CSE 881 FINAL PROJECT

Project Title: User Authentication with Activity Patterns

Project Type : Empirical Study

Difficulty Level : Moderate

Justification for your rating:
Data Collection +1
Data Preprocessing +1
Evaluation +.5

Summary of Team Member Participation:

Fill out the following table for each team member with a rating from 1 to 3 (1: poor, 2: satisfactory, 3: good). For "responsive to emails" and "attendance at project meetings", the rating must be provided by averaging the rating provided by other members of the group.

Name	Responsive to emails*	Attended project meetings*	Participate in data collection /preprocessing	Participate in coding	Participate in analysis/ experiments	Writing final report	Class presentation	Completed Assigned Tasks
Achsah Junia Ledala	3	3	3	3	3	3	3	3
Luke Stanton	3	3	3	3	3	3	3	3
AKM Tauhidul Islam	3	3	3	3	3	3	3	3

Team Member Roles and Contributions:

Name	Roles and Contributions
Achsah Junia Ledala	Help with project idea, Perform data preprocessing, Perform data analysis and evaluation, Help with all reports and presentation.
Luke Stanton	Help with data collection, Perform data preprocessing, Perform data analysis, help writing all reports and presentation, Help to come up with project idea
AKM Tauhidul Islam	Helped with the project idea, performed data preprocessing, data analysis and evaluation, and participated preparing the reports and the presentation.

I approve the content of the final report (please add your signature below):

Achsah Junia Ledala: Achsah Ledala 12-06-2018
Luke Stanton: Luke Stanton 12-06-2018
AKM Tauhidul Islam: AKM Tauhidul Islam 12-06-2018

User Authentication with Activity Patterns

Achsah Junia Ledala
Department of Computer Science
Michigan State University
East Lansing, MI, 48824, USA
517-402-9909
ledalaac@msu.edu

Luke Stanton
Department of Computer Science
Michigan State University
East Lansing, MI, 48824, USA
stanto85@msu.edu

AKM Tauhidul Islam
Department of Computer Science
Michigan State University
East Lansing, MI, 48824, USA
islama@msu.edu

ABSTRACT

More and more people today frequently use wearable devices that can track their activity patterns. For instance, carrying around a Fitbit creates time series data of how many steps the wearer takes, at what time of day the steps are taken, and more. This type of data has been frequently used for user authentication, assuming they are unique to a user [1-4]. We use similar data to create a model that attempts to authenticate users based on the activity patterns they provide from their fitness tracking device. We used an open source data set that contains the activity patterns of 30 different people [6] over the course of a few months, using a Fitbit. We used the data to train a classification model with support vector machine and adaptive boosting. We also used the data off of the FitBit one of our group members has been wearing throughout the year. With privacy becoming a major concern in the world today, a simple effective way to authenticate users of computer systems becomes a must. As described in the sections below, we had some difficulties at first getting any meaningful results but after using some larger data sets and different techniques we were able to show that using activity patterns as a mean to authenticate users does show some promising results.

Keywords

User Authentication, Activity Patterns, Activity Recognition, Support Vector Machine, Adaptive boosting, Bagging, Random Forests.

1. INTRODUCTION

Computer security and privacy are becoming more and more of a concern today. With the challenges that come with remembering strong passwords and the susceptibility to compromise that comes with weak passwords. The need for a simple and effective way to authenticate users continues to grow.

More and more people today are adopting some form of activity tracking device into their daily lives. It can be in the form of a device specialized to track activity patterns, or it can simply be the smartphone they are carrying around that tracks their activity.

Putting together the data points that these activity tracking devices create could potentially be used as input to be used to train a classifier for a user authentication system.

In this paper we will describe how the data collected from a FitBit could potentially be used in an authentication system. We explore the use of Support Vector Machine, Adaboost, Bagging, Random Forest and Clustering to achieve this goal.

We explore local models specific to the user they are meant to authenticate as well as global models that would be used to authenticate any user, without the need of user specific models.

The problem we are looking to research is the idea of how fitness data could be used to authenticate users of computer systems. Previous work [4,5] are examples of research using data mining techniques such as Support Vector Machine and Adaboost to build classifiers for authentication systems. They mentioned in these papers that they use those methods because they have shown promising results in other applications, and it did show good results in their applications. In our research we are trying to build similar authentication systems with different data collected as input.

The rest of the report is organized as follows. Section 2 presents a number of existing works on user authentication. Section 3 presents further detail on the problem statement and the challenges. The compared techniques are briefly presented in section 4. Experimental results are presented in section 5. And the report concludes in section 6.

2. Related Work

A related piece of research we found is described in [4] as “CABA” where the researches built an authentication system based on health data extracted from the users with an array of medical sensors. The users would be hooked up to a blood pressure monitor, a heart rate monitor, and more that would would take their health information and feed it into an authentication system that would continuously make sure the person hook up to the workstation is indeed the person who is authorized to use the workstation. An issues with this system is just how invasive it is to run. The users need devices on them that are constantly monitoring their blood pressure as well as other health attributes.

As described in [5] the researches built an authentication system that was built around features extracted from users eyes when a visual stimuli is presented to it. The benefit of a system like this is that the means of authentication is effortless for a user. In this case the user simply had to look at a dot on a screen and the system would do the rest. With our system the goal would be to achieve a similar form of effortless authentication on the users side as all they would need to do is pass the data they have on their fitness tracking device to the system to be authenticated.

As the related pieces of work explore using Support Vector Machine and Adaboost to classify the users of the system, those are the methods we first tried with our research.

3. PROBLEM STATEMENT

As mentioned in the introduction, the problem we are interested in researching is whether fitness data can be used as a means to authenticate users. The idea being many people carry around devices that track their daily activity and this data could potentially create a "fitness signature" that could be used to authenticate users.

The type of data we collected is in the form of activity data collected from FitBits. The main features of the data are in the forms of the number of steps the user has taken, the distance they have moved, and the amount of time they spent during different types of activity (lightly active, very active, and fairly active).

The data mining tasks that we are performing are clustering and classification. Specifically the classification methods that we are implementing are Support Vector Machine, AdaBoost, and Random Forest.

4. METHODOLOGY

We have applied a number of machine learning techniques on the activity patterns obtained through the sensors. In this section, these techniques are briefly described along with the reasoning for choosing them.

4.1 Classification

Classification is the problem of identifying to which of the set of labels a new observation belongs, based on the training data of observations whose label is known.

Classification technique can be applied to our problem of authenticating a user by training a model using the activity data of users from the past and looking at new activity data they make in the future.

The various techniques of Classification used in our project are:

Support Vector Machine: It is a supervised learning model used for classification. This model creates a hyperplane that is used to separate different classes.

Bootstrap aggregating: Also called as Bagging, it is a Machine Learning ensemble algorithm which increases the accuracy by reducing the variance and helps to avoid overfitting.

Random Forests: It is an ensemble learning method for Classification that operates by constructing a number of decision trees and sampling them randomly to avoid overfitting. As this algorithm chooses random features at every step of iteration, there is a less probability of the features being correlated and thus the accuracy should be high.

Adaptive Boosting: It converts weak learners to strong learners by assigning higher weights to training data that has been misclassified. Since it focuses on the 'harder to classify' data at every step by assigning higher weights, we believe that it could give us good results in authenticating the users.

We built a local classification model for authenticating each user in the dataset using Support Vector Machines and Adaptive Boosting by assigning a positive label to the correct user and zero labels to incorrect users. We also built a Global Classification model for every user by giving them individual labels using Support Vector Machines, Random Forests, Bagging and Adaptive Boosting.

4.2 Clustering

Unlike the classification approach, the clustering approach focuses more on the underlying similarity of a set of entries in a dataset. A subset of entries form a cluster if they are close to each other based on given constraints. Because of the weak distinctive nature of the activity patterns, such dataset seems like a good candidate for clustering.

Therefore, we have further studied the similarity of the feature vectors via unsupervised learning. The objective was to extract highly similar activity patterns corresponding to a small number of users because the activity patterns are not as unique as the biometric properties. Once a pattern can be mapped to a subset of the population, it will be more effective to authenticate a user based on a small number of patterns.

We have applied hierarchical clustering for the similarity search for two reasons. First, the dendrogram generated by the hierarchical clustering technique would help understand the similarity of the patterns in different distance levels. Second, there is no need to provide the number of clusters manually. Therefore, the true activity patterns will emerge based on similarity.

It is expected that a user will be associated to more than one activity patterns because of different habits on different times of a day and/or different days of a week. Therefore, we need to use longer periods of activity patterns to authenticate a user while using the resulting clusters. The majority of the predicted labels should be used for such technique.

5. EXPERIMENTAL EVALUATION

We have conducted extensive experiments to evaluate effectiveness of the applied techniques for user authentication. The rest of this section presents the experimental setup, properties of the datasets and the experimental results.

5.1 Experimental Setup

The experiments were conducted on an Intel(R) Core(TM) i7 CPU @ 3.30GHz with 8GB physical memory. The classification techniques used for the experiments were Support Vector Machine, Bagging, Random Forest and Adaptive Boosting along with the Hierarchical Clustering. The performance of the classification techniques were evaluated based on prediction accuracy. On the other hand, the clustering technique was evaluated based on purity of the clusters. All the experiments were implemented in python.

5.2 Datasets

Activity Data:

The Dataset had two separate files for the 2 months of Daily Activity for 30 users and it had 14 different attributes that had not been normalized. The characteristics of the data after preprocessing is that the data from the two files have been merged into a single file and normalized using the min-max normalization and the data has also been standardized. So, the final Activity data has 14 attributes for each user and 1397 instances.

Time Series Data:

The data had 1 attribute (Steps) and 46183 instances of Step Activity of 30 users for 60 days where each instance corresponds to hourly step count. We preprocessed the Time Series data of 30 users for 60 days to have 24 attributes corresponding to each hour of the day for each user. So, the final dataset contains 1925 instances each instance pertaining to a user's step activity for a single day.

Fitbit Data:

We had a FitBit that collected data from before and throughout the semester. We combined this data with the data set we found on the Google database search to create a local model for the owner of the FitBit. Our original attempt had about 30 users each with about 2 months of data at best, in turn we would have around 60 days worth of training data for a single user at most for their local model. When we tried this with our own fitbit data we had over 200 days worth of data to train the model on.

5.3 Experimental Results

In this section, experimental results of the classification techniques used are presented followed by those of the clustering based technique.

5.3.1 Classification

We built a Global Classification model using popular techniques like Support Vector Machine, Bagging, Adaptive Boosting and Random Forest. We demonstrated that the user can be authenticated using these methods with the following accuracy for Activity Data:

SVM- 68.5%

Adaptive Boosting- 71.43%

Bagging- 65.71%

Random Forest-71.4%

Figure 1 shows the histogram of various models and their corresponding accuracies.

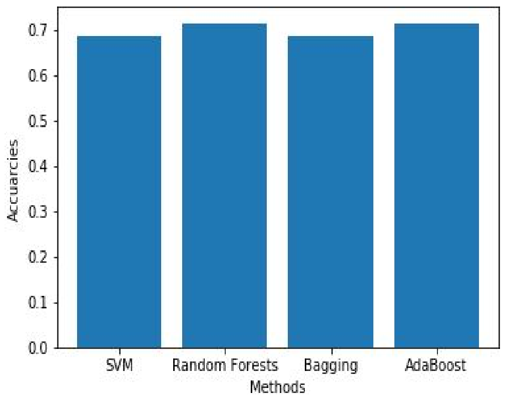


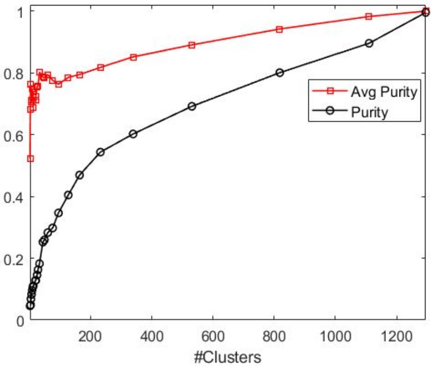
Figure 1: The accuracies of different models for Global Authentication model.

Furthermore, we tried classification on the local models with Support Vector Machine and AdaBoost. At first when we tried to create a model with the daily activity patterns for the user it was not working very well at all, it would almost always falsely reject the authorized user. We believe one of the limitation that made our initial attempts to make a classifier with the data unsuccessful was the limited number of samples from the users. As mentioned in the previous section, the original data only had about 60 days that could be used as training data. When we combined it with the data from our own FitBit, we had training data with over 200 days. When we trained a local model for the owner of the FitBit with the 200 days, in testing it correctly authenticated the true user 100% of the time and incorrectly authenticated an imposter as the genuine user 0% of the time.

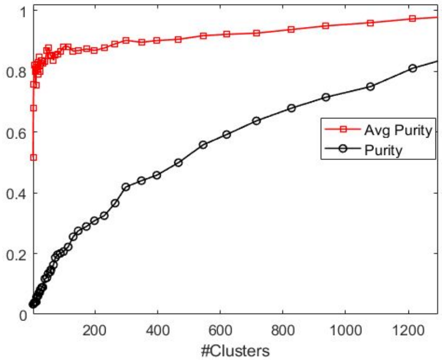
5.3.2 Clustering

We have also performed clustering to evaluate the similarity of the feature vectors. Both time series and daily activity based feature vectors are used for the clustering. The agglomerative hierarchical clustering was chosen to observe the gradual similarity of the feature vectors with increasing distance measure. The linkage criterion was set to average linkage to reduce susceptibility to noise. Purity of the clusters are measured based on the ground truth labels.

Figure 2 shows the change in weighted and average purity of the clusters as the number of clusters changes. The number of clusters is decreased as the intracluster distance is increased. For both time series and daily average activity data, the weighted purity of the clusters drops rapidly as the distance increases. This is because there are few large clusters where the majority of the feature vectors are from different users. In other words, those patterns are common to most users. However, we prefer tightly bound clusters which provide distinct patterns of one/few users. For that purpose, we measured the average purity of the clusters. The Figure shows that the purity of such small clusters are significantly high even for larger distance. Those clusters could be used to authenticate a user based on his/her activity patterns for a period of time. We intend to investigate this in the future.



(a) daily activity data



(b) time series data

Figure 2: Purity of the clusters generated from activity patterns

5.4 Discussion

At first when we built the classifier against the activity data and applied Support Vector Machine to the data to build the local classification model it did not return very encouraging results. Many times it was hard to get the model to recognize a single user. It is worth noting that while from the start it was very unsuccessful at correctly recognizing the correct user, we never ran into issues of the local model incorrectly classifying the wrong users as the right one. That is, from the start we had about a 0% false acceptance rate. After applying the same support vector machine technique with the larger data set we got from our personal FitBit the results got much more encouraging. The model correctly classified the owner of the FitBit for 5 testing samples and correctly classified all the other users as not the owners of the FitBit for 25 samples.

6. CONCLUSIONS

In this project we have investigated the continuous user authentication problem by using user activity patterns collected through activity sensors. We have generated both time series and non time series feature vectors from the collected user activities and applied them to create models for user authentication. The models were generated based on frequently used techniques such as SVM, Adaboost, Bagging and Random Forest. In addition, we created a clustering based techniques to exploit similarity of the activity patterns among users to authenticate them based on their activities within a window. Our experimental results provide valuable insight to develop an effective user authentication technique based on their activity patterns although we couldn't supersede the state-of-the-art user authentication technique based on biomedical signals of a user. We intend to work on this project to make further improvements based on the gained understanding to further improve the authentication accuracy.

7. REFERENCES

[1] Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in Proc. IEEE Symp. Secur. Privacy, 2012, pp. 553–567.

[2] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith, "Smudge attacks on smartphone touch screens," in Proc. 4th USENIX Workshop Offensive Technol., 2010, vol. 10, pp. 1–7.

[3] C. Ma, D. Wang, and S. Zhao, "Security flaws in two improved remote user authentication schemes using smart cards," Int. J. Commun. Syst. , vol. 27, no. 10, pp. 2215–2227, 2014.

[4] Mosenia, Arsalan, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha. "CABA: Continuous authentication based on BioAura." IEEE Transactions on Computers 66, no. 5 (2017): 759-772.

[5] Ivo Sluganovic, Marc Roeschlin, Kasper B. Rasmussen, Ivan Martinovic: "Using Reflexive Eye Movements for Fast Challenge-Response Authentication" Department of Computer Science, University of Oxford.

[6] https://zenodo.org/record/53894#.W99X-9FVK1F

[7] https://thenextweb.com/insider/2016/03/31/5-technologies-will-flip-world-authentication-head/

Appendix

The source code and datasets are available in the following link
https://github.com/cse-course-projects/cse881_FS18_user_authentication