Keystore Dynamics aplicado a la comparación de muestras de texto libre

Sebastián Sznur[1]

1. Ingeniero en Informática. Docente e investigador adjunto en la Universidad FASTA. Catamarca 735 1E, Mar del Plata, CP7600.

Resumen

Uno de los mecanismos utilizados para aumentar la seguridad de los sistemas informáticos es la aplicación de técnicas biométricas. Keystroke dynamics es una rama de la biometría que se dedica al estudio del reconocimiento del patrón de tecleo de una persona. Presenta las ventajas de ser no intrusivo, barato de implementar y utilizable luego de la etapa de autenticación. Este trabajo se centra en el análisis de los patrones de tecleo para determinar si dos muestras de texto escritas por usuarios en su labor cotidiana pertenecen o no a la misma persona.

Abstract

One of the many mechanisms used to increase the security of computer systems is the application of biometrics. Keystroke dynamics is a branch of biometrics that is dedicated to the study of keystrokes pattern recognition of a person. It has the advantages of being non-intrusive, inexpensive to implement and usable after the authentication. This paper focuses on analyzing keystroke patterns to determine whether or not two samples of written text belongs to the same user.

Palabras Clave: Biometría, Patrón de tecleo, Seguridad informática, Clasificación.

1. Introducción

En cualquier institución, empresa u organización, uno de los activos más importantes es la información y uno de los desafíos constantes es garantizar que dicha información sea administrada y utilizada de la manera en que se decidió y por quienes tienen los permisos para hacerlo. El gran incremento de sistemas informáticos ha simplificado la vida de las personas, aumentando también su dependencia a las computadoras y redes digitales. Estos avances traen aparejados nuevas amenazas en el acceso a la información, fraude y suplantación de identidad. [1]. Los sistemas que se centran en enfocar la seguridad en el inicio de sesión ponen un mecanismo de seguridad importante y confiable a la hora de dejar entrar a un usuario al sistema, pero no aseguran de ninguna manera que a lo largo de toda la sesión dicho usuario continúe siendo el mismo. Por otro lado, es igual o más difícil engañar al sistema de identificación constantemente mientras dure la sesión que sólo al inicio de ella [2].

Uno de los mecanismos utilizados para aumentar la seguridad de los sistemas informáticos es la aplicación de técnicas biométricas. Biometría es un término utilizado para indicar un conjunto de características que se cree que son únicas tanto de la fisiología cómo del comportamiento de un individuo, y por esa razón, difíciles de duplicar. [3]

Las técnicas de biometría basadas en características fisiológicas son más estables y, en condiciones normales, no varían con el tiempo. Dentro de este grupo se encuentran las huellas dactilares, las retinas, el iris, las venas de la mano, entre otras. Las características que describen comportamiento, en cambio, pueden verse influenciadas por la situación en la que se encuentra la persona. Dentro de este grupo se encuentra la firma manuscrita, la forma de ca-

minar y la forma en que se tipea en un teclado.

Keystroke dynamics es una rama de la biometría que se dedica al estudio del reconocimiento del patrón de tecleo de una persona. Presenta las ventajas de ser no intrusivo (pues los usuarios van a escribir en el teclado de cualquier forma), relativamente barato de implementar (no requiere hardware adicional) y además sigue siendo utilizable luego de la etapa de autenticación [4].

Dentro de keystoke dinamics las investigaciones pueden centrarse en el análisis de "texto fijo" o "texto libre" [4]. En el análisis de texto fijo todos los usuarios tipean las muestras utilizando siempre el mismo texto (como ser contraseñas o frases cortas), mientras que el análisis de texto libre implica un análisis de todo lo escrito. Esta última opción resulta más atractiva porque amplía el abanico de aplicaciones, pero tiene la desventaja de no contar con características fijas y fácilmente extraíbles, pues no se sabe a priori lo que los usuarios van a escribir.

Este trabajo se centra en el análisis de los patrones de tecleo para determinar si dos muestras de texto libre, escritas por usuarios en su labor cotidiana, pertenecen o no a la misma persona.

2. Trabajos relacionados

En líneas generales las investigaciones han probado que cada persona tiene su patrón de tecleo y que en mayor o menor medida éste puede ser utilizado para diferenciarlas.

Algunas investigaciones se centran en la extracción de características mientras que otras lo hacen directamente en la clasificación o identificación de los usuarios. En cuanto a la extracción de características, se ha analizado qué patrones deben tenerse en cuenta en la extracción y los posibles problemas que surgen [5]. También se han utilizado algoritmos genéticos para seleccionar las características más relevantes, logrando así reducir drásticamente los errores de clasificación [6]. Dentro de las características utilizadas, se han encontrado mejoras en el rendimiento al utilizar varios tiempos combinados (por ejemplo, el tiempo de presión de una tecla y el tiempo de liberación) [7] [8].

Es importante el tamaño de la muestra y el de los conjuntos de muestras [8][9]. Algunas investigaciones basadas en la autentificación [10] llegaron a la conclusión de que el nombre de usuario y la contraseña deben tener una longitud superior a ocho caracteres. Con respecto al tamaño del conjunto de muestras, varía dependiendo del método utilizado. Se ha reportado una excelente performance utilizando un clasificador de Bayes [11], pero considerando cuidadosamente el tamaño del conjunto de muestras. En caso de ser reducido, se aconseja el método K-Means. Las redes neuronales, aunque han probado su buen rendimiento, requieren una gran cantidad de muestras para su correcto funcionamiento [12].

En algunas publicaciones se advierte que el ritmo de escritura varía a lo largo del tiempo [13], y también se afirma que son varios los factores que lo alteran: condiciones físicas, mentales y del ambiente, pero que una vez que el usuario está familiarizado con lo que escribe las características se vuelven más estables [14]. También en otra investigación se marca la relación que hay con el haberse habituado a lo que se está escribiendo [7]. Esto podría derivar en una mayor facilidad para clasificar a las personas que cuenten con cierta experiencia.

Para tratar de minimizar el impacto que puedan tener las condiciones que alteran los factores de tecleo, se plantea que, pese a que el tiempo total que lleva escribir un párrafo puede variar, las relaciones entre los tiempos internos se mantienen [15]. Este trabajo tomó la decisión de separar lo ingresado en trígrafos (es decir, tomando las teclas de a tres), ordenarlo por tiempo y calcular la distancia entre distintas muestras. Mediante este sistema han logrado identificar y autentificar a los usuarios. Hay que tener en cuenta que estas características no provienen de cada usuario por separado sino que utilizaron distancias que miden la relación que hay entre una muestra y otra, no son valores que dependen de una muestra sino de un par de muestras.

La mayoría de los trabajos se centran en el análisis de palabras o frases cortas y se encuadran dentro de la categoría 'texto fijo', haciendo escribir a todas las personas el mismo texto. Sólo algunos tratan la temática correspondiente al 'texto libre'. [3]

Una de las primeras investigaciones en este último orden llegó a poder distinguir el estilo de cuatro usuarios [16]. Más desarrollado se puede encontrar en Gunetti [4], que siguiendo con el método utilizado en Bergadano [15], lo aplican al análisis de texto libre, teniendo solamente en cuenta los tiempos de presión de tecla. Han demostrado que sólo unas líneas sirven para identificar y clasificar a los usuarios con alto nivel de exactitud. Siguiendo esta línea se probó que se pueden utilizar solamente los dígrafos y trígrafos. [16]

Teniendo en cuenta lo investigado se procedió a:

- Obtener un conjunto de datos amplio que permita realizar investigaciones en el análisis de texto libre. No se encontró en internet un conjunto de datos disponible y pareció oportuno realizar uno que pueda ser utilizado para múltiples investigaciones y además fuese en español.
- Comenzar por indagar y mejorar un punto que hasta ahora sido poco estudiado y es primario: la comparación entre dos muestras de texto, para poder determinar si pertenecen o no a un mismo usuario. Hasta el momento las investigaciones se centran en la identificación, autenticación o clasificación de usuarios, pero poco dicen del análisis entre dos muestras de texto sin tener en cuenta datos previos. [16]

Se partió de unos de los métodos que habían presentado mejores resultados y además implicaba una visión novel del problema: la obtención de características que surgen de la relación de dos muestras presentada por Gunetti [4].

3. Materiales y métodos

Las entidades mínimas que se utilizaron para el análisis fueron los dígrafos y trígrafos. En la temática se llama dígrafo al par compuesto por dos teclas (presionadas consecutivamente) y un tiempo (el que transcurrió entre que se presionaron ambas teclas). Por ejemplo al escribir en un teclado la palabra ELLA se pueden obtener 3 dígrafos: el digrafo EL (con tiempo x1), el dígrafo LL (con tiempo X2) y el dígrafo LA (con tiempo x3).

De manera análoga, un trígrafos está compuesto por tres teclas presionadas consecutivamente y por los tiempos que transcurrieron entre que se presionaron las teclas. Del ejemplo anterior se pueden extraer 2 trígrafos: ELL (con los tiempo x1 y x2) y LLA (con los tiempo x2 y x3).

Lo que se buscó desde un principio es utilizar solamente los dígrafos y los trígrafos, sin tener en cuenta el orden en el que fueron escritos, la palabra que componían, ni ninguna otra característica. Se quiere obtener de cada muestra de datos simplemente el listado de los dígrafos y trígrafos que la componen.

3.1 Obtención de datos

No es fácil obtener las pulsaciones de teclas de los usuarios, especialmente las de texto libre. Son varios los problemas más comunes para obtener dicho conjunto de datos. Primero, los usuarios no quieren compartir sus datos por cuestiones relacionadas a la privacidad. Segundo, esos usuarios deben escribir libremente mientras realizan sus tareas cotidianas y sin supervisión, pues las acciones deben ser lo más reales posibles.

Se realizó un software para obtener los datos de usuarios reales mientras realizan sus tareas en su computadora. El programa podía ser encendido y apagado a voluntad del usuario y solo se enviaban los datos ya procesados (dígrafos y trígrafos) para preservar su privacidad. El software se distribuyó entre compañeros de trabajo y personas conocidas. Fueron personas con características diferentes, pero coinciden en que todas usaban una computadora en su labor diaria.

Después de un tiempo, se logró obtener un conjnto de datos grande con usuarios reales. Las principales características de los datos obtenidos son: 17 usuarios diferentes, 373 muestras de texto, 2.726.203 dígrafos totales y 2,553.494 trígrafos totales.

3.2 Extracción de características

Para comparar las muestras entre si, primero es necesario extraer las características que sirvan para el análisis. En el presente trabajo se utilizaron las distancias propuestas por Gunetti [4].

Dadas dos muestras, la idea es compararlas sin importar el texto que se ha escrito en cada una de ellas. Para esto se extraen los dígrafos en

común de ambas muestras junto con sus tiempos, siendo esta la información que comparten. En el caso de que algún dígrafo en común se presente más de una vez en alguna muestra, el mismo se extrae una sola vez con el tiempo promedio de sus ocurrencias. Los dígrafos no comunes entre las dos muestras se ignoran. Luego debe aplicarse alguna medida de distancia a dicha información. Se proponen dos distancias que devuelven un valor real entre 0 y 1 que se denominan distancia R (por relativo) y distancia A (por absoluto). Se obtienen así dos características que salen, no de una sola muestra, sino de la relación entre dos muestras.

Para obtener la distancia R entre dos muestras M1 y M2 se toman los dígrafos que las mismas tienen en común, y se los ordena por tiempo, como se muestra en el ejemplo de la figura 1. A continuación se calcula el grado de desorden entre dichas muestras que consiste en sumar las distancias entre las posiciones de cada dígrafo. Claramente, si los dígrafos están en el mismo orden en ambas muestras, el grado de desorden es 0, mientras que el máximo desorden se encuentra cuando están en el orden inverso. En el ejemplo el grado de desorden es 0+2+0+2=4. Si la cantidad de dígrafos en común entre dos muestras es k, es conveniente normalizar su desorden para poder compararlo con otras muestras que tengan un valor diferente de k. Esto puede hacerse dividiendo el grado de desorden por el valor de máximo desorden de un vector de k elementos. De esta forma se puede comparar el desorden de vectores de diferente tamaño. Luego de esta normalización, el grado de desorden cae entre 0 (si está ordenado) y 1 (si está en el orden inverso).

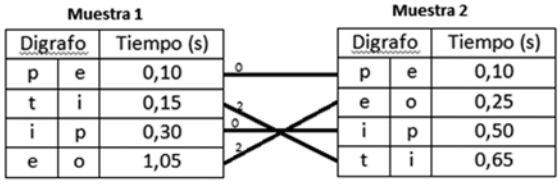

Figura 1: *Ejemplo del cálculo del grado de desorden entre dos muestras*

En otras palabras, si dos muestras M1 y M2 comparten k dígrafos, la distancia R de M2 a M1 es la suma de las distancias de cada dígrafo de M2 a la posición del mismo dígrafo en M1, dividido por el máximo desorden mostrado por un vector de k elementos.

Lo explicado anteriormente se aplica de la misma forma a los trígrafos.

El problema que presenta la distancia R es que pasa por alto cualquier valor absoluto del tiempo de las muestras que considera, considerando solamente el valor relativo. Por ejemplo, en el caso de que el tiempo de cada dígrafo en M1 sea exactamente el doble que en M2, la distancia R daría 0. Esta distancia falla al discriminar entre muestras de dos intrusos diferentes que tengan ritmos de tecleo similares, incluso si uno de ellos es mucho más rápido que el otro.

A diferencia de la distancia R, la distancia A solo considera el valor absoluto de la velocidad de tecleo de cada par de dígrafos idénticos en las dos muestras que se están comparando. Si se está comparando el mismo dígrafo de dos muestras M1 y M2, ocurriendo con tiempos t1 y t2 respectivamente, se dice que los dígrafos son similares si:

$$1 \leq \frac{\max(t1,t2)}{\min(t1,t2)} \leq t$$

para alguna constante t mayor a 1 (un valor óptimo para esa variable es 1,25 según los autores del método). Entonces, siendo n el número de dígrafos similares entre M1 y M2, N el número total de dígrafos en común entre M1 y M2, la distancia A se define como 1 – (n/N). Como consecuencia, si no hay pares similares de dígrafos entre las muestras, la distancia A vale 1. Si todos los dígrafos en común son similares, la distancia A es 0.

Una medida de distancia que puede resultar más significativa entre dos muestras es la cumulativa que se obtiene combinando los valores de la distancia R (o la distancia A) obtenida usando dígrafos con la obtenida usando trígrafos. Para hacer esto, se debe tener en cuenta la cantidad de dígrafos y trígrafos en común entre las muestras. Si dos muestras M1 y M2 comparten N n-grafos y M m-grafos, con N>M, y siendo Rn la distancia R para n-grafos, Rm la distancia R para m-grafos se define la distancia R cumulativa

como:

$$R_{n,m} = R_n + R_m \times \frac{M}{N}$$

Del mismo modo se define la distancia A cumulativa.

Además de las seis distancias explicadas anteriormente y extraídas de Gunetti [4], para esta investigación se incorporaron otras cuatro para intentar mejorar la performance. Cuando se calculan los trígrafos se tienen dos tiempos: el lapso que transcurre entre la primer y segunda tecla (l12) y el lapso que transcurre entre la segunda y tercer tecla (l23). La opción más común es utilizar la suma de esos dos lapsos como el tiempo del trígrafos, o sea:

$$t = I_{12} + I_{23}$$

Pero hay otras opciones que brindan diferente información que también puede llegar a ser útil.

En este trabajo se utilizaron otras dos formas diferentes de obtener el tiempo del trígrafo: la multiplicación de los lapsos:

$$t = I_{12} \times I_{23}$$

que se llamó "trigrafosMult" y la división del primero por el segundo:

$$t = I_{12}/I_{23}$$

que se llamó "trigrafosRatio". A estas dos nuevas medidas se le calcularon las distancias A y R, quedando así cuatro distancias nuevas y un total de diez distancias.

3.3 Clasificadores

Un clasificador es un algoritmo utilizado para asignar un elemento entrante no etiquetado en una categoría concreta conocida. Generalmente cuentan con dos etapas: la primera de entrenamiento, en la que se le presentan los elementos de entradas con sus categorías deseadas para que el modelo aprenda a etiquetar correctamente, y una segunda etapa de ejecución en la que se le presentan elementos nuevos (desconocidos) que el clasificador tiene que etiquetar.

Ya que solo hay dos posibilidades (que el par de muestras pertenezca al mismo usuario o que no) se debió escoger entre clasificadores binarios: la categoría de salida que asigna solo puede ser 0 (o negativo) o 1 (o positivo).

Dado que en la bibliografía se han obtenido resultados con varios clasificadores diferentes, no se tiene ninguna respuesta concluyente en cuanto a cual se desempeña mejor, ya que son dependientes de lo que representa el conjunto de datos que se está analizando. Por dicho motivo y por no encontrarse en la bibliografía ninguna investigación que haya presentado los datos de la misma manera, se utilizó un conjunto de clasificadores para evaluar cuál de ellos se desempeñaba mejor.

Los clasificadores utilizados fueron:

Redes Neuronales Artificiales (RNA)
k-Nearest-Neighbors (kNN)
Arboles de decisión (dtree)
Análisis Discriminante (Disc)
Ensemble Methods (EnM)

Los Ensemble Methods se construyen con un modelo base y un algoritmo. Para este trabajo se utilizaron para el modelo base Árboles de decisión con los algoritmos Bootstrap aggregating (Bag), Boosting (Logit-Boost, Gentle-Boost y Ada-Boost) y kNN con el algoritmo SubSpace.

3.4 Desempeño

Para evaluar el desempeño de los clasificadores binarios se tienen varias medidas [18]. En este trabajo se utilizaron Exactitud (o accuracy), Razón de Falsos Positivos, Sensibilidad y Valor Predictivo Positivo (VPP).

En primer lugar hay que considerar el espacio de resultados posibles: se tienen elementos positivos y negativos, que a su vez pueden ser etiquetados por el clasificador de manera correcta o incorrecta.

Los elementos a priori negativos que fueron etiquetados como tales se denominan Verdaderos Negativos (VN)

Los elementos a priori negativos que fueron etiquetados como positivos se denominan Falsos Positivos (FP).

Los elementos a priori positivos que fueron etiquetados como negativos se denominan Falsos Negativos (FN).

Y los elementos positivos que fueron etiquetados correctamente se denominan Verdaderos Positivos (VP).

La exactitud (accuracy) de un clasificador viene determinada por la cantidad de aciertos que tuvo sobre la cantidad de elementos totales:

$$Acc = \frac{VN + VP}{VN + FP + FN + VP}$$

La Razón de Falsos Positivos (FPR) determina la probabilidad de que un elemento negativo sea considerado como positivo, o sea es la cantidad de Falsos Positivos sobre la cantidad total de elementos negativos reales:

$$FPR = \frac{FP}{FP + VN}$$

La sensibilidad viene dada por la probabilidad de clasificar correctamente un elemento positivo, o sea la cantidad de elementos identificados correctamente como positivos sobre la cantidad de elementos positivos totales que se tenía:

$$S = \frac{VP}{VP + FN}$$

El Valor Predictivo Positivo es la probabilidad de que un elemento sea realmente positivo una vez conocido que el clasificador lo etiquetó como tal, o sea determina la precisión en la recuperación de la información.

$$VPP = \frac{VP}{VP + FP}$$

Un valor cercano a 1 asegura que cuando el clasificador etiqueta un elemento como positivo, éste realmente sea positivo.

3.5 Validación cruzada

Unos de los efectos no deseados que pueden tener los algoritmos de aprendizaje que emplean los clasificadores es el sobreajuste (overfitting). Esto ocurre cuando el clasificador sabe clasificar casi perfectamente los datos utilizados en el entrenamiento pero falla cuando se le presentan datos nuevos, no pudiendo generalizar correctamente. En este caso las métricas de desempeño pueden ser excelentes pero el clasificador va a estar comportándose de manera errática en la realidad. [19]

Si bien se sabe que ninguna estimación de la exactitud va a ser correcta todo el tiempo, se busca utilizar un método que se adapte bien a los sesgos y tendencias de los datos típicos del mundo real. Para lograr esto, el conjunto de datos suele dividirse en datos de entrenamiento y datos de evaluación: se entrena con el primer conjunto y se calculan las medidas de desempeño con el segundo. [20]

Para garantizar que los resultados son independientes de la partición entre datos de entrenamiento y evaluación se utilizan técnicas de validación cruzada.

En este trabajo se utilizó la técnica k-fold con k=5. Los datos de muestra se dividieron en 5 subconjuntos. Uno de los subconjuntos se utiliza como datos de validación y los 4 restantes como datos de entrenamiento. El proceso de validación cruzada es repetido durante 5 iteraciones, con cada uno de los posibles subconjuntos de datos de validación. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Todos los resultados reportados fueron obtenidos mediante este procedimiento.

4. Experimentos y resultados

Se tiene una base de datos con 373 muestras pertenecientes a 17 usuarios. Como las distancias de una muestra consigo misma son siempre 0 y además el cálculo de las distancias es simétrico (d(m1,m2)=d(m2,m1)), se cuenta con un total de 69.378 pares de muestras (373*372/2). De esos pares, 8.669 corresponden a muestras de un mismo usuario (o sea datos con etiquetas

positivas o 1) y los restantes 60.709 corresponden a pares de muestras de diferente usuario (datos con etiquetas negativas o 0).

Primero se procedió a evaluar si las nuevas medidas propuestas en este trabajo mejoran el desempeño. Se corrieron pruebas con todos los clasificadores, primero con las 6 distancias propuestas por Gunetti [4] y luego se agregaron las 4 restantes. Los resultados pueden verse en la Tabla 1.

Cada clasificador tiene distintos parámetros que, de acuerdo al problema, permiten que se adapten mejor a la solución. Se fue probando para cada clasificador distintas variantes hasta encontrar la mejor.

Se observa que el resultado en la mayoría de los clasificadores mejora levemente al utilizar las 10 distancias.

Un punto a tener en cuenta es que al encontrarse desbalanceada la cantidad de elementos positivos y negativos en los datos (debido a la naturaleza del problema) se puede observar rápidamente que la exactitud no nos va a decir todo lo que deseamos: un clasificador que etiquete a todos los elementos con 0 va a tener accuracy=87,5%. Ahí es cuando se vuelven relevantes las medidas mencionadas anteriormente: FPR, Sensibilidad y VPP.

En la tabla 2 se muestran los valores de las demás funciones de desempeño para los tres mejores clasificadores de la tabla 1, utilizando las 10 distancias.

Tabla 2: Valores de las funciones de desempeño para los 3 mejores clasificadores

	RNA	*EnM-Bag-Tree*	*EnM-Log-Tree*
Accuracy	93,98%	94,19%	93,66%
FPR	2,71%	2,44%	2,95%
Sensibilidad	70,78%	70,62%	69,89%
VPP	78,87%	80,52%	77,17%

Los resultados obtenidos son bastante alentadores. No hay grandes diferencias entre estos 3 clasificadores en cuanto a los resultados, aunque EnM-Bag-Tree se destaca levemente por sobre el resto.

FPR en ningún caso supera el 3%, lo que significa que son escasas las posibilidades de tener dos muestras de diferentes personas clasificadas erroneamente. Debido a la gran cantidad de ejemplos negativos disponibles los clasificadores no mostraron mayores problemas en este sentido.

La sensibilidad en los tres casos se aproxima al 70%. Esto quiere decir 7 de cada 10 veces que le presentemos al clasificador dos muestras de una misma persona las va a etiquetar correctamente.

Por último se tiene un VPP del 80%. Si el clasificador nos devuelve un resultado positivo, hay un 80% de probabilidades de que esté en lo correcto.

Una característica de los algoritmos utilizados para los Ensemble Methods es que se le puede dar un peso a los errores cometidos. Por ejemplo, si se tratase de un sistema de diagnóstico médico, a fines de un rápido tratamiento, son menos graves los falsos positivos (gente sana diagnosticada con una enfermedad) que los falsos negativos (gente enferma con diagnóstico sano). Con este objetivo en mente se le puede pedir al modelo que reduzca un tipo de error en pos de aumentar el otro.

Para probar los límites del clasificador y lograr una mayor certeza al obtener un resultado positivo, se fueron modificando paulatinamente los costos de los errores para mejorar el VPP. En la tabla 3 se muestran los resultados para EnM-Bag-Tree.

Tabla 3: variación de las funciones de desempeño para el clasificador *EnM-Bag-Tree* al ir variando el costo de los errores

Accuracy	94,19%	94,00%	93,65%	93,29%	92,37%	92,84%	92,56%	92,37%
VPP	80,52%	86,16%	89,43%	91,46%	93,23%	94,88%	95,63%	95,92%
Sensibilidad	70,62%	61,97%	55,75%	51,03%	47,79%	45,11%	42,43%	40,66%

Se puede observar que tolerando una sensibilidad del 50% el VPP sobrepasa el 90%. Es decir, se identificarán la mitad de los pares de

Tabla 1: Valores de la función Accuracy para todos los clasificadores utilizando 6 y 10 distancias.

	RNA	EnM-Bag-Tree	EnM-Log-Tree	EnM-Gen-Tree	EnM-Rob-Tree	EnM-Ada-Tree	EnM-Sub-kNN	D-Tree	kNN	Disc
6 dist.	93,75%	93,78%	93,65%	93,49%	93,41%	93,45%	90,17%	91,72%	90,68%	91,92%
10 dist.	93,98%	94,19%	93,66%	93,52%	93,14%	93,43%	91,66%	92,20%	91,62%	91,88%

muestra positivos que se presentan al modelo, pero si se identificó como tal hay amplias posibilidades de que se esté en lo cierto.

5. Conclusiones

Se partió con el objetivo de poder discernir, por medio del análisis del patrón de tecleo, si dos muestras diferentes de texto corresponden o no a una misma persona. Para abordar el problema primero se tuvo que diseñar e implementar un software que permitiese obtener una cantidad abundante de datos de usuarios escribiendo normalmente en sus computadoras. Se logró obtener un conjunto de datos con 373 muestras pertenecientes a 17 personas.

Se siguió por el camino planteado en Gunetti, se lo adaptó para el problema planteado y se aportaron nuevas formas de calcular las distancias que resultaron en una mejora si complementan con las anteriores.

Se buscó a través de diez clasificadores encontrar el mejor modelo que se ajuste al problema. Los mejores resultados fueron dados por un Ensemble Method compuesto por Árboles de decisión y con el algoritmo Bootstrap aggregating.

Los resultados obtenidos son alentadores. Es posible, si se tienen los patrones de tecleo de dos muestras de texto, analizarlas por los clasificadores aquí vistos y decir con cierto grado de certeza si fueron o no escritas por la misma persona. Se logró una exactitud superior al 94% y el modelo elegido se puede adaptar para conseguir una seguridad superior al 90% en los resultados positivos obtenidos.

Este es el primer paso en la búsqueda de nuevas técnicas que por medio de keystroke dynamics refuercen la seguridad de un sistema. A futuro se buscará, con el mismo conjunto de datos y adaptando el método visto, implementar la clasificación de usuarios (dada una muestra, saber a qué usuario le corresponde) y la autentificación (dada una muestra, saber si corresponde o no al usuario que dice ser).

Bibliografia

[1] Monrose, F., & Rubin, A. D. (2000, 12). Keystroke dynamics as a biometric for authentication. Future Generation Computer Systems, 16(4), 351-359. doi: 10.1016/S0167-739X(99)00059-X

[2] Sim, T., Zhang, S., Janakiraman, R., & Kumar, S. (2007, 12). Continuous Verification Using Multimodal Biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4), 687-700. doi: 10.1109/TPAMI.2007.1010

[3] Ahmed, A.A.; Traore, I. (2013). Biometric Recognition Based on Free-Text Keystroke Dynamics. Cybernetics, IEEE Transactions on , vol.PP, no.99, pp.1,1, 0 doi: 10.1109/TCYB.2013.2257745

[4] Gunetti, D., & Picardi, C. (2005, 12). Keystroke analysis of free text. ACM Transactions on Information and System Security, 8(3), 312-347. doi: 10.1145/1085126.1085129

[5] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. New York: Wiley.

[6] Yu Enzhe y Cho Sungzoon (2003). Biometrics-based Password Identity Verification: Some Practical Issues and Solutions. Http://dmlab.snu.ac.kr.

[7] Araújo L. C. F., Lizárraga M. G., et al. (2005). Autentificación personal por dinámica de tecleo basada en lógica difusa. IEEE COMPUTER SOCIETY.

[8] Robinson, J., Liang, V., Chambers, J., & Mackenzie, C. (1998, 12). Computer user verification using login string keystroke dynamics. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 28(2), 236-241. doi: 10.1109/3468.661150

[9] Paeckock Alen, Ke Xian, et al. (2004). Typing Patterns: A Key to User Identification. IEEE COMPUTER SOCIETY.

[10] Cheng-Huang Jiang, Shiuhpyng Shieh, et al. (2007). Keystroke statistical learning model for web authentication. Proceedings of the 2nd ACM symposium on Information, computer and communications security. Singapore, ACM.

[11] Cho Tai-Hoon (2006). Pattern Classification Methods for Keystroke Analysis. SICE-ICASE International Joint Conference.

[12] Obaidat M. S. y Sadoun Balqies (1997). Verification of Computer Users Using Keystroke Dynamics. IEEE COMPUTER SOCIETY.

[13] Kacholia Varun y Pandit Shashank (2004). Biometric Authentication using Random Distributions (BioART). shashankpandit.com.

[14] Mroczkowski Piotr (2004). Identity Verification using Keyboard Statistics. Linkoping University, Electronic Press.

[15] Bergadano, F., Gunetti, D., & Picardi, C. (2002, 12). User authentication through keystroke dynamics. ACM Transactions on Information and System Security, 5(4), 367-397. doi: 10.1145/581271.581272

[16] Shepherd S J (1995). Continuous Authentication by analysis of keyboard typing characteristics.IEEE COMPUTER SOCIETY.

[16] Alsultan, A. Warwick K. (2013). Keystroke Dynamics Authentication: A Survey of Free-text Methods. IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No 1

[18] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). Introduction to Information Retrieval. 1st ed. Cambridge: Cambridge University Press. Cambridge Books Online. Web. 13 March 2014. http://dx.doi.org/10.1017/CBO9780511809071.017

[19] Devijver, P. A., and J. Kittler (1982). Pattern Recognition: A Statistical Approach, Prentice-Hall, Londres.

[20] Kohavi Ron (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Morgan Kaufmann. P 1137-1143.

Autor

Sebastián Sznur, Ingeniero en Informática. Docente e investigador adjunto en la Universidad FASTA. Catamarca 735 1E, Mar del Plata, CP7600. Cel: (0223)154267166. szsebas@gmail.com

Colaboradores

Sebastián Fink, Ingeniero en Informática. Auxiliar de investigación graduado en la Universidad FASTA. safink@gmail.com

Marcos Jesús Vivar, Auxiliar de investigación alumno en la Universidad FASTA. m4rk1ch@gmail.com

Sebastián García, Ingeniero en Informática. Investigador titular en la Universidad FASTA. eldraco@gmail.com

El trabajo se enmarca bajo el proyecto de investigación “Biometría del Comportamiento” de la Facultad de Ingeniería de la Universidad FASTA.