5 presents the hybrid-based methodology proposed for the Privacy Controller Module, while Section 6 discusses the experimental results and Section 7 concludes our paper.

2 Related Work

Among many uses of ML algorithms for addressing context-aware related challenges, recently the user behavior recognition and recommendation system in context-aware environment began to receive a significant research attention.

Concerning the behavior or activity recognition systems, approaches [5, 6] focused on applying supervised learning algorithms to predict user behavior over time, unfortunately these approaches suffer from scalability with respect to diversity and number of user behavior along with huge range of context data. Moreover, the user activity recognition in these approaches analyzes inertial sensors data which are often subject to privacy concern which clearly does not consider users’ privacy protection requirement and specifications. To overcome these issues, works [7, 8] proposed unsupervised learning techniques to recognize user activities from unlabeled data using data mining techniques. However, a large unlabeled dataset along with a certain amount of labeled date are required for an efficient activity recognition method. Authors in [9, 10, 11] combined the strengths of both supervised and unsupervised approaches by proposing semi-supervised learning methods for activity recognition. Their techniques are initialized using small labeled training datasets then a continuous enhancement for the model is conducted using unlabeled data. In recent more related works [12-14], the analysis of various context-aware models considering the temporal, spatial or social contexts for building relevant context-aware systems shows that tree-based classification models have high prediction results over other context-aware models when applied on mobile phone data consisting of multi-dimensional contexts. In [12], the authors analyzed the various context-aware models by considering the temporal, spatial or social contexts for building relevant context-aware systems utilizing individual’s smartphone data. The study shows that tree-based classification models have high prediction results over other context-aware models when applied on mobile phone data consisting of multi-dimensional contexts. In [13], a personal training method has been proposed with the purpose of reducing the flow of user data to the cloud. The aim is to keep the data under user control and thus mitigating the risk of breaching privacy or misusing data. The model is trained over two steps. First, a shared model is built in the cloud using a small set of shared data. Then, the shared model is retained locally using personal data. Finally, the authors in work [14] discussed the use of data-centric social context using mobile phone data to provide personalized services. Specifically, using phone call activities to determine interpersonal relationships and thus provide personalized phone call services to intelligently manage the call interruptions.

Distinguishable from those recognition and recommendation proposals, in our work, we address several challenges considering privacy protection recommendation system using three modules based on hybrid techniques. Firstly, the increasing popularity of context-aware services results in a huge amount of services being employed by users and thus huge data to be protected, which expands the privacy behavior recognition space features significantly. Therefore, the selection of effective privacy protection recommendation features becomes even harder. Moreover, the diversity of context data grows extremely large since users are more likely to access more services at anytime and from anywhere, consequently the privacy recommendation model becomes more complex while satisfying the quality of the service. Finally, the recognition of complex activities like the user privacy is particularly challenging and researchers have investigated the use of ontologies to represent these complex activities. However, ontological recommendation-based recognition is limited to the fact that it must be driven from basic observations (such as current location, environmental conditions, or surrounding objects) that could also be a concern to user privacy. Our proposed method is based on hybrid techniques to solve these problems. It is meant to empower users with appropriate mechanisms to manage and control their privacy preferences by monitoring their behavior. Moreover, the usage of personal data through an automated decision-making process to relieve the user from manually controlling these complex privacy preferences, which has not been a focus of the privacy work mentioned above. To the best of our knowledge, this is the first attempts to deploy a system of three modules based on hybrid-methods to provide privacy protection in context-aware environment.

3 System Architecture

As explained before, we are proposing a context-aware system that manages user context with the aim of protecting user privacy. Our proposed approach will take decisions on behalf of the user with respect to service providers accessing their context. The approach will use the general user privacy preferences per service provider to set the service user privacy preferences. Once these preferences are set, they will be communicated to the operating system, as it is considered the context provider. Thus, the proposed approach does not incur extra overhead when services are provided. The architecture of the proposed privacy protection system comprises three modules: Privacy Preferences Manager (PPM), Service Classifier Module (SCM) and Privacy Controller Module (PCM). Figure 1 illustrates the proposed approach architecture.

Before we discuss the system architecture, we need to identify the set of all user’s context shown in Figure 1. Context data whether explicitly specified by users, like name and date of birth, or implicitly obtained by sensors, like location and temperature, is classified into two categories, *Approximated* data and *Actual* data. The actual data normally requested by the service provider for authentication purposes like the email address and phone number. Thus, it must be provided as it is to the service providers. Approximated data, on the other hand, is used to attain the required service while maintaining user privacy. For example, if we have the year of birth to be 1975, then we can approximate that value to be between 1970 and 1980. Therefore, the data that can’t be approximated and must be released and shared as it is are considered *Actual data*, otherwise they are *Approximate date*.

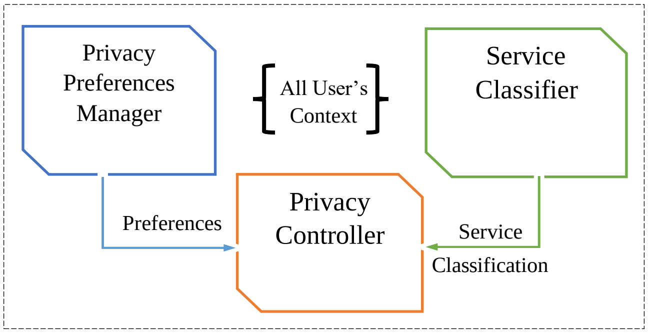

Figure 1: Privacy-Aware System Architecture

The user set their privacy preferences and sensitivity level of their personal data (i.e. privacy preferences) are maintained in PPM. The preferences can be set to *Sharable*, *Not Sharable* or *To Be Determined*. These preferences are continuously updated either by the user or PCM when new context is requested by the service provider or when new service is needed and added by the user. The SCM determines the classification of the service provider in three trust zones (Trusted, Untrusted and Under-investigation). Similarly, to the privacy preferences the service classification is continuously changing based on the service provider’s usage behavior for the user’s data through systematized procedure that considers how the context is used and shared in the ongoing activities of the service providers. Finally, PCM is responsible for notifying the users when their personal data are disclosed by service providers. Also, it takes full control of the context released and shared to service providers while sustaining the user privacy and without disturbing the services provided. In this work, we focus on the PCM which is the core of our proposed system, more information can be found in [4].

Generally, the system triggered by a request received by PCM from a service provider to access a set of the user context. Then, the PCM will act based on the following three criteria:

1. **Context data type** (approximated data or/and actual data).
2. **Preferences** provided by PPM.
3. **Service Classification** provided by SCM.

PCM set a decision regards releasing and sharing the requested context. The decision falls under one of four options: Allow, Deny, Approximate or Change Preference. Allow decision, clearly, permits releasing and sharing the requested context to the service. Deny decision, on the other hand, prevents service from accessing the requested context. The approximation decision only applies for the *Approximate date* and it permits releasing and sharing the requested context as approximation context. The last decision, which is Change Preference, recommends the user to change their privacy preferences setting or in some cases apply the changes automatically if the service provider is classified as *Trusted*. The decision matrix is shown in Table 1 and elaboration of this matrix is depicted using flowchart in Figure 2.

Context / Preference	**Service Provider Classification**		
	Trusted	Untrusted	Other
Actual / Sharable	Permit	Recommend Deny	Recommend Deny
Actual / Not Sharable	Deny	Deny	Deny
Approx. / Sharable	Permit	Recommend Approximate	Recommend Approximate
Approx. / Not Sharable	Deny	Deny	Deny
Approx. / Approximate	Permit	Rec. Deny	Approximate

Table 1. Decision Matrix

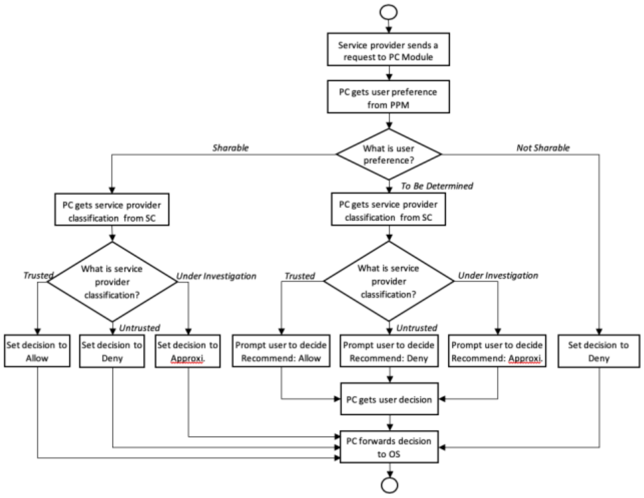
Service provider sends a request to PC Module
PC gets user preference from PPM
What is user preference?
Sharable
Not Sharable
To Be Determined
PC gets service provider classification from SC
PC gets service provider classification from SC
What is service provider classification?
What is service provider classification?
Trusted
Untrusted
Under Investigation
Trusted
Untrusted
Under Investigation
Set decision to Allow
Set decision to Deny
Set decision to Approxi.
Prompt user to decide Recommend: Allow
Prompt user to decide Recommend: Deny
Prompt user to decide Recommend: Approxi.
Set decision to Deny
PC gets user decision
PC forwards decision to OS

Figure 2: Decision flowchart

As illustrated in Table 1 and Figure 2, PCM release and share the requested context if the user preference is *Sharable* and the service is trusted either as actual or approximated data. Whereas if the user preference is *Not Sharable* then PCM deny releasing and sharing the requested context regardless the service classification or context data type. For all other cases, PCM promotes the user with a recommendation to either permit or deny releasing and sharing the requested context. Approximating the context if possible is also the responsibility of the PCM. Roughly speaking, the user privacy preferences and the service classification are variables that change continuously. Consequently, these changes directly affect the PCM recognition and recommendation decision. Thus, to have an effective privacy protection system, PCM must accurately recognize user privacy activities. In order to perform that, we have defined a statistical method along with LR learning algorithm that models these activities and tasks. The methodology of utilizing these two techniques to implement PCM is described later.

4 Real-World Dataset

For our present experiment we used a real-world dataset collected by volunteers from University of Colorado Boulder (CU Boulder, USA), Kuwait Institute for Scientific Research (KISR, Kuwait) and Katholieke Universiteit Leuven (KU Leuven, Belgium). This dataset collected from 756 users deploying 2,093 context-aware services from 5 main categories: medical, social network, shopping, education and utilities. A total of 1,360,864 records were collected over 8-months period from 01-05-2019 to 31-12-2019. We have excluded the dataset collected from the users consisting gray period (i.e. no date records from using the services for one week). In average, 1,800 privacy activities collected per user for 5 minutes every day to assure the consistency distribution of the dataset. Moreover, a minimum of 1,050 services were restricted to be used by all the users during this period to assess the efficiency service classification and privacy recommendation system. The collected dataset is divided into four subsets ***U***, ***S***, ***C*** and ***A***. The ***U*** subset consists of all user relevant variables from the collected dataset (i.e. user gender, age, country ...etc.). Similarly, ***S*** subset consists of all service-related variables (i.e. provider, developer, cost …etc.). All the variables from the collected dataset that is related to user context have been listed in subset ***C***. Finally, subset A contains the time-base user activities and behavior. For the convenience of PCM modeling, the intersection of subsets (***U***, ***S***, ***C*** and ***A***) has been considered and so-called subset ***P***. Other variables from subsets ***U***, ***S***, ***C*** and ***A*** are not considered in this module and are out of this paper scope. Parts 2 and 3 of our work will address these subsets variables and more. The description of variables in this subset is shown in Table 2.

Variable	Type	Description
Date_Time	Date/Time	Date and time of the user activity
Context_ID	Text	Context name or reference
Context_Cat	Binary	1=Approximated data and 0=Actual data
User_ID	Number	User identification number
User_Pref	Binary	1=Sharable; 0=Not Sharable and Null=Other
Service_ID	Text	Service name
Service_Clas	Binary	1=Trusted; 0=Not Trusted and Null=Other
Service_Cat	Integer	1=Medical, 2=Social network, 3=Shopping, 4=Education and 5=Utilities
Activity	Integer	User behavior activity; 1=Allow, 2=Deny, 3=Approximate and 4=Change

Table 2. Variable Description of subset *P*.

User_Id, Service_Id and Context_Id variables are nominal and the rest of the variables in subset ***P*** are event-base variables. As shown, *Activity* variable represents the user privacy behavior on the context based on the service and consists of four action: *Allow*, *Deny*, *Approximate* and *Change* preference. Thus, the LR learning algorithm need to predict the user privacy behavior. Based on that, ***P*** is divided into four sets: training, validation, prediction and optimization to assess the efficiency of LR learning algorithm for PCM. The dataset division considering sets and dates is illustrated in Figure 3. As shown the dataset is partitioned using simple random method, where 60% of the data is used for training, 35% for validation and 5% for prediction and optimization. Moreover, the first two sets (training and validation) contain the user variables collected from 01-05-2019 to 16-12-2019. LR model is trained on these two sets to predict the user privacy behavior (*Activity*) for the period from 17-12-2019 to 27-12-2019. Afterword, the LR model is optimized on the dataset collected from 28-12-2019 to 31-12-2019.

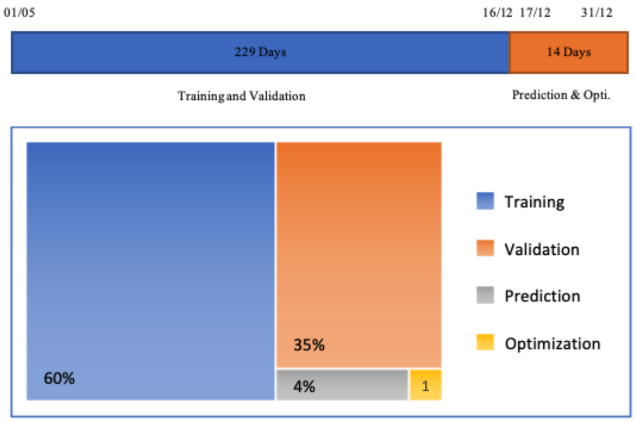

Figure 3: Dataset Partition based on Percentages

After completing LR training, SM is applied on the learned models to achieve a better predictive performance. This step found to be helpful, the prediction results are used to statically filter the prediction activities that found not likely to be taken by the user.

5 Hybrid-Techniques Methodology for PCM

We believe that to build an effective and automated privacy protection system, one must adopt the same method used to menace this privacy, which is the context-aware methods. For example, the context-aware methods help PCM by considering the previous privacy behavior activity performed by the user for the same context type and service. Another example, context-aware methods help the PCM in assembling the user's privacy preferences and context data revealing practice to recommend a set of protection action scenarios for the user privacy profile settings. Motivated by that, we are proposing a context-aware methodology that manages user context with the aim of protecting user privacy for the privacy controller module, see Figure 4.

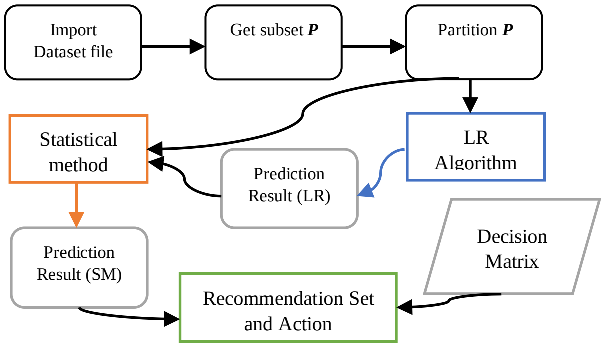

Figure 4: PCM Methodology

Our proposed methodology for implementing PCM consists of two techniques, Logistic Regression (LR) learning algorithm and Statistical Method (SM). This hybrid-technique is utilized to perform tasks associated to PCM which includes dataset processing, partitioning, activity prediction and recommendation.

5.1 Logistic Regression (LR) learning algorithm

Logistic Regression (LR) is supervised learning algorithm and considered as the most important probabilistic discriminative classifier techniques employed for analyzing binary and proportional response datasets. One of the main advantages of LR is that it produces piece-wise linear decision boundary for multi-class classification. Another advantage is that learning parameters for LR model requires numerical optimization, thus unconstrained optimization issues can be addressed using LR. In our proposed system, we use simple LR learning algorithm to recognize and predict the user privacy behavior activities. We trained our parameters using gradient descent algorithm to reduce the computational complexity and optimize prediction errors.

Let $\boldsymbol{P_V}$ be a subset of set $\boldsymbol{P}$ that is represented by matrix of size $\boldsymbol{M}$ by $\boldsymbol{N}$, where $\boldsymbol{M}$ is the number of services and $\boldsymbol{N}$ is the user privacy action on the requested context. $\boldsymbol{P_V}$ contains every privacy preference for the context from user but not the privacy behavior activity (***B***), which is collected based on the applied service and its classification. ***B*** is the user privacy behavior taken by user for the requested context by the service. In our dataset, this is represented by the *Activity* variable which is an integer in the range from 1 to 4. We will consider the subset of $\boldsymbol{P_V}$ that intersected with ***B***. Let ***b*** be the user privacy behavior predicted by LR which consists of the same values as in ***B***. The objective of our model is to minimize the mean squared error (***E***) between the collected behavior (***B***) and the predict (***b***) one as shown in Equation 1.

$$\frac{1}{M}\sum(B-b)^2 \qquad (1)$$

The above equation is used to assess the LR efficiency. For modeling the learning algorithm, we use Equations (2) and (3).

$$R_t(y \mid x) = \frac{1}{1 + e^{-f(x)}} \qquad (2)$$

$$f(x) = \alpha_0 + \alpha_1 * x + \epsilon \qquad (3)$$

Where $\boldsymbol{y}$ is the output, $\boldsymbol{x}$ is the input variable for time $\boldsymbol{t}$ and $\boldsymbol{\varepsilon}$ is the error between the predicted values and the actual one. $\boldsymbol{\alpha_0}$ is a constant that does not have an intuitive interpretation while $\boldsymbol{\alpha_1}$ is the amount by which the output $\boldsymbol{y}$ changes when input $\boldsymbol{x}$ changes. In RL, the probability is always between 0 and 1. It is compared with the respect to the selected threshold to assess the number of points classified correctly. Most commonly used threshold is 0.5 and we have considered this value in our methodology.

5.2 Statistical Method (SM)

Now when considering user privacy activity, LR prediction does not depend on the prediction of another activity. This is due to the fact that normally users continuously perform different activities depend on the context and service, but they tend to perform the same activity for the same context and service before changing activity. In this work, we integrated Statistical Method (SM) to address this issue and to improve LR learning algorithm.

In this paper, we use SM to achieve two main objectives. The first one is to statistically predict the most likely *Activity* (***b'***) to be perform by the user based on the given context, privacy preference and service classification input set ***P*** during dates $\boldsymbol{d_1}$ to $\boldsymbol{d_2}$. Considering this, for context ***i*** and service ***j*** the single corresponding user *Activity* $\boldsymbol{b_{ij}}$ is predicted with probability of $\boldsymbol{s_{ij}}$ such that the sum of s is equal to 1 as shown in Equation 4.

$$\sum_{j=1}^{M} s_{ij}\ of\ b_{ij} = 1 \quad for\ i = 1,2\,..N \qquad (4)$$

The second objective is to achieve a better predictive performance for the prediction result from LR learning algorithm. Basically, the prediction probability of LR ($\boldsymbol{P_{lr}}$) and SM ($\boldsymbol{P_{sm}}$) will be used to find better prediction ($\boldsymbol{P_o}$) probability compared to the actual activity as shown in Equation 5.

$$P_o = (\mu_1 * P_{lr}) + (\mu_1 * P_{sm}) \qquad (4)$$

Where $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are positive constants less than 1 and represent the weight rate of LR and statistical method, respectively.

To evaluate the best prediction probability, we use F1 score [15] which is a weighted average of precision (δ) and recall (λ) rate. The definitions of these two rates (δ and λ) is given in Equation 6 and 7 respectively. While Equation 8 is used for the F1 score.

$$\delta = \frac{|\{B\} \cap \{b\}|}{|\{B\}|} * 100 \qquad (6)$$

$$\lambda = \frac{|\{B\} \cap \{b\}|}{|\{b\}|} * 100 \qquad (7)$$

$$F1 = 2 * \frac{\delta * \lambda}{\delta + \lambda} \qquad (8)$$

Where ***{b}*** is the prediction set of the user privacy behavior result from LR, statistical method or from both method and ***{B}*** is the collected and actual user privacy behavior activity set.

That is to say, the best prediction set with best probability of these three values ($\boldsymbol{P_{lr}}$, $\boldsymbol{P_{sm}}$ or $\boldsymbol{P_o}$) will be considered as the prediction result for the PCM recommendation process. In other words, the best prediction result with higher probability and performance will be considered by PCM to perform the privacy decision for releasing and sharing the context along with the necessary recommendation and changes related to the preferences and service classification.

6 Experimental Result

For investigating privacy controller module (PCM) of the privacy protection system in the context-aware environment that is proposed in this paper, we performed extensive experiments to evaluate the prediction performance of the proposed hybrid technique. Moreover, we investigated variations in performance using standalone LR algorithm compared to the hybrid methodology.

Before we review the experimental results, we need to demonstrate the privacy *Activity* performed by users in the collected dataset. The histogram shown in Figure 5 demonstrates the frequency distribution of different activities performed by users according to the collected dataset. As shown, both *Allow* and *Deny* activities govern a total of 74.4% over all recorded activities with 35.7% aa *Allow* and 38.7% as *Deny*. *Approximate* activity found to be 19% of the total recorded activities while 6.6% related to changes in privacy preferences. Thus, most of the recorded activities (93.4%) tend to perform the same action for the same context and service.

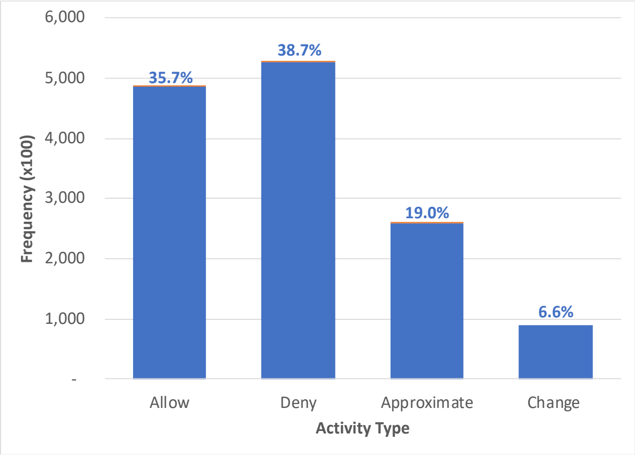

Figure 5: Frequency of different user Activities.

Overall, we found that the user privacy behavior depends mostly on the service category rather than the context type, see Figure 6 and 7.

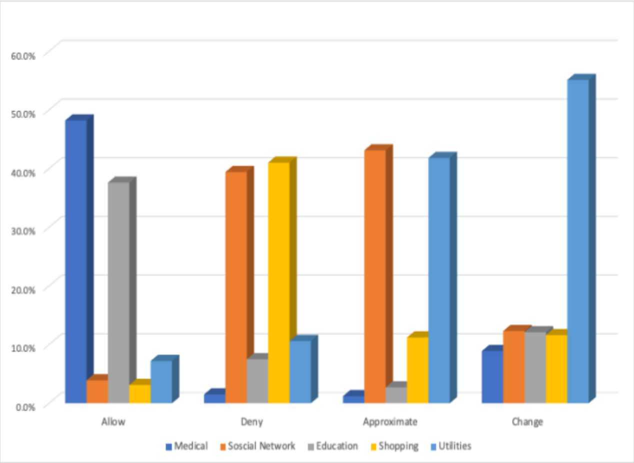

Figure 6: User Activities based on Service Category.

As shown in Figure 6, users prefer to allow releasing and sharing the context if the requested services are from medical or education categories. The highest percentage of releasing and sharing the requested context by the user was for the medical service with 48.2%. The next higher percentage, which is nearly 38%, of the Allow activity are for the educational services. While for the shopping and social network service, users mostly perform *Deny* activity with 41% and 39% respectively. We also found that almost 85% of the activities for approximating the requested context is performed by the user for social network and utilities services combined, and 55.1% of privacy changes were recorded for the utility's services alone.

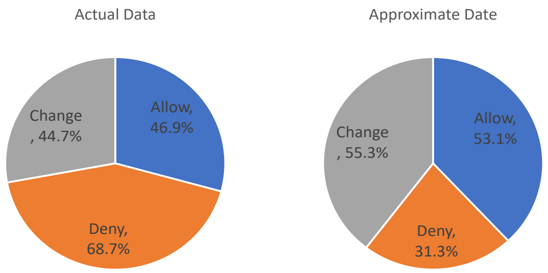

Figure 7: User Activities based on Context data type.

On the other hand, Figure 7 illustrates the user privacy activity distribution based on the context data type. Needless to say, the *Approximate* activity is not applicable for the actual context data for that reason it wasn't considered in this observation and data analysis. Based on the figure, user privacy preferences (*Allow* and *Change*) slightly varies if the context data type is considered. However, the *Deny* action is mostly recorded for the actual data context with nearly 69%. That reflects the reality that most of the user tend to protect their actual data context and prefer to release them only if necessity.

Our experiments aim at evaluating the prediction performance of LR learning algorithm and hybrid methodology, together with comparing the results with the respect to the F1 score. Next figures, Figures 8 and 9, show the performance comparison between LR learning algorithm and the proposed hybrid methodology (i.e. LR integrated with SM). We can see that LR shows good performance in predicting the user activity with nearly 92.7% accuracy. The algorithm score 80.8% in Precision (δ), 67.5% in Recall (λ) and 73.5% in F1.

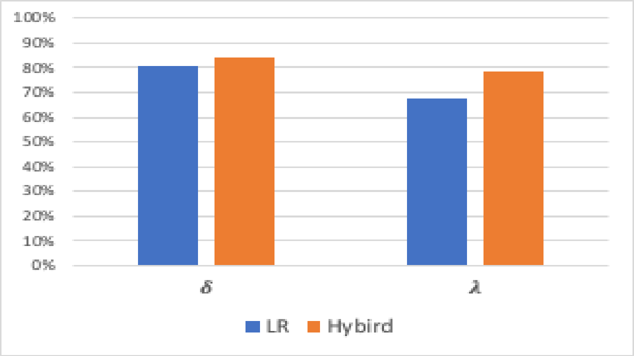

Figure 8: Precision (δ) and Recall (λ) scores.

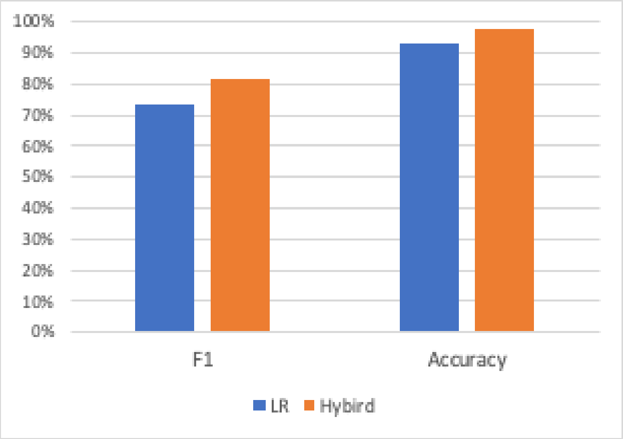

Figure 9: Performance scores for LR and Hybrid algorithms.

However, one can observe a significant improvement in performance for the hybrid methodology compared to the LR algorithm for the same dataset. The final performance scores are 3.84% and 17% better than before for δ and λ correspondingly. Overall, hybrid methodology preform prediction with accuracy up to 97.9% with F1 score of 81.3%. In another word, the integrating SM in LR learning method enhance the accuracy of the activity prediction by 5.61%.

The setup parameters for controlling LR learning algorithm [15] and the SM are shown in Table 3.

Algorithm	Parameter	Setup value
LR	No. Iteration	10,000
LR	Convergence Error	1 E -5
LR	Threshold	0.5
SM	μ_1	0.77
SM	μ_2	0.23

Table 3. Experiment setup of algorithm parameters.

The experiments are conducted on Linux platform using Python programming language. The training and validation process for the investigated algorithm is about 45 minutes running on a 48-cores machine. Yet, using the same machine, the prediction process running time is about 9 minutes.

From the above experiments results, it can be observed that the hybrid-based methodology for implementing privacy controller model proposed in this paper are effective for recognizing and recommending user privacy behavior in context-aware environment.

7 Conclusions and Future Work

In this paper, a privacy protection system in context-aware environments is presented. Part 1 of our proposed system is illustrated in this paper, which includes deploying hybrid methodology based on a statistical technique (SM) and Logistic Regression (LR) learning algorithm for Privacy Controller Module (PCM). It also includes an innovative architecture along with implementation for its modules. The efficiency of the proposed methodology has been assessed through large-scale real-world dataset provided by institutes from Kuwait, the United States and Belgium. Results confirm the efficiency of our approach with respect to privacy recognition and protection classification methods with accuracy of 97.9%. The proposed methodology compared with LR algorithm, the performance measurement F1 score of this increased by 11% and the accuracy enhanced by 5.61%.

Currently, we have completed the implementation of the other two modules (i.e. Privacy Preferences Manager and Service Classification) of the proposed system however they were not tested yet. Our next step involves testing the efficiency of these two modules as well as enhance the efficiency of PCM presented in this work. Near future work will mainly be focused on assembling these three modules to integrate the automated privacy protection system and deploy it in a real context-aware environment to assess its efficiency.

ACKNOWLEDGMENTS

The authors of this work would like to thank the volunteers from University of Colorado Boulder (CU Boulder, USA), Kuwait Institute for Scientific Research (KISR, Kuwait) and Katholieke Universiteit Leuven (KU Leuven, Belgium) that collaborated to the collection of data used in our experiments.

REFERENCES

[1] Gregory Abowd, Anind Dey, Peter Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a Better Understanding of Context and Context-Awareness. In Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing (HUC ’99). Springer-Verlag, Berlin, Heidelberg, 304–307.

[2] Byun Hee and Cheverst, Keith. 2004. Utilizing Context History To Provide Dynamic Adaptations.. Applied Artificial Intelligence. Vol. 18. 533-548. DOI:https://doi.org/10.1080/08839510490462894.

[3] Janrain, 2018. https://www.janrain.com/resources/industry-research/consumer-attitudes-toward-data-privacy-survey-2018, last accessed 2020/06/28.

[4] Alawadhi, Ranya and Hussain, Tahani. 2019. A Method Toward Privacy Protection in Context-Aware Environment. Procedia Computer Science. 151. 659-666. DOI:https://doi.org/10.1016/j.procs.2019.04.088.

[5] Jennifer Kwapisz, Gary Weiss, and Samuel Moore. 2011. Activity Recognition using Cell Phone Accelerometers. SIGKDD Explor. Newsl. Vol. 12, 74–82. DOI:https://doi.org/10.1145/1964897.1964918

[6] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition using Body-worn Inertial Sensors. ACM Comput. Surv. 46, 3, Article 33, 33 pages. DOI:https://doi.org/10.1145/2499621

[7] Yongjin Kwon, Kyuchang Kang and Changseok Bae. 2014. Unsupervised Learning for Human Activity Recognition Using Smartphone Sensors. Expert Systems with Applications, Vol. 41, 6067-6074, https://doi.org/10.1016/j.eswa.2014.04.037.

[8] Dorra Trabelsi, Samar Mohammed, Faicel Chamroukhi, Latifa Oukhellou and Yacine Amirat. 2013. An Unsupervised Approach for Automatic Activity Recognition Based on Hidden Markov Model Regression. Automation Science and Engineering, IEEE Transactions on. Vol. 10, 829-835, DOI:https://doi.org/10.1109/TASE.2013.2256349.

[9] Zahraa Abdallah, Mohamed Gaber, Bala Srinivasan and Shonali Krishnaswamy. 2018. Activity Recognition with Evolving Data Streams: A Review. ACM Comput. Surv. Vol. 51, 4, Article 71, 36 pages.DOI:https://doi.org/10.1145/3158645.

[10] Shashi Phoha, Nurali Virani, Pritthi Chattopadhyay, Soumalya Sarkar, Brian Smith and Asok Ray. 2014. Context-aware Dynamic Data-driven Pattern Classification. Procedia Computer Science, Vol. 29, 1324-1333. https://doi.org/10.1016/j.procs.2014.05.119.

[11] Özgür Yürür, Chi Harold Liu, Zhengguo Sheng, Victor Leung, Wilfrido Moreno and Kin K. Leung. 2016. Context-awareness for Mobile Sensing: A Survey and Future Directions. IEEE Communications Surveys & Tutorials, Vol. 18, 68–93. doi: 10.1109/COMST.2014.2381246.

[12] Iqbal Carker, A. S. Kayes and Paul Watters. 2019. Effectiveness Analysis of Machine Learning Classification Models for Predicting Personalized Context-

aware Smartphone Usage. Journal of Big Data, Vol. 6, 57-85. https://doi.org/10.1186/s40537-019-0219-y.

[13] Sandra Servia-Rodriguez, Liang Wang, Jianxin Zhao, Richard Mortier and Hamed Haddadi. 2017. Personal Model Training under Privacy Constraints. arXiv, Vol. 40, 24–38.

[14] Iqbal Sarker. 2018. Understanding the Role of Data-Centric Social Context in Personalized Mobile Applications. EAI Endorsed Transactions on Context-aware Systems and Applications, Vol. 5, 155671. https://doi.org/10.4108/eai.18-6-2018.155671.

[15] David Hosmer Jr, Stanley Lemeshow and Rodney Sturdivant. 2013. Applied Logistic Regression (3rd. ed.). Wiley, New York, NY.