International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 1, February 2017, pp. 505~512
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i1.pp505-512

Optimized Active Learning for User's Behavior Modelling based on Non-Intrusive Smartphone

Ika Kusumaning Putri[1], Deron Liang[2], Sholeh Hadi Pramono[3], Rahmadwati[4]
[1]Department of Computer Science and Information Engineering, National Central University, Taiwan
[1,2,3,4]Department of Electrical Engineering, University of Brawijaya, Malang, Indonesia

Article Info

Article history:

Received Sep 9, 2016
Revised Nov 15, 2016
Accepted Nov 30, 2016

Keyword:

Active learning
Non-intrusive authentication
Support vector machine
User authentication

ABSTRACT

In order to protect the data in the smartphone, there is some protection mechanism that has been used. The current authentication uses PIN, password, and biometric-based method. These authentication methods are not sufficient due to convenience and security issue. Non-Intrusive authentication is more comfortable because it just collects user's behavior to authenticate the user to the smartphone. Several non-intrusive authentication mechanisms were proposed but they do not care about the training sample that has a long data collection time. This paper propose a method to collect data more efficient using Optimized Active Learning. The Support Vector Machine (SVM) used to identify the effect of some small amount of training data. This proposed system has two main functionalities, to reduce the training data using optimized stop rule and maintain the Error Rate using modified model analysis to determine the training data that fit for each user.Finally, after we done the experiment, we conclude that our proposed system is better than Threshold-based Active Learning. The time required to collect the data can reduced to 41% from 17 to 10 minutes with the same Error Rate.

Corresponding Author:

Ika Kusumaning Putri,
Department of Electrical Engineering,
University of Brawijaya,
Jl. Veteran, Malang 65145, East Java, Indonesia.
Email: ikaemanguka@gmail.com

1. INTRODUCTION

In recent years, smartphone are used more frequently for various applications other than telecommunication. The business activity like online money transaction and any other things makes us store our personal data on the smartphone. That phenomenon makes us more concerned about security issues in our smartphones. We need a protection for accessing our personal sensitive data on our smartphones.

The current protection mechanisms of mobile devices are usually based on PIN codes or passwords, and biometric based methods. Fingerprint is one of the biometric systems that have a strong features with high precision. It can be applied on desktop authentication [1] and mobile authentication [2] by using the fingerprint devices and mounted fingerprint in mobile device. Fingerprints and password entry require explicit action from the user, which is not convenient in a frequent use. The password-based authentication methods used on mobile devices have a surfing issue, people can use social engineering to get the password. Some crackers also can use the brute-force method to get our PIN code. A biometric method like face recognition also have some issues when people input our face only through our photo.

According to recent surveys [3-5], 60% to 80% of users choose to turn these verification features off simply because of its inconvenience. In order to improve the convenience and security of the mobile devices, non-intrusive authentication mechanisms are desirable [6]. Not only for mobile, non-intrusive authentication also used for desktop security. Some desktop protection research [7], [8] exploits mouse dynamics that is

effective for continuous verification because it does not need special hardware to collect biometric data. Non-intrusive authentication does not need an interface because it collects user's behavior in the background then continue authenticating to protect data on the smartphone.There are several method Non-Intrusive authentication that has been proposed, such as using the orientation sensor or a touch screen and using Batch Learning [9-11] and Active Learning [12] method to collect the training data that construct the Non-intrusive authentication mechanism. They use smartphones as sensors to identify user behavior while operating the smartphone. Their training data collected from the way of user use their phone. These research took a flick touch data from the user as the training data. The data is retrieved using a display list then analyzed as user behavior.

The Batch Learning and Threshold-based Active Learning provide non-intrusive authentication to the smartphone, although each of them has its weakness.Batch processing is a model of data processing, with collecting data first, and organize clustering data into groups called batches [13], [14]. The Batch Learning Authentication needs about 240 training data for each user to build the model that will be used as the authentication to achieve high recognition rate. However, to get a lot of data takes a long time data collection and it is not efficient for the user. The Active Learning select the data from the learning process to improve performance with less training data [12]. Threshold-based Active Learning uses a different way of collecting data, the amounts of training data for each user, variety of positions and type of flick is different according to each user's habit. Active Learning will collect training data in some stages if the validation accuracy in the previous stage does not achieve a good result yet. The data collection of Active Learning depends on accuracy in the validation process. Threshold-based Active Learning data collection time faster than Batch Learning because it just takes the important data according to user habits. However, the error rate is greater than Batch Learning which means it has lower recognition rate. Both existing types of research have each other's weaknesses, Batch Learning has a long data collection time and Threshold-based Active Learning has a bad recognition rate.

The value of Equal Error Rate (EER) and collection data time of the previous research shown in Figure 1. Batch Learning have best EER than the other points but the collection time about 20 minutes is not efficient to user while collecting the data. The rectangle and triangle point is theThreshold-based Active Learning that using 90% and 95% as the threshold to stop the collecting data.The Threshold-based Active Learning (TBAL) have faster-collecting data than the Batch Learning. It is possible to collect data from a user in the real implementation.

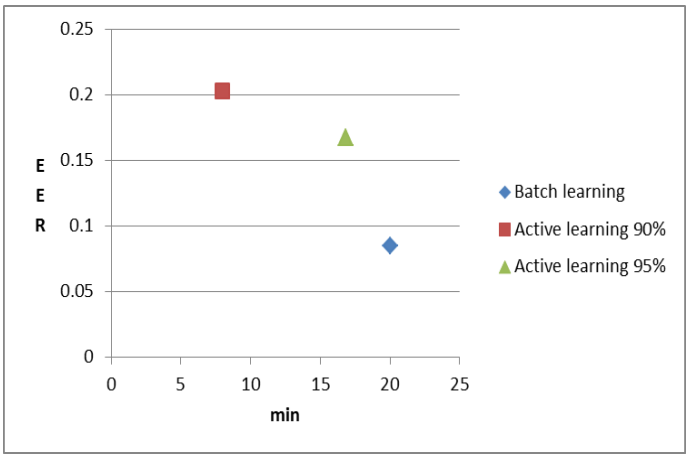

Figure 1. EER and Data Collecting Time of Batch Learning and Threshold Based Active Learning

Figure 2 show that TBAL not effective to reduce the training data. Yellow line in that figure is 11 stages that means it collect the same training data number with the Batch Learning. There are 19 of 45 users that cannot effectively reach the stop rule. It still has more training data than the batch learning.The main purpose of this research are increasing the recognition rate and reduce the amount of training data so we just need less time to collect data. The expected error rate should be less than or equal and the time taken expected to be less than or equal to TBAL95%. To demonstrate the feasibility of the proposed approach, an app has been implemented on the Android 4.3 operating system to collect the biometrics from the touch sensor of 45 users when they operate the smartphones in their hands.

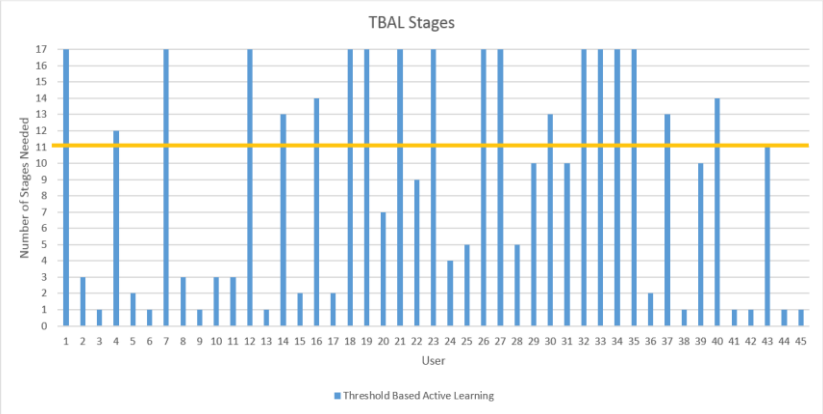

Figure 2. The Effectiveness of Threshold-based Active Learning

2. RESEARCH METHOD

Figure 3 shows the design scenario of this proposed research. In this section, the data collection stages, stopping criteria to collect user's training data and model analysis are presented as follows. This system collects 60 training data in the first stage then do the training using SVM. If the SVM validation accuracy met the criteria to stop, then the system will stop collecting data and start creating Model using SVM. But, if the SVM validation accuracy didn't meet the stopping criteria, the system will analyze the training data in that stage to know which behavior that fit user. 18 training data will be collected based on behavior ratio that already analyzed and start training again using all stage training data.

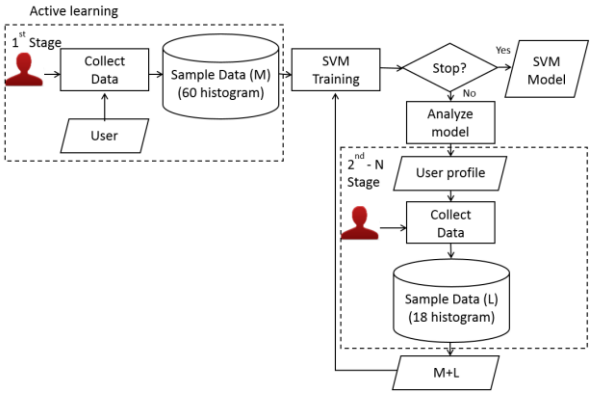

Figure 3. Research Design

2.1. Data Collection Stages

The user's operational behavior can be divided into six categories according to their posture and flick type as shown in Figure 4. Every user has a different ratio for each operational behavior based on their habits. The TBAL use different amount of data required between the operational behaviors. We decided sitting and standing as the posture type because these postures common used by everyone when operating the smartphone than lying, walking or running posture. We are trying to build as much as possible for behavior operation to get the specific user's behavior variation. Therefore we decided to use short, medium and long flick type for user's behavior operation instead of using one flick type or just long and short as a flick type.

In the first phase, the system took the same amount of data from the 6 types of behavior, it used to detect user's behavior while operating the smartphone. The second phase will take the data according to user behavior at the previous stage. Thedata collection amount for each different behavior depends on the ratio obtained from the analysis of behavior SVM model at the previous stage. The system uses each user profile to indicate which operation will be carried out to create a model. As shown in Figure 5, the user asked to stand up and scroll the screen to the numbers 15 and click on it. The asked numbers represent the type of

flick required. In the user profile, we must determine two things, first how much data should be operated by the user and the second posture and number indicating flick type of the user's operational behavior.

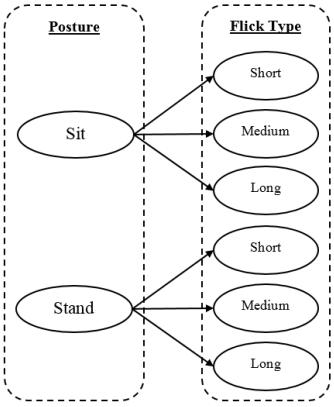

Figure 4. Behavior Operation

Figure 5. Display Application

2.2. Stopping Criteria

Some people have a different rate to reach their stability while using the smartphone. They can stable from the first session or they will stable after some stages that indicated by the SVM validation accuracy. If they reach the highest validation accuracy, the system does not need to collect other data anymore. But, if the validation accuracy is still improving, we need to look another stage until reaches the highest accuracy. Most users have the same pattern on the SVM validation accuracy. If their validation accuracy high and will not rise again significantly, the model of the highest accuracy that has been formed is the best model, so we do not need to add more data and it will reduce the collecting data time. Figure 6 shows some user that have SVM validation accuracy more than or equal to 95% will stable and not decrease or increase significantly. In this stable condition, we can make the data collection stop earlier because the early model already has a good model and do not need to add more data.

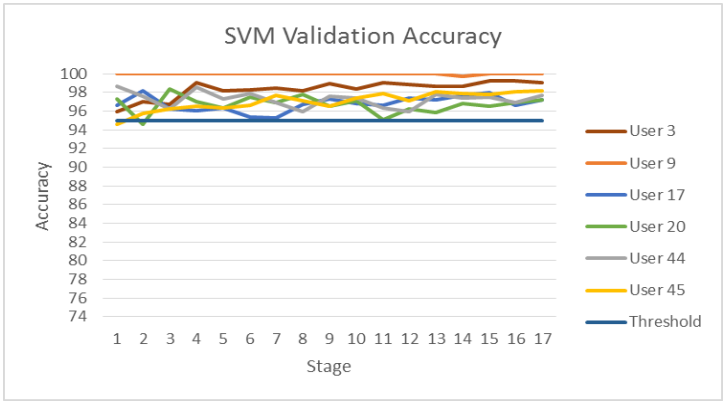

Figure 6. Stable SVM Validation Accuracy

We set the stopping criteria for collecting data to reduce the number of training data. If the stage reaches the criteria, the collecting data will be stopped. In this research, we set the stopping criteria until the SVM validation accuracy of the previous stage does not improve anymore or stop until SVM validation accuracy reach more than or equal to 95% and set 11 stages as a maximum number of the stage.We choose 11 stages as the maximum number of stages according to the number of training data in the batch learning because batch learning gives the best EER with the same number of data in the 11^{th} stage. Table 1 shows the stop rule pseudocode before collecting the next data using 3 window stages.

Table 1. Pseudocode of Data Collection Stop Rule

```
if stage n-2 ≥ stage n-1
        then best SVM Accuracy=stage n-2
else
        then best SVM Accuracy =stage n-1
end if
threshold= best SVM Accuracy * 1%
if (stage n- best SVM Accuracy) > threshold AND SVM Accuracy
< 95
        then go to next stage
else
        then stop
end if
```

2.3. Model Analysis

The TBAL uses the model analysis to determine the next stage training data. It uses SVM classifier in the previous stage to determine what behavior is included in SVM support vector. Then the ratio of each behavior on SVM support vector is calculated to obtain the amount of data on individual behavior as additional data in the stage.In this research, we use the SVM classifier using soft margin to separate the non-linear separable classes. To determine which behavior that will be used for next stage, we analyze the closest point to the hyperplane because SVM support vector is not the most sensitive members of the class since there were still any members that closer to the hyperplane. Since we use the closest points to find the fittest user's behavior, we just find 18 closest points to the hyperplane. Then we analyze all point to get next stage's behavior training data.

Although the optimal hyperplane constructed by support vector, there are some points that closer to the hyperplane than the support vector. Because of their position closer to the hyperplane, they are more sensitive than other and potentially misclassified. Figure 7 shows the support vector and the points between the margin and hyperplane in the SVM soft margin that have more sensitive effect to the classification. Therefore, we decided to choose them to get more accurate classification.

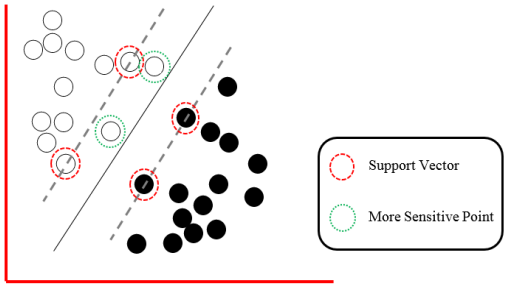

Figure 7. Support Vector and More Sensitive Point in SVM Soft Margin

3. RESULTS AND DISCUSSION

The new model analysis applied to all users and we did not limit the number of stages to know the impact of the new model analysis to all stages. We use 17 stages that same with the maximum number of stages in Threshold-based Active Learning to analyze the difference of SVM validation accuracy when using the different model analysis. Figure 8 shows that the average of SVM validation accuracy increasing to almost all users when using the new model analysis which uses the closest point to determine the ratio of each behavior operation.

When we apply the threshold based stop rule to this new model analysis, the collecting data will be stopped earlier than before because the value of SVM validation accuracy has improved. As shown in Figure 9, about 20 of 45 users can stop earlier than using the support vector model analysis. The average number of the stage that needed for all user when using support vector model analysis is about 9.5 stages and it decreases to 6.5 stages while using the closest point model analysis. The average of collecting data time can reduce about 23% from 17 minutes to 13 minutes. The results reveal that reducing the time of the model analysis based on closest points to the hyperplane is significant.

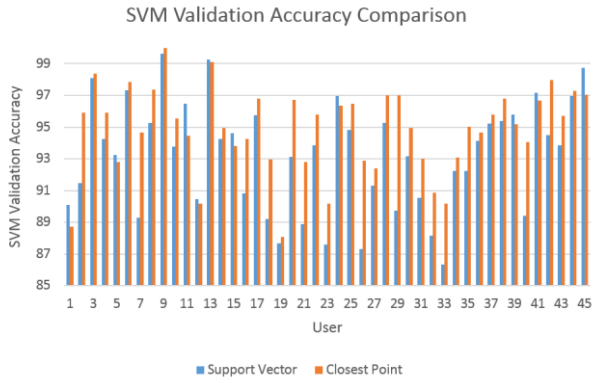

Figure 8. SVM Validation Accuracy Comparison

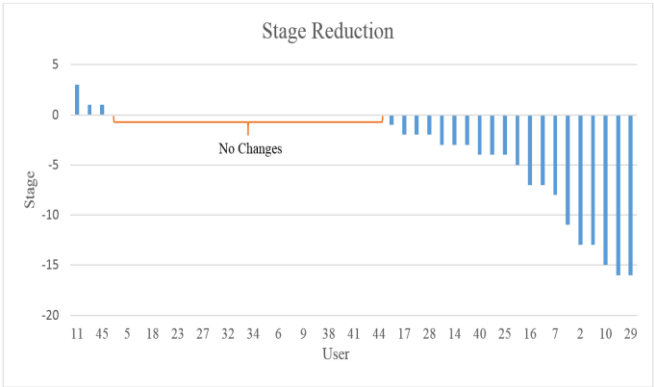

Figure 9. Stages Reduction

After knowing the model analysis effectiveness, we applied the optimized stop rule to the system to limit the number of stages.To determine whether the results of the experiment is successful, we did a comparison of the time data collection between the experimental results with TBALwhen it reaches the same EER. From this comparison, as shown in Figure 10, it is known that there is a reduced data collection time from Threshold Based Active Learning to Optimized Active Learning as much as 41%. Applying new model analysis and optimized stop rule to the system gives faster authentication time than TBAL. The results show that this approach is more feasible for authentication purpose than the TBAL because faster authentication time gives more comfortable to the users.

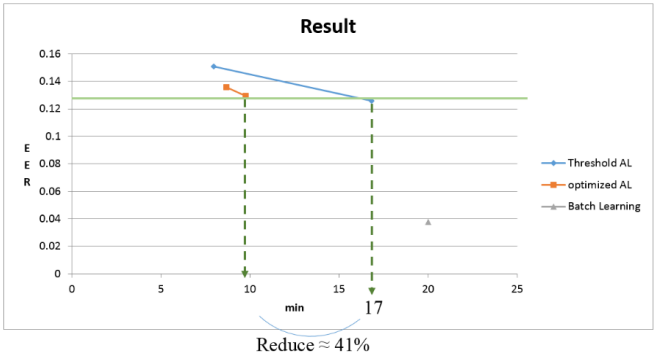

Figure 10. EER and Data Collection Time Comparison

4. CONCLUSION

In this research, we improve the concept of Active Learning for collecting the training data with analyzing each user behavior when they operate the smartphone to shorten the data training and collecting data time until it could be acceptable to the user.In order to achieve this goal, we develop the Active Learning by changing the several phases including the model analysis, stopping criteria and combining it with the Threshold-based Active Learning's stop rule. In the first phase, we changing the stopping criteria of the collecting data with checking if the accuracy will not improve significantly is the better stop rule to get the optimal result.The second phase, we change the model analysis that used to determine the ratio of each behavior when selecting the training data. We choose the closest point to the SVM hyperplane to calculate the behavior ratio.This study shows that 43 of 45 person's training data can reduce using the new Active Learning method. The Optimized Active Learning has less collecting data time than Threshold-based Learning and Batch Learning. We can reduce the collecting data time until 41% from the Threshold-based Active Learning.

ACKNOWLEDGEMENTS

The authors would like to thank Sally Pan and all members of Intelligent Software Systems Laboratory of National Central University Taiwan for insightful discussions during this research and also reviewers for their valuable comments.

REFERENCES

[1] R. Chouhan, A. Mishra and P. Khanna, "Fingerprint Authentication by Wavelet-based Digital Watermarking", *International Journal of Electrical and Computer Engineering*, Vol. 2, No. 4, pp. 523~528, 2012.
[2] J. Han, "Fingerprint Authentication Schemesfor Mobile Devices", *International Journal of Electrical and Computer Engineering*, Vol. 5, No. 3, pp. 579~585, 2015.
[3] O. Mazhelis, J. Markuula, and J. Veijalainen, "An integrated identity verification system for mobile terminals", *Information Management & Computer Security*, vol. 13, no. 5, pp. 367-378, 2005.
[4] Smart Credit. "Consumer Reports survey on mobile phones and security", 2011 Press. available from: http://www.smartcredit.com/blog/2011/09/02/consumer-reports-surveyon-mobile-phones-and-security/ (2011/11/15).
[5] C. Theriault, "Survey says 70% don't password-protect mobiles", 2011 Press. available from: http://nakedsecurity.sophos.com/2011/08/09/freesophos-mobile-security-toolkit/ (2011/11/11).
[6] N. Clarke, S. Karatzouni, and S. Furnell, "Flexible and transparent user authentication for mobile devices", *IFIP Advances in Information and Communication Technology*, 297/2009, 1-12, 2009.
[7] C.C. Lin, C.C. Chang, and D. Liang, "A New Non-intrusive Authentication Approach for Data Protection Based on Mouse Dynamics", *International Symposium on Biometrics and Security Technologies*, Taipei, Taiwan, March 26-19, pp. 9-14, 2012.
[8] A.A.E. Ahmed, and I. Traore, "A New Biometric Technology Based on Mouse Dynamics", *IEEE Trans. on Dependable and Secure Computing*, vol. 4, no. 3, pp. 165-179, 2007.
[9] H. Gamboa, and A. Fred, "A User Authentication Technic Using a Web Interaction Monitoring System", *Lecture Notes in Computer Science* (Pattern Recognition and Image Analysis), vol. 2652, pp. 246-254, 2003.
[10] C. C. Lin, C. C. Chang, D. R. Liang, and C. H. Yang, "A Preliminary Study on Non-Intrusive User Authentication Method Using Smartphone Sensors", *Applied Mechanics and Materials*, vol. 284, pp. 3270-3274, 2013.
[11] C.C. Lin, C.C. Chang, and D. Liang, "A Novel Non-intrusive User Authentication Method Based on Touchscreen of Smartphones", to appear in *Journal of Internet Technology*, vol. 16, p. 1-10, 2015.
[12] E. Chen. "Using Active Learning to Collect User's Behavior for Training Model. Base on Non-intrusive Smartphone Authentication", Master Thesis, National Central University, 2015.
[13] J. Liu and J. Hu, "Dynamic batch processing in workflows: Model andimplementation", Future Generation Computer Systems, vol. 23, no. 3, pp. 338–347, 2007.
[14] L. Pufahl and M. Weske, "Batch Activities in Process Modeling andExecution", in Service-Oriented Computing. Springer, 2013, pp. 283–297.

BIOGRAPHIES OF AUTHORS

Ika Kusumaning Putri received the MSc degree in Department of Computer Science and Information Engineering, National Central University, Taiwan in 2016 as an International Dual Degree Master student between University of Brawijaya, Indonesia and National Central University, Taiwan. She completed her Bachelor degree in Department of Informatics Engineering, University of Brawijaya, Indonesia in 2014. Her research interest area in the areas of software engineering and information technology, mobile development, and intelligence systems.

Deron Liang is a Professor in the Department of Computer Science and Information Engineering, National Central University, Taiwan. He received a BS degree in electrical engineering from National Taiwan University in 1983, an MS and a PhD in computer science from the University of Maryland at College Park, USA in 1991 and 1992 respectively. He alsoholds joint appointment with the Institute of Information Science (IIS), Academia Sinica, Taipei, Taiwan, Republic of China. He was with IIS from 1993 till 2001. Dr. Liang's current research interests are in the areas of software fault-tolerance, system security, distributed systems, objectoriented and system reliability analysis. Dr. Liang is a member of ACM and IEEE.

Sholeh Hadi Pramono is a senior lecturer in Department of Electrical Engineering, University of Brawijaya, Indonesia. He got Doctor degree from University of Indonesia, Indonesia. Hepresently work in Telecommunication Laboratory, University of Brawijaya, Indonesia as anoptical telecommunication specialist. His research interest area in the areas of optical telecommunication, technology of antenna, distributed systems, and telecommunication architecture.

Rahmadwati is a lecturer in Department of Electrical Engineering University of Brawijaya, Indonesia. She got PhD degree from University of Wollongong, Australia. Her research interest area in the areas of Image Processing and Intelligence Systems.