Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 184 (2021) 52–59

Procedia Computer Science

www.elsevier.com/locate/procedia

The 12th International Conference on Ambient Systems, Networks and Technologies (ANT)
March 23-26, 2021, Warsaw, Poland

The Efficiency of Learning Methodology for Privacy Protection in Context-aware Environment during the COVID-19 Pandemic

Ranya Alawadhi[a], Tahani Hussain[b,*]

[a]*Kuwait University, P.O.Box 5969 Safat 13060, Kuwait*
[b]*Kuwait Institute for Scientific Research, P.O.Box 24885 Safat 13109, Kuwait*

Abstract

When the COVID-19 coronavirus hit, the context-aware application users were willing to relax their context privacy preferences during the lockdown to cope their lives while staying home. Such disturbance in the privacy behavior affected the performance of Machine Learning (ML) algorithm that is trained on normal behavior. In this paper, we present the impact of the pandemic on the efficiency of the learning algorithm implementation of a privacy protection system. The system is composed of three modules, in this work we focus on Privacy Preferences Manager (PPM) module which is implemented using hybrid methodology based on a Statistical Model (SM) and Logistic Regression (LR) learning algorithm. The efficiency of the hybrid methodology is assessed using two real-world datasets collected prior and during the COVID-19 pandemic. The results show that the pandemic significantly impacted the efficiency of the hybrid methodology by 13.05% and 15.22% for the accuracy and F1 score respectively.

Keywords: Privacy, Behavior Recognition, Context-aware, Machine Learning, Logistic Regression, COVID-19, Protection, Intelligent System;

1. Introduction

As context-aware applications are becoming increasingly popular, there are also mounting demands for flexible and adaptable services. While these applications allow users to receive personalized services, sharing context-data

* Corresponding author. Tel.: +965-2498-9717; fax: +965-2498-9409.
E-mail address: thussain@kisr.edu.kw

10.1016/j.procs.2021.03.017

with such applications can leads to privacy breaches. Many of these applications access more sensitive data than necessary. For example, it was found in a study that analyzed 843 apps that 56% of the apps that access "Multimedia Storage" do not need this data to properly work, 33% of the apps that access "Wi-Fi Connections" do not need this data to properly work, and 24.4% of the apps that access "Contact" do not need this data to properly work [1]. Users expresses dismay and outrage when confronted with the behavior of these applications [2]. Consequently, protecting user privacy includes protecting context-data. However, relying solely on users to configure privacy preferences may not achieve the optimum privacy level the users seek as lake of knowledge negatively influences the privacy perception [1]. Since user's privacy behavior is shaped by their personality and sensitivity toward privacy, addressing privacy protection issues in context-aware environment is considered both a challenging and a complex problem.

Some work has been proposed to predict and set user privacy preferences [6-8]. One work developed privacy profiles to configure permissions [6], while another used crowdsourcing [7]. A more recent work used matrix factorization with local differential privacy [8]. Other approaches [9-11] utilized machine learning algorithms to address behavior or activity recognition challenges. In [9], they proposed semi-supervised learning methods for activity recognition. Whereas in [10], researches employed a deep learning model for unsupervised activity recognition. The authors in [11] analyzed the effectiveness of various machine learning classification models for predicting personalized usage utilizing individual's phone log data. Despite the success of machine learning for developing automated and intelligent systems, no work yet properly covered the impact of COVID-19 pandemic on the well-known pre-developed machine learning based systems, which is considered in this work.

COVID-19 pandemic has forced people to alter their daily behavior and consider a series of extensive measures to cope with the locked down procedure enforced by many governments. People were forced to download and use different context-aware mobile applications to carry on with their lives while staying home. Consequently, users were willing to relax their context privacy preferences in trade off using the optimum solution to endure the predicament. Unfortunately, such disturbance in user behavior affected automated and intelligent system causing degrading the efficiency of the learning machine algorithms that run behind the scenes. This is due to the fact that these learning algorithms is trained on normal behavior, and during the pandemic these behaviors have changed, some changed little while others dramatically. Thus, human behavior recognition remains challenging and significant area in developing automated systems.

In a previously related work [3], a context-aware privacy protection system was proposed to automate the users' context sharing decision-making process by monitoring the user privacy behavior and personal data usage. The main objective is to control the release of context-data to protect user privacy. The tasks of the proposed system are carried out by three modules: Privacy Preference Manager (PPM), Service Classifier (SC), and Privacy Controller (PC). Each of these modules is addressed with a Machine Learning (ML) algorithm. Due to the complexity of the system, the implementation of the whole system is divided into three parts. Each part focuses on the implementation details, experiment environment and results of one specific module. The implementation of the PC module, part 1 of our work, was presented in [4]. The module was implemented using a hybrid methodology based on a Statistical Model (SM) and Logistic Regression (LR) learning algorithms. In this paper, part 2 of our work, we present a modified implementation of this hybrid methodology for the PPM module. We also assess its efficiency using a large-scale real-world dataset provided by institutes from Kuwait, United Sates and Belgium. Finally, we distinguish our contribution by investigation the impact of the COVID-19 pandemic on the efficiency of the proposed hybrid methodology using two data sets, one is collected prior and the other during COVID-19 pandemic. These data sets are selected to demonstrate the efficiency of the prediction of a model trained with a normal user behavior when significant abnormal changes occur in user behavior.

The rest of the paper is organized as follows: Section 2 presents the system architecture, Section 3 discusses the learning methodology for the PPM module, Section 4 describes the datasets, Section 5 discusses the results, and finally Section 6 concludes the paper and provides directions for future work.

2. System Architecture - PPM Module

The architecture of proposed privacy-aware protection system, in this work, consists of three modules: Privacy Preference Manager (PPM), Service Classifier (SC), and Privacy Controller (PC), see Fig. 1. Briefly, the system is triggered by a request received by PC module from a service provider to access a set of the user context. Then, PC

will act based on the following three criteria: **User Context** set, **Preferences** provided by PPM module and **Service Classification** provided by SC module. The decision falls under one of four options: Allow, Deny, Approximate or Change. Allow and Deny decisions, clearly, permits or prevent service provider from accessing the requested context respectively. Approximate decision only applies on the Approximated subset and it permits sharing an approximation of the requested context. The last decision, which is Change, 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. In this work, we focus on the PPM module, more information about the system architecture and methodology can be found in [3].

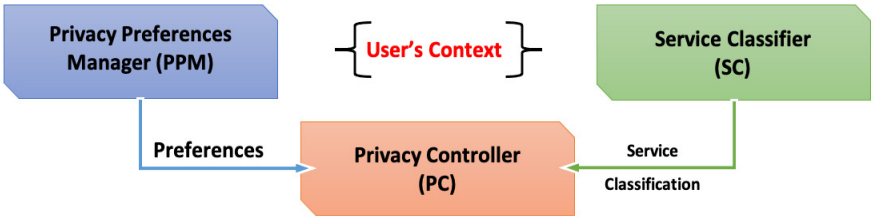

Fig. 1. Privacy-Aware Protection System Architecture.

PPM module is associated with two variables **User Context** and **Preferences** as shown in Fig. 1. **User Context** is a set that consists of two subsets of explicit (i.e. specified by the user) and implicit (i.e. obtained by the sensors) context data: *Actual and Approximated*. The context that must be released and shared as it is, usually, for authentication purposes is classified in the *Actual* subset, otherwise it is classified in the *Approximated* subset. **Preferences** are the privacy sensitivity level of each context-data. The context privacy preference can be set to *Sharable*, *Not Sharable* or *To Be Determined*. *Sharable* preference means releasing the context to the service provider if requested. *Not Sharable* preference means explicitly don't release the context to the requester. The last preference type, *To Be Determined*, means that the user/system would like to have control over a particular context depending on the application in use.

Generally, PPM module is responsible for maintaining and continuously updating the **Preferences** set. This module is responsible to act on behalf of users for setting, modifying and updating the preferences based on user behavior and interaction with service provider (*Activity*) or on PC module recommendations. PPM module is initiated when it receives a new context request or when a new context is added to the **User Context** set.

3. Learning Methodology for PPM Module

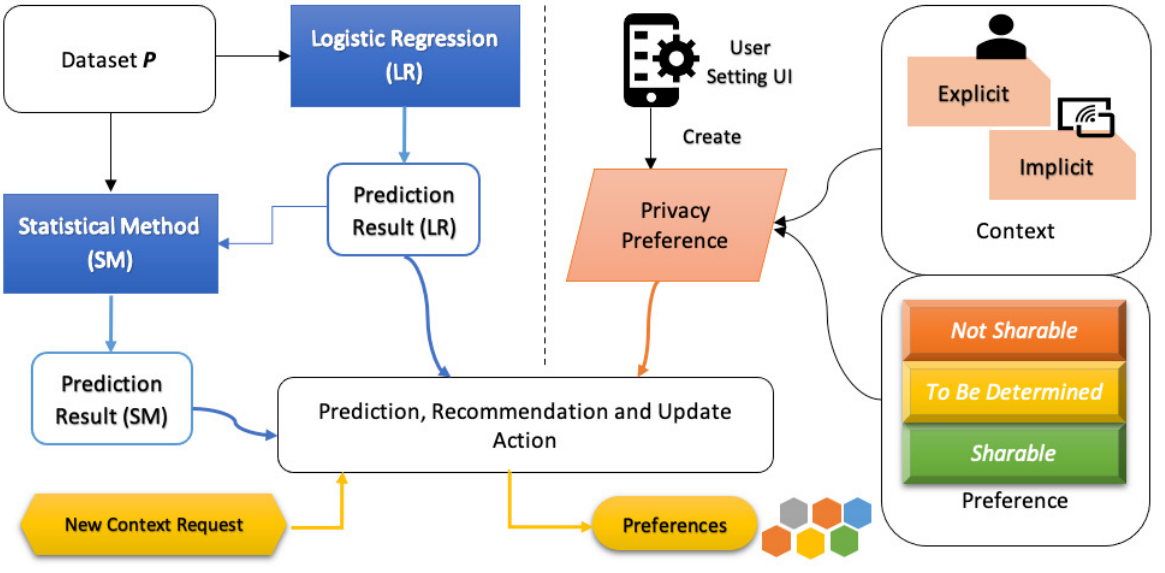

Fig. 2. PPM Module Methodology.

As stated before, PPM module is responsible for maintaining and continuously updating the **Preferences** set. Roughly speaking, this continuous changes directly affect the user behavior recognition system implemented for PPM modules. Thus, for efficient privacy preferences management, the module must accurately recognize the user privacy preference behavior. For that reason, PPM module is implemented by utilizing a hybrid methodology of two techniques: Logistic Regression (LR) learning algorithm and Statistical Method (SM), see Fig. 2. This hybrid mythology is efficient with an accuracy of 97.9% when adopted for PC module [4]. In this work, minor adjustment has been applied to this methodology for PPM module.

First, we use LR learning algorithm to predict the user privacy preference for a context. Then, we use SM to statistically investigate the most likely users' privacy preferences of personal contexts to predict recommendations of user privacy preference. Principally, the main recommendation and user preference profile action are factors that will be considered by PPM to select the best prediction probability from using LR only or the hybrid methodology (LR supported by SM).

In our proposed system, we use simple LR learning algorithm to recognize and predict the user privacy preferences behavior activities. We trained our parameters using gradient descent algorithm to reduce the computational complexity and optimize prediction errors. The objective of our model is to minimize the mean squared error between the collected dataset (***P***) and the prediction by LR ($\boldsymbol{P_{LR}}$) for context set ***M*** as shown in Equation 1.

$$\sum_{i=1}^{M} (P - P_{LR})^2 \tag{1}$$

In LR, the probability is compared with a threshold to assess the number of points classified correctly, we set the threshold in our methodology to 0.5.

Since different privacy preference settings depend on the context and service in use, SM is integrated with LR to statistically predict ($\boldsymbol{P_{SM}}$) the most likely preference performed by the user $\boldsymbol{d_{ij}}$ based on the given context ***i*** data type *j* (*Actual and Approximated*), see Equation 2.

$$\sum_{i=1}^{M} d_{ij} = 1 \qquad for\ j = 1, 2 \tag{2}$$

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 recorded user privacy preferences behavior and action. 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 PPM module recommendation and update action. Equation to find $\boldsymbol{P_o}$, classification ***Accuracy*** and ***F1*** score (average of precision and recall rate) are descried and formulated in [4].

4. Real-World Dataset

A total of 1,971,015 real-world dataset records from 756 volunteers lived in USA, Belgium and Kuwait, who have used 2,138 context-aware services, have been collected over 10-months period from 01-05-2019 to 28-02-2020 to be used in evaluating the efficiency of the learning system presented in [4]. Kuwait announced the first confirmed case of COVID-19 in February 24, 2020 [13]. After that, the dataset records during the pandemic time was provided through an information visualization mobile application installed by a total of 33 volunteers from Kuwait who participated in earlier 2019-dataset studies [12]. A total of 4,326 records were collected and more than 172 context-aware services were used over a 6-months period, from 24-03-2020 to 30-09-2020. In general, we have excluded the dataset collected from the users consisting gray period (i.e. no date records from using the services for one week). For the convenience of PPM modeling, we are considering the variables descried in Table 1, so-called subset ***P***. Other variables are not considered in this module and are out of this paper scope.

In this work, three experiment runs were conducted using three version of set ***P*** to evaluate the accuracy of the algorithm and to assess the impact of COVID-19. The first version, set $\boldsymbol{P_1}$, includes all the dataset records collected before the pandemic, i.e. dataset collected from 01-05-2019 to 28-02-2020 to assess the accuracy of the methodology prior the pandemic. The second version, set $\boldsymbol{P_2}$, is used to evaluate the accuracy of the methodology for the pandemic dataset only. Thus, the set includes all the dataset collected from 24-03-2020 to 30-09-2020. The last set version, set

P_3, is used to explore the impact of the pandemic on the efficiency of the hybrid methodology considering normal and abnormal user contextual privacy preferences behavior. It equal to set P and consists all the dataset records collected before and after the pandemic.

Table 1. Variable Description of subset *P*.

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 and 0=Actual
User_ID	Number	User identification number
User_Pref	Binary	1=Sharable; 0=Not Sharable and Null=To Be Determined
User_Gen	Binary	1=Male and 0=Female
User_Age	Integer	User age
User_Cont	Text	User country
Activity	Integer	User behavior activity; 1=Allow, 2=Deny, 3=Approximate and 4=Change

All the sets (P_1, P_2 and P_3) are used in SM to statistically predict the most likely user privacy preferences setting and configuration performed by the user for the context-data based on the context-data type (*Actual or Approximated*). For LR learning algorithm, the sets (P_1, P_2 and P_3) are divided into four subsets: training, validation, prediction and optimization, see Fig. 3.

Fig. 3. Dataset Partition based on Percentages.

For set P_1, the user variables records collected over 273 days are used for the training and validation processes of the LR learning algorithm. Records collected from 29-01-2020 to 28-02-2020 (25 days) are used for the prediction and optimization processes of the algorithm. Set P_2, on the other hand, uses records collected over 112 days and 56 days for the training and validation processes of the LR learning algorithm respectively. Due to the absences of historical datasets during the pandemic, only records collected over 9 days could be used for the prediction and optimization processes. However, for the superset set P_3, RL learning algorithm is trained and validated using records collected over 443 days. The prediction and optimization process use records from set P_3 collected over 21 days.

5. Results

Before we review the results, Fig. 4 demonstrates the frequency distribution of the preference setting for each set. As shown, during the pandemic (P_2) users set 12% more *Sharable* preferences comparing to their setting prior the pandemic (P_1). *To be Determined* preference found to be ranging between 41% to 43% of the user setting. However, the *Not Sharable* preference distribution drop during the pandemic reflects the reality that most of the user are willing to relax their context privacy preferences in trade off using the optimum solution to endure the predicament. Finally,

the huge different between the sets size justifies the insignificant impact of the noticeable changes during the pandemic (P_2) on the overall dataset (P_3).

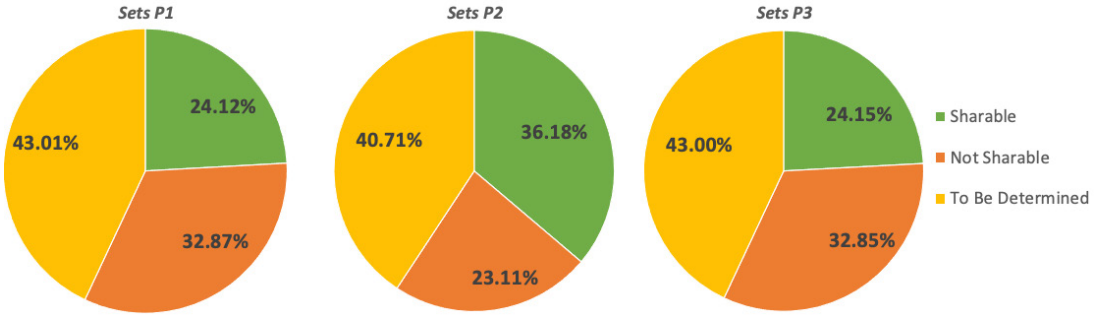

Fig. 4. Frequency Distribution of Preferences per Set.

We have conducted two experiment runs considering three datasets (P_1, P_2 and P_3). The first run is performed to assess the performance of the LR algorithm only compared to the hybrid algorithm adopted for PPM module. The second run is conducted to assess the impact of COVID-19 pandemic on the efficiency of the Hybrid algorithm. Accuracy and F1 score [4] are used to evaluate the prediction efficiency of the algorithms.

Fig. 5 illustrates the efficiency comparison result of LR and the Hybrid algorithms using huge historical set P_1 (i.e. prior the pandemic) and small-scall historical set P_2 (i.e. during the pandemic).

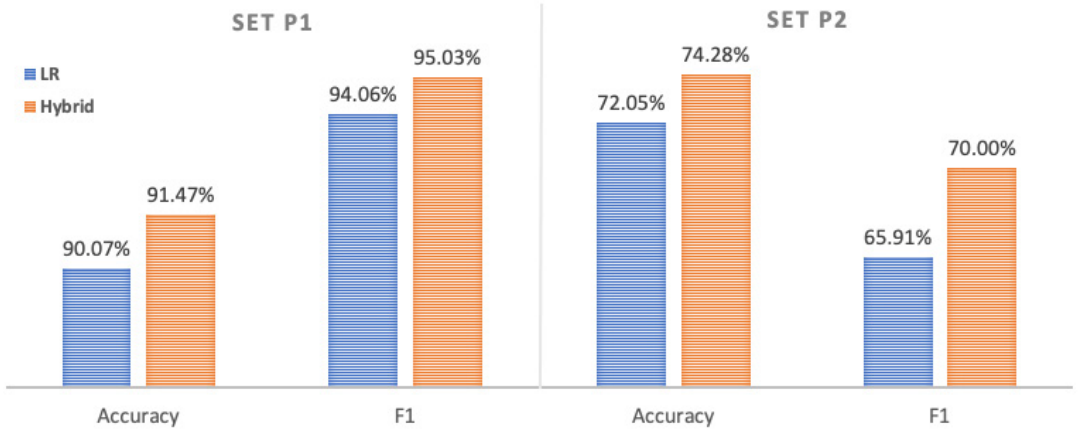

Fig. 5. Performance Scores for LR and Hybrid Algorithms Prior (P_1) and During (P_2) the Pandemic.

As illustrated, both algorithms proven to be efficient when considering normal user privacy preferences behavior prior the pandemic with Accuracy and F1 score above 90%. However, the hybrid algorithm preform prediction with accuracy up to 91.4% and enhanced the LR algorithm robustness by 2.23% prior the pandemic. On the other hand, the algorithms fairly predicted the user privacy preferences behavior during COVID-19 pandemic with at most 74.28% and 70% for Accuracy and F1 score correspondingly. This is justified by the fact that the performance of the learning algorithm relies heavily on the size of the training subset available, which is small for set P_2. But that is not necessarily true as shown in Fig. 6, which demonstrates the performance of the algorithms considering normal and abnormal user contextual privacy preferences behavior set P_3.

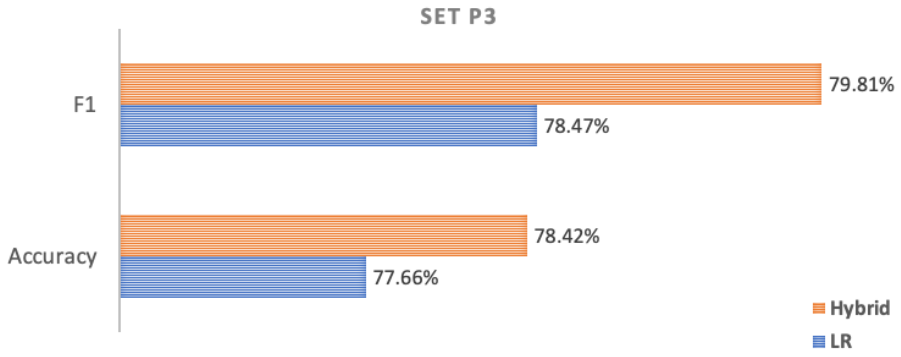

Fig. 6. Performance Scores for LR and Hybrid Algorithms for set P_3.

Based on the figure, both algorithms predicted the user privacy preferences behavior with accuracy less than 80%. However, it is observed that the integrating SM technique slightly enhance the LR algorithm robustness by 1.5%. Overall, the results from the first experiment run show that the hybrid algorithm is more efficient and accurate compared to LR algorithm for implementing privacy preference manager module in the context-aware environment.

The results from the second experiment run is aimed to assess the impact of COVID-19 on the performance scores comparison for the Hybrid learning algorithm of PPM module considering dataset before (P_1) and after including the pandemic dataset (P_3) is demonstrated in Fig. 7.

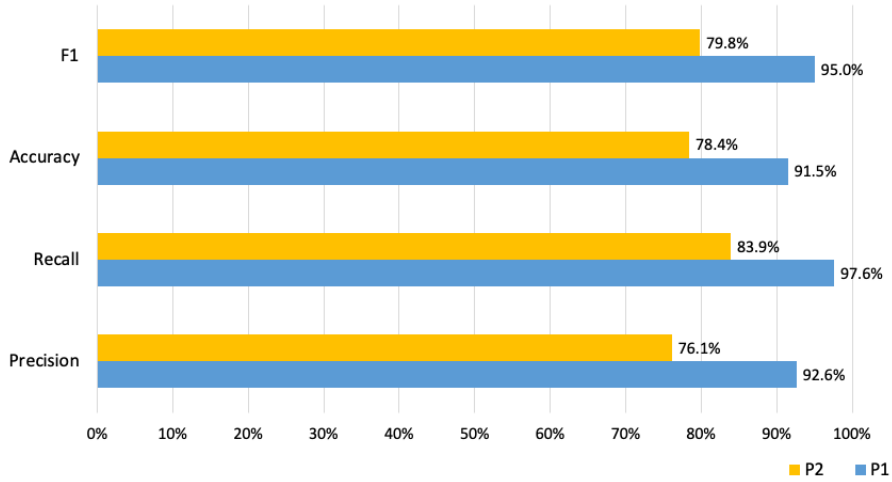

Fig. 7. Performance Scores for Hybrid Algorithms Before (P1) and After (P3) the Pandemic.

From Fig. 7, we can notice the significant impact of the pandemic dataset on the performance and prediction scores of the hybrid algorithm. The algorithm score 92.6% in precision, 97.6% in recall and 95.0% in F1 prior the pandemic with accuracy reaches 92%. Nevertheless, a significant deterioration in performance compared to the dataset including the pandemic data (P_3) for the same algorithm. The final performance scores are 16.5% and 13.7% worse than before for precision and recall correspondingly. In general, hybrid algorithm preform prediction with accuracy less by 13.1% and less robustness by 15.2% with COVID-19 pandemic dataset compared to before the pandemic.

From the above results, we found that the users have rapidly altered their privacy preferences during the COVID-19 pandemic. Moreover, although the learning algorithms are trained to predict different behavior and respond to changes, yet they don't perform well when the behavior differs too much from what they were trained on. We think that this deviations from user normal behaviors was not apprehended by the learning algorithms even with the

availability of large historical dataset. As a conclusion, COVID 19 pandemic reduces the efficiency of user behavior learning algorithms by disturbing the prediction accuracy.

6. Conclusion and Future Work

Part 2 of our proposed privacy protection system in context-aware environments, which includes deploying hybrid methodology based on a statistical technique (SM) and Logistic Regression (LR) learning algorithm for Privacy Preferences Manager (PPM) module is presented. The efficiency of the proposed hybrid methodology prior and during the COVID-19 pandemic has been assessed through two real-world datasets provided by institutes from Kuwait, USA and Belgium. Results show that dramatic changes in user behavior during the pandemic over a short period of time do not allow learning algorithms to evolve properly and negatively impacted the prediction accuracy by 13.1%.

Our next step involves exploring other machine learning algorithms to enhance the prediction accuracy along with assembling the three modules to integrate the automated privacy protection system and assess its effectiveness by deploying it in a real context-aware environment.

Acknowledgements

The authors of this work would like to thank the volunteers from USA, Kuwait and Belgium that collaborated to the collection of data used in our experiments. They extend their gratitude for the users who participated in installing the proposed visualization mobile application and sharing their contextual data during COVID-19 pandemic (2020).

References

[1] Furini M., Mirri S., Montangero M. and Prandi C. (2019) “Privacy Perception and User Behavior in the Mobile Ecosystem,” In: P*roceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, New York, NY, USA, Sep. 2019, 177–182, doi: 10.1145/3342428.3342690.

[2] Shklovski I., Mainwaring S., Skúladóttir H. and Borgthorsson H. (2014) “Leakiness and Creepiness in App Space: Perceptions of Privacy and Mobile App Use,” In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2014, 2347–2356, doi: 10.1145/2556288.2557421.

[3] Alawadhi, R. and Hussain, T. (2019) “A Method Toward Privacy Protection in Context-Aware Environment.” *Procedia Computer Science* **151**: 659-666.

[4] Hussain, T. and Alawadhi, R. (2020) “A Privacy Protection System in Context-aware Environment: The Privacy Controller Module.” In *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services, 39–46*

[5] B. Liu B., Andersen M., Schaub F., Almuhimedi H., Zhang S., Norman Sadeh N., et al. (2016) “Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions,” In: *Proceedings of the Twelfth USENIX Conference on Usable Privacy and Security*, USA, Jun. 2016, 27–41.

[6] Toch E. (2014) “Crowdsourcing Privacy Preferences in Context-aware Applications,” *Personal and Ubiquitous Computing* 18 (1): 129–141. DOI:https://doi.org/10.1007/s00779-012-0632-0

[7] Asada M., Yoshikawa M. and Y. Cao Y. (2019) “When and Where Do You Want to Hide?” – Recommendation of Location Privacy Preferences with Local Differential Privacy. In: Foley S. (eds) Data and Applications Security and Privacy XXXIII. DBSec 2019. Lecture Notes in Computer Science, vol 11559. Springer, Cham. https://doi.org/10.1007/978-3-030-22479-0_9

[8] Abdallah Z., Gaber M., Srinivasan B. and Shonali Krishnaswamy. (2018). “Activity Recognition with Evolving Data Streams: A Review,” ACM Computing Survey 51 (4): Article 71, 36 pages.DOI:https://doi.org/10.1145/3158645.

[9] Bai L., Yeung C., Efstratiou C. and Chikomo M. (2019). “Motion2Vector: Unsupervised Learning in Human Activity Recognition using Wrist-sensing Data,” In: *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers.* ACM, 537–542.

[10] Carker I., Kayes A. and Watters P. (2019). “Effectiveness Analysis of Machine Learning Classification Models for Predicting Personalized Context-aware Smartphone Usage,” *Journal of Big Data* 6: 57-85. https://doi.org/10.1186/s40537-019-0219-y.

[11] Prakash A., Sharma P., Sinha I. and Singh U. (2020), "Spread & Peak Prediction of Covid-19 using ANN and Regression (Workshop Paper)," In: *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, New Delhi, India, 2020, 356-365, doi: 10.1109/BigMM50055.2020.00062.

[12] Hussain, T.and Ismail, A. (2020) “The COVID-19 Impacts on Contextual Privacy Behavior in Kuwait.” In *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia, 187–191*

[13] Central Agency for Information Technology – CAIT, “COVID-19 Updates”. https://corona.e.gov.kw/En/Home/CasesByDate (accessed October 5, 2020).