Biometric Authentication using Mouse Dynamics

MACHINE INTELLIGENCE AND EXPERT SYSTEM

Arnab Biswas(Roll-21EC65R01)
Alisha Oraon(Roll-21EC65R04)
Kavitha Annepu(Roll-21EC65R20)
Yashwanth Solasa(Roll-21EC65R08)
Ritwik Das(Roll-21EC65R02)

Introduction

In the modern world, most computers use a password or any other kind of recognition system to identify the user. But if an authorised user is detected, then, the authentication system does not authorise again during the active season. This may create a huge security threat because a non-technical person neighbours know about the security process and very frequently leave their computer unlocked with an active season.

This type of carelessness allows three types of attacks:

1. A user of low clearance can gain access to a terminal with higher clearance and access files or functions of their network.
2. Users with the same or higher clearance can conceal their identity by performing malicious actions under the guise of a co-worker.
3. A person who is affiliated with the company in any way can gain access to the internal network.

To avoid these kinds of problems, authentication techniques based on biometric are introduced. In this project, we have implemented a method for a continuous user authentication system for PCs using mouse dynamics.

Biometric Authentication

Definition:

Human recognition can be done by using physiological or behavioural characteristics. Biometrics offer automated methods of identity verification or identification on the principle of measurable physiological or behavioural characteristics. The characteristics are measurable and unique. Thus, biometrics play an important role in recognizing a human being.

Types of Biometrics:

The biometrics which is used to recognise a person is two types:

1. Physiological biometrics
2. Behavioural biometrics

Physiological biometrics:

Physiological biometrics involves physiological characteristics of a human being used as biometric such as voice, DNA, fingerprint, IRIS pattern or hand geometry. These biometrics are more reliable and accurate. They are not affected by any mental conditions such as stress or illness.

- Fingerprint: Fingerprint recognition refers to the automated method of identifying or confirming the identity of an individual based on the comparison of two fingerprints. Fingerprint recognition is one of the most well-known biometrics, and it is by far the most used biometric solution for authentication on computerized systems. The reasons for fingerprint recognition being so popular are the ease of acquisition, established use and acceptance when compared to other biometrics, and the fact that there are numerous (ten) sources of this biometric on each individual.
- Face: Face recognition is a method of identifying or verifying the identity of an individual using their face. Face recognition systems can be used to identify people in photos, video, or in real-time. Law enforcement may also use mobile devices to identify people during police stops.
- Iris recognition or iris scanning is the process of using visible and near-infrared light to take a high-contrast photograph of a person's iris. It is a

form of biometric technology in the same category as face recognition and fingerprinting.

- Voice Recognition: Voice recognition is commonly used to operate a device, perform commands, or write without having to use a keyboard, mouse, or press any buttons. Today, this is done on a computer with ASR (automatic speech recognition) software programs.

Behavioural biometrics:

Behavioural biometrics involves the behavioural characteristics of a human being. These biometric characteristics are acquired over time by an individual and are at least partly based on acquired behaviour. Thus, it is something known to an individual and can be exploited for authentication purposes.

- Mouse Dynamics: Mouse Dynamics are tiny patterns and variations in mouse- or pointer-based hand and finger movement that occur naturally as users interact with their screen pointer. This behavioural biometric is characterized by the way an individual moves the mouse or clicks on the screen of the desktop/laptop. Mouse actions like mouse movements, clicks, drag and drop etc. can be used as useful features. This behavioural biometric also has issues with the variability of features over time.
- Keylog Dynamics: The behavioural biometric of Keystroke Dynamics uses the manner and rhythm in which an individual types characters on a keyboard or keypad. The keystroke rhythms of a user are measured to develop a unique biometric template of the user's typing pattern for future authentication. Keystrokes are separated into static and dynamic typing, which are used to help distinguish between authorized and unauthorized users.[6] Vibration information may be used to create a pattern for future use in both identification and authentication tasks.

Various Components of Biometric Recognition system:

- Feature extraction captures the data generated by standard input devices such as a mouse or a keyboard.
- Feature extraction and classifier module that constructs the users signature based on his behavioural biometrics.
- A signature database consisting of behavioural signatures of registered users.

Problem Statement:

The objective of the project is to create a continuous user authentication system for PCs/laptops to prevent threats against intruders, using biometrics involving mouse dynamics. The program was implemented to authenticate the user by K- Nearest Neighbor Classifier by the neutral, happy and sad mood data.

Mouse Dynamics

For getting an idea about mouse dynamics we need to look into different types of mouse actions that can occur from user interaction with the PC through a mouse. The main strength of mouse dynamics biometric technology is in its ability to constantly monitor legitimate and illegitimate users based on their session based usage of a computer system.

Different types of mouse actions are

- Mouse Move: Mouse move is a simple movement involving no clicks. Mouse move can be between two click events or non-click events.
- Drag and Drop: It is the action that starts with a mouse button held down followed by a movement and finally the button released. Generally, it is used to move/copy a file to a particular location.
- Point and Click: It is a movement of the mouse ending in a click.
- Silence: This action suggests no mouse movement.

In order to capture these kinds of mouse actions we would require the data collection software to capture events like:

1. Mouse move
2. Mouse pointer location
3. Mouse wheel movement
4. Mouse Pressed
5. Mouse released

Using the five events mentioned in the table we can extract the different mouse actions e.g. left click, left double click, right-click, right double click, drag and drop.

K-NN Algorithm

The K Nearest Neighbor algorithm comes under the Supervised Learning category and is used for Classification (most commonly) and Regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

The algorithm's learning is

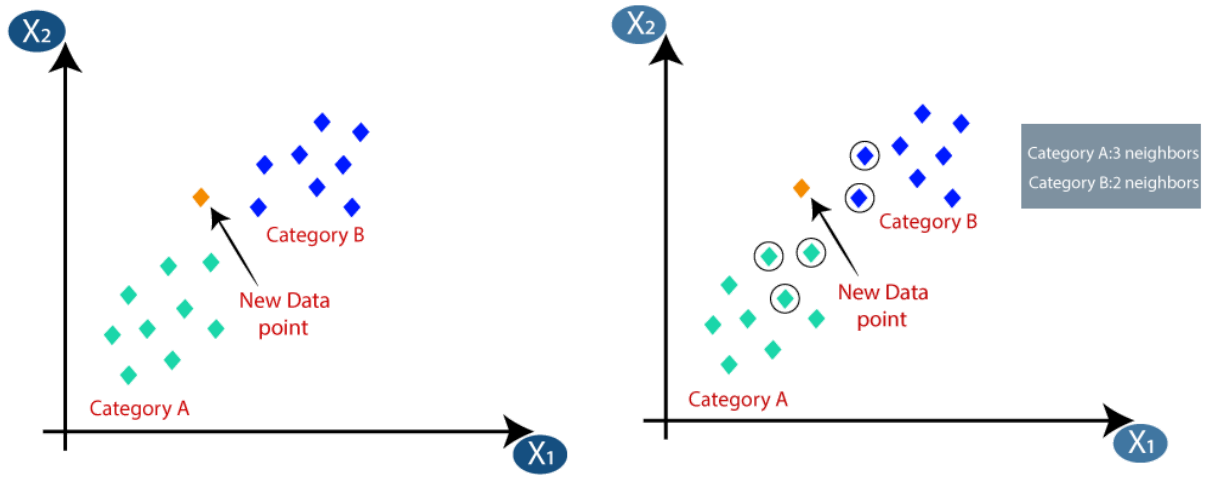
1. Instance-based learning: Here we do not learn weights from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data.

2. Lazy Learning: Model is not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance.

3. Non - Parametric: In KNN, there is no predefined form of the mapping function.

Principle

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



Firstly, we will choose the number of neighbours, so we will choose k=5. Next, we will calculate the Euclidean distance between the data points. Euclidean distance is the most

popular distance metric. We can also use Hamming distance, Manhattan distance, Minkowski distance as per our need. By calculating the Euclidean distance we got the nearest neighbours, as three nearest neighbours in category A and two nearest neighbours in category B. As we can see the 3 nearest neighbours are from category A, hence this new data point must belong to category A.

For classification: A class label assigned to the majority of K Nearest Neighbors from the training dataset is considered as a predicted class for the new data point.

For regression: Mean or median of continuous values assigned to K Nearest Neighbors from training dataset is a predicted continuous value for our new data point.

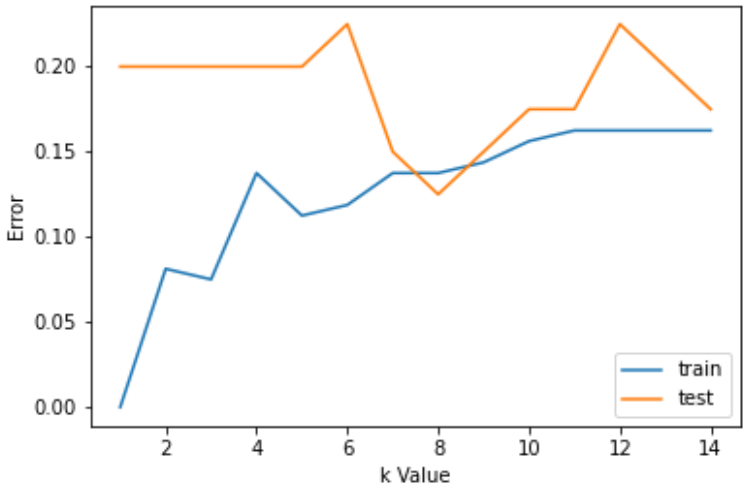
Model Representation

Here, we do not learn weights and store them, instead, the entire training dataset is stored in the memory. Therefore, model representation for KNN is the entire training dataset.

Choosing K value

K is a crucial parameter in the KNN algorithm. Some suggestions for choosing K Value are:

1. **Using error curves:** The figure below shows error curves for different values of K for training and test data.



At low K values, there is overfitting of data/high variance. Therefore test error is high and train error is low. At K=1 in train data, the error is always zero, because the nearest neighbour to that point is that point itself. Therefore though training error is low test error is high at lower K values. This is called overfitting. As we increase the value for K, the test error is reduced. But after a certain K value, bias/ underfitting is introduced and test error goes high. So we can say initially the test data error is

high(due to variance) then it goes low and stabilizes and with further increase in K value, it again increases(due to bias). The K value when test error stabilizes and is low is considered as the optimal value for K. From the above error curve we can choose K=8 for our KNN algorithm implementation.

2. Also, domain knowledge is very useful in choosing the K value.
3. K value should be odd while considering binary(two-class) classification.

Required data preparation

1. Data Scaling: To locate the data point in multidimensional feature space, it would be helpful if all features are on the same scale. Hence normalization or standardization of data will help.
2. Dimensionality Reduction: KNN may not work well if there are too many features. Hence dimensionality reduction techniques like feature selection, principal component analysis can be implemented.
3. Missing value treatment: If out of M features one feature data is missing for a particular example in the training set then we cannot locate or calculate distance from that point. Therefore deleting that row or imputation is required.

Advantages

1. Simple to implement.
2. Can handle large amounts of predictors.
3. More effective if the training data is large.
4. Training is fast.
5. Robust to noisy data.

Disadvantages

1. Always needs to determine the value of K which may be complex sometimes.
2. The computation cost is high because of calculating the distance between the data points for all the training samples.
3. No training stage, all the work is done during the test stage.
4. Slower compared to other models as processing is delayed until prediction is needed (Classification time is long).

Methodology:

In this project, the data of mouse dynamics was acquired to authenticate the user by training and testing using the KNN classifier from the data acquired by continuous authentication data of our group (group 12). The data was collected by all 5 members of our group for weeks. We got the data of 5 members of our group and assigned each member as one class. We used KNN-algorithm for training the model and then we calculated the accuracy.

Algorithm

1) Dataset collection :

- Reference:- Folder named '\data'.
- Collected using the mouse.jar application.
- Continuous data were collected over a period of time by various users.

2) Extracting Dataset :

- Reference:- extractor.py is used for extracting data
- Pre-processes the raw data and transforms it to contain pause time and statistical features for all combinations of mouse dynamics.
- The basics of mouse movement: X-coordinate, Y-coordinate, Theta value etc are extracted.

3) KNN Classifier :

- Data obtained from extractor.py is used as input.
- Data for each user is assigned a particular class value (0,1,2,..).
- Train-test split is done separately for each class to ensure the train and test set contain appropriate proportions of each class.
- The whole data is then merged while maintaining the train-test split.
- Select the number K of the neighbours
- Calculate the Euclidean distance of K number of neighbours
- Take the K nearest neighbours as per the calculated Euclidean distance.
- Among these k neighbours, count the number of the data points in each category.

- Assign the new data points to that category for which the number of neighbours is maximum.
- Accuracy is calculated from the predicted data and test data class label.

Libraries Used:

- Numpy
- Pandas
- Sklearn

Observation:

The accuracy obtained using k-NN algorithm=100% and accuracy obtained on cross validation=100%.

Analysis:

1. The data extracted of group members through the extractor. The function was assigned different class labels. Training and testing were done on splitting the data using the k-NN algorithm, giving an accuracy of 100%, which implies the classifier is working perfectly fine and give the accurate result for our mouse dynamics extracted data.
2. Also, on cross-validation, we get a cross-validation accuracy of 100%, which thus proves the effectiveness of the algorithm and the code implemented by our group.