Detector de mentiras utilizando señales EEG

P.Mariño y A.R.Telli
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, UNL-CONICET

Resumen

Como la detección de mentiras utilizando la variabilidad electroencefalográfica ha despertado mucho interés en los últimos años, este informe utiliza señales de electroencefalogramas (EEG), identificando bandas de frecuencias sensibles a los estímulos de un sujeto, junto con la recopilación de otras características extraídas de las señales EEG, para poder predecir la veracidad o falsedad del evento proporcionado por dicho sujeto. Para realizar dicha clasificación se utilizaron dos clasificadores clásicos que fueron el método de los k vecinos más cercanos (KNN), y el algoritmo de Naive Bayes. Este informe proporciona tanto un modelo de extracción de características para detectar mentiras como una comparación entre el porcentaje de predicción de dichos clasificadores. Los resultados mostraron que con la normalización de los datos y el uso de características específicas en nueve bandas de frecuencia se alcanzo un rendimiento óptimo del 79.33 con el uso de 36 características. Sin embargo, la precisión mostró variabilidad al incluir características adicionales, lo que destaca la complejidad de la detección de mentiras basada en EEG.

1. Introducción

La detección de mentiras ha sido un tema de interés a lo largo de los siglos, desde la observación del lenguaje corporal hasta los modernos polígrafos. Recientemente, el uso del electroencefalograma (EEG) ha emergido como una herramienta prometedora para detectar mentiras, registrando la actividad eléctrica del cerebro.

Este informe explora el desarrollo y aplicación de un detector de mentiras basado en señales EEG, evaluando su precisión y metodología. Se basa en un artículo científico que utilizó un dispositivo EEG con el electrodo "FP1"para recopilar señales [1], que luego se transformaron al dominio frecuencial mediante la Transformada de Fourier (FFT). Posteriormente, se dividieron en bandas y se extrajeron características relevantes usando operaciones estadísticas. Aunque el artículo original utilizó teoría difusa y "Machine Learning"para la clasificación, en nuestro análisis se emplearon clasificadores más simples sugeridos por el profesor "L.E.Di Persia".

La recolección de datos fue un proceso complejo debido a la falta de acceso a la base de datos original. Se encontró una base de datos adecuada en un articulo científico [2], que utilizaba el electrodo "FP1", muestreada a 500Hz, con 15 sujetos sometidos a 5 sesiones de pruebas de mentiras y verdades, resultando en 150 señales EEG.

Se realizaron varias pruebas con resultados mixtos, que se discutirán en detalle. Este informe presenta los métodos y recursos utilizados, análisis de resultados y conclusiones obtenidas.

2. Métodos

En esta sección se presentan los métodos utilizados para llevar a cabo la detección de mentiras mediante el análisis de señales EEG.

2.1. Recolección de Datos:

Primero se cargaron las señales EEG desde la base de datos seleccionada en el software GUI Octave. Para la adquisición de los datos EEG la base de datos seleccionada utilizo doce electrodos (Fp1, Fp2, F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, Oz) de un sistema Internacional 10-20. Dichos datos se digitalizaron a una frecuencia de muestreo de 500 Hz. Todos los electrodos estaban referenciados al lóbulo de la oreja derecha.

El protocolo experimental dividió a los participantes en dos grupos de 15 (inocentes y culpables). Se utilizaron seis joyas diferentes como estímulos. Los inocentes guardaban y memorizaban una joya, mientras que los culpables elegían y robabanüna de dos joyas. Luego, los participantes veían imágenes de las joyas en una pantalla y presionaban 'Sí' o 'No' según reconocieran los ítems. Este proceso evaluaba las respuestas EEG a estímulos relacionados con la culpabilidad o inocencia.

En este informe se utilizaron exclusivamente los datos proporcionados por el electrodo situado en la posición FP1. Esta elección se debe a que el trabajo de referencia empleado utilizada este electrodo y además los sujetos de la base de datos estaban expuestos a estímulos visuales, y las respuestas neuronales a estos estímulos se generan en zonas del cerebro más cercanas a los ojos [3]. Por lo tanto, el electrodo FP1, ubicado en la parte frontal del cráneo, es particularmente adecuado para captar la actividad eléctrica relevante para el análisis.

2.2. Transformada Rápida de Fourier (FFT):

Para convertir las señales EEG del dominio temporal al dominio frecuencial, se aplicó la Transformada Rápida de Fourier (FFT). La FFT descompone una señal en sus componentes de frecuencia, lo que permite analizar la distribución de energía en diferentes bandas de frecuencia. La ecuación básica para la FFT es:

$$X(f) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi fn/N} \tag{1}$$

2.3. Densidad Espectral de Potencia (PSD):

Luego, se calculó la Densidad Espectral de Potencia (PSD) para cada señal. La PSD estima cómo se distribuye la potencia de la señal a través de las distintas frecuencias y se obtiene aplicando la siguiente ecuación:

$$P(f) = \frac{1}{N} |X(f)|^2 \tag{2}$$

2.4. Extracción de Características:

Una vez que la señal está definida en el dominio frecuencial utilizando la FFT, procedemos a realizar la extracción de características. Comenzamos calculando la densidad espectral de potencia de la señal (PDS) y luego dividimos la señal en 9 bandas de frecuencias distintas: Gamma, Theta, Alpha, Low Alpha, High Alpha, Beta, Low Beta, High Beta y Total. Las frecuencias de cada banda se encuentran detalladas en la Figura 1.

Bandas de Frecuencias (Hz)	Características	Descripción
Gamma 25-40	Gamma.Min, Max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
Theta 4-7	Theta.Min, Max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
Alpha 8-13	Alpha.Min, Max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
Alpha 8-10	LAlpha.Min, Max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
HAlpha 10-12	HAlpha.Min, Max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
Beta 14-30	Beta.Min, Max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
LBeta 14-25	LBeta.Min, max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
Hbeta 25-35	HBeta.Min, max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
Total 4-35	Total. Min, max, ave, y SD	Máximo, Mínimo, Promedio, SD (Desvió Standard),
Toda la Banda 0-500	Entropía, Media, Índice de Frecuencia Dominante	Entropía, Media, Índice de Frecuencia Dominante

Figura 1: Descripción de las bandas de frecuencias y sus características.

Para cada una de las 9 bandas, calculamos las siguientes características: máximo, mínimo, desviación estándar y valor promedio. En total, para estas 9 bandas tendremos 36 características.

Además, para toda la señal con su FFT aplicada y calculada su PDS, se extrajeron 3 características adicionales:

• La entropía:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \tag{3}$$

Es una medida de la incertidumbre o desorden en un sistema, refiriéndose en señales a la cantidad de información y su grado de imprevisibilidad.

• El valor medio:

$$\bar{f} = \frac{\sum_{i=1}^{n} f_i P(f_i)}{\sum_{i=1}^{n} P(f_i)} \tag{4}$$

Es el promedio ponderado de los valores de la señal f_i multiplicados por sus probabilidades $P(f_i)$, mostrando el valor central de la señal considerando su distribución probabilística.

• El Índice de Frecuencia Dominante:

$$f_{\text{max}} = \arg \max_i X_{\text{pds}}(i) \tag{5}$$

Es el valor de la frecuencia i que maximiza la densidad espectral de potencia $X_{\text{pds}}(i)$, identificando la frecuencia más prominente en la señal para caracterizar su contenido espectral y propiedades de frecuencia.

En conclusión, por cada señal EEG tendremos 39 características; 4 características para cada una de las 9 bandas, osea 36 características y las 3 características mas mencionadas anteriormente.

2.5. Normalización de los Datos:

En el proceso de análisis, una vez se recopilaron todas las características, se procedió a normalizar cada una de ellas. La normalización juega un papel crucial en el procesamiento digital de señales, siendo fundamental para mejorar tanto la eficiencia como la precisión del algoritmo de clasificación. Este proceso es esencial ya que elimina las variaciones de escala, permitiendo una comparación más significativa entre las distintas características.

La normalización es especialmente útil cuando se trabaja con características o variables que se encuentran en escalas numéricas diferentes. Al ajustar estos valores a una escala común, se facilita su interpretación y análisis, contribuyendo así a una toma de decisiones más informada y precisa.

Para nuestros análisis, estamos utilizando clasificadores que se entrenan con datos de la base de datos. Durante el proceso de entrenamiento y predicción, estamos aplicando la normalización a los datos. Esto significa que ajustamos los valores de las características para que estén dentro de un rango específico, generalmente entre 0 y 1. Esta práctica simplifica el proceso de clasificación y hace que nuestras predicciones sean más fiables. En resumen, al normalizar los datos, estamos obteniendo resultados consistentes y confiables en nuestras tareas de análisis y predicción.

$$\text{Valor normalizado} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{6}$$

2.6. Clasificación:

En esta sección no se entrara muy en profundidad como funcionan los algoritmos que utilizamos ya que excede los conceptos de la asignatura. Los clasificadores sirven para al pasar una señal, teniendo datos de entrenamiento, predecir si esa persona esta mintiendo. Los datos de entrada que los algoritmos si o si deben tener, los datos de entrenamiento y las clases, quiere decir que le pasamos todas las señales que querríamos que entrene y si esa señal es una señal en donde el sujeto miente o dice la verdad, que esta ultima es la clase. Los dos clasificadores que utilizamos, fue:

• Naive Bayes: El clasificador de Naive Bayes es un modelo de aprendizaje automático que se basa en el teorema de Bayes y asume que las características son independientes entre sí. Durante el entrenamiento, calcula las probabilidades condicionales de cada característica dada cada clase de salida. En la fase de predicción, utiliza estas probabilidades para calcular la probabilidad de que una instancia pertenezca a una clase dada las características observadas, y elige la clase con la probabilidad más alta.

• KNN: El clasificador de K-Nearest Neighbors (KNN) es un algoritmo de aprendizaje automático que determina la clase de un nuevo punto de datos basándose en la clase de sus vecinos más cercanos en el conjunto de entrenamiento. Funciona calculando la distancia entre el nuevo punto y todos los puntos de entrenamiento, seleccionando luego los K puntos más cercanos. Para este algoritmo se debe elegir un K apropiado, que se adapte y tenga mejor rendimiento con el conjunto de datos específico.Es importante no elegir un K demasiado pequeño para evitar el sobreajuste, ni demasiado grande para prevenir el subajuste. Se eligió un K impar ya que si optábamos un K par, podría pasar, llegado el caso, que exista como un .[em]pate.[en]tre los puntos mas cercanos.

Un clasificador propuesto pero descartado, fue el del articulo mencionado, que utiliza teoría difusa, pero como la base de datos es distinta a la del articulo y son temas que nos exceden, quedo descartado por completo.

3. Pruebas, resultados y discusión

3.1. Introducción

Para comenzar esta sección, es importante aclarar ciertas cuestiones. Se realizaron varias pruebas de distintos tipos. Nuestra base de datos posee 15 sujetos con 5 sesiones cada uno, por lo que las primeras pruebas se hicieron con los 15 sujetos, variando las sesiones de prueba.

3.2. Primer Conjunto de Pruebas

Por ejemplo, se tomaron las sesiones 1, 2, 4 y 5 para entrenamiento y la sesión 3 para prueba. Lógicamente, en ningún entrenamiento se utilizaron sesiones que después íbamos a probar, ya que sería en vano. En esta primera serie de pruebas, se variaron tanto las sesiones de entrenamiento como las de prueba. Además, se hicieron pruebas con los datos normalizados y no normalizados para analizar los resultados obtenidos.

3.3. Segundo Conjunto de Pruebas

En una segunda instancia, se tomaron 14 sujetos con sus 5 sesiones para entrenamiento, y se tomó 1 sujeto con sus 5 sesiones para prueba. Por ejemplo, se tomó el sujeto 15 para prueba y los otros

14 para entrenamiento, utilizando en ambos casos las 5 sesiones. Se realizaron pruebas con ambos clasificadores mencionados, nuevamente utilizando datos normalizados y no normalizados.

3.4. Consideraciones Adicionales

En las figuras anteriores, se puede ver cómo se ha variado el experimento. Antes de definir y escribir acerca de los resultados obtenidos, es importante aclarar que el valor de K utilizado para el algoritmo de KNN varió según el tipo de prueba: se usó K=5 al variar las sesiones y K=7 al variar los sujetos. Aunque se podían utilizar algoritmos de detección del mejor K mediante validación cruzada, se sugirió por parte del docente L.E. Di Persia probar con algunos valores de K y ver cuál funcionaba mejor. Así se hicieron pruebas preliminares para encontrar el mejor K.

3.5. Interpretación de Resultados

Los resultados obtenidos se deben interpretar en porcentajes. Por ejemplo, si se indica un 5 % en "PromedioKNN", significa que el clasificador KNN tiene una tasa de aciertos del 5 %. Esto aplica también para "VerdadKNN", "MentiraKNN", "VerdadNB", "MentiraNB", "PromedioKNN" y "PromedioNB".

Al referirnos a "sesiones de entrenamiento", nos referimos a las sesiones utilizadas para entrenar el clasificador. De igual manera, "sesiones de prueba" se refiere a las sesiones utilizadas para probar el clasificador. Lo mismo aplica para "sujetos de prueba" y "sujetos de entrenamiento".

3.6. Resultados y Análisis

A continuación, se comentarán los resultados obtenidos de estas pruebas.

Pruebas "A"	PRUEBA 1	PRUEBA 2	PRUEBA 3	Prueba 4	Prueba 5
Sesiones de entrenamiento:	2,3,4,5	1,3,4,5	1,2,4,5	1,2,3,5	1,2,3,4
Sesion de prueba:	1	2	3	4	5
KNN	5	5	5	5	5
NB					
VerdadKKN:	73,33	76,33	53,33	60	60
MentiraKNN	60	80	40	60	66,67
VerdadNB	13,33	13,33	80	13,33	26,67
MentiraNB	86,67	100	46,67	93,33	100
PromedioKNN	66,66	76,66	46,67	60	63,33
PromedioNB	50	56,57	63,33	53,33	63,33

Pruebas "B"	PRUEBA 1	PRUEBA 2	PRUEBA 3	Prueba 4	Prueba 5
Sesiones de entrenamiento:	2,3,4,5	1,3,4,5	1,2,4,5	1,2,3,5	1,2,3,4
Sesion de prueba:	1	2	3	4	5
KNN	5	5	5	5	5
NB					
VerdadKKN:	80	66,67	60	73,33	80
MentiraKNN	73,33	100	80	93,33	86,87
VerdadNB	13,33	13,33	80	13,33	26,67
MentiraNB	86,67	100	46,67	93,33	100
PromedioKNN	76,67	83,33	70	83,33	83,33
PromedioNB	50	56,37	63,33	53,33	63,33

Pruebas "C"	PRUEBA 1	PRUEBA 2	PRUEBA 3	Prueba 4	Prueba 5
Sesiones de entrenamiento:	1,3,4,5	2,3,4,5	1,2,4,5	1,2,3,5	1,2,3,4
Sesion de prueba:	2	1	3	4	5
KNN	5	5	5	5	5
NB					
VerdadKKN:	86,67	40	60	53,33	60
MentiraKNN	46,67	33,33	53,33	40	53,33
VerdadNB	14,33	20	56,67	13,33	20
MentiraNB	100	73,33	86,67	93,33	100
PromedioKNN	66,67	36,67	46,67	46,67	56,67
PromedioNB	57,165	46,67	66,67	53,33	60

Pruebas "D"	PRUEBA 1	PRUEBA 2	PRUEBA 3	Prueba 4	Prueba 5
Sesiones de entrenamiento:	1,3,4,5	2,3,4,5	1,2,4,5	1,2,3,5	1,2,3,4
Sesion de prueba:	2	1	3	4	5
KNN	5	5	5	5	5
NB					
VerdadKKN:	66,67	80	60	73,33	66,67
MentiraKNN	93,33	73,33	73,33	93,33	73,33
VerdadNB	14,33	20	86,67	13,33	20
MentiraNB	100	73,33	46,67	93,33	100
PromedioKNN	80	76,67	66,67	80	70
PromedioNB	57,165	46,67	66,67	53,33	60

Figura 2: Resultados cambiando las sesiones.Primer conjunto de pruebas.

Se observa en la figura 2, los resultados intercambiando las sesiones de entrenamiento y de prueba.Se observa algunos cuadritos rellenos de rojo, estos quieren decir que su tasa de aciertos fue menor que el 50 %.

- Cuadro A: Resultado sin normalizar los datos y con 36 características.
- Cuadro B: Resultado con los datos normalizados y con 36 características.
- Cuadro C: Resultado sin normalizar los datos y con 39 características.
- Cuadro D: Resultado con los datos normalizados y con 39 características.

Pruebas "E"	PRUEBA 1	PRUEBA 2	PRUEBA 3	PRUEBA 4	PRUEBA5
Sujeto de entrenamiento	3	1	15	7	8
Sujeto de prueba	[1;2]&[4;15]	[2;15]	[1;14]	[1;6]&[8;15]	[1;7]&[9;15]
KNN	7	7	7	7	7
VerdadKKN:	60	20	40	0	20
MentiraKNN	60	20	20	80	40
VerdadNB	20	0	0	0	0
MentiraNB	60	80	80	100	800
PromedioKNN	60	20	30	40	30
PromedioNB	40	20	40	50	50

Pruebas "F"	PRUEBA 1	PRUEBA 2	PRUEBA 3	PRUEBA 4	PRUEBA5
Sujeto de entrenamiento	3	15	8	7	1
Sujeto de prueba	[1;2]&[4;15]	[1;14]	[1;7]&[9;15]	[1;6]&[8;15]	[2;15]
KNN	7	7	7	7	7
VerdadKKN:	20	80	0	100	100
MentiraKNN	80	60	100	100	0
VerdadNB	20	0	0	0	0
MentiraNB	60	80	100	100	80
PromedioKNN	50	70	50	100	50
PromedioNB	40	40	50	50	40

Pruebas "G"	PRUEBA 1	PRUEBA 2	PRUEBA 3	PRUEBA 4	PRUEBA5
Sujeto de entrenamiento	3	1	15	7	8
Sujeto de prueba	[1;2]&[4;15]	[2;15]	[1;14]	[1;6]&[8;15]	[1;7]&[9;15]
KNN	7	7	7	7	7
VerdadKKN:	60	20	40	0	20
MentiraKNN	60	20	20	80	40
VerdadNB	20	0	0	0	0
MentiraNB	60	80	80	100	80
PromedioKNN	60	20	30	40	30
PromedioNB	40	20	40	50	50

Pruebas "H"	PRUEBA 1	PRUEBA 2	PRUEBA 3	PRUEBA 4	PRUEBA5
Sujeto de entrenamiento	3	15	8	7	1
Sujeto de prueba	[1;2]&[4;15]	[1;14]	[1;7]&[9;15]	[1;6]&[8;15]	[2;15]
KNN	7	7	7	7	7
VerdadKKN:	20	80	0	100	100
MentiraKNN	60	60	100	100	0
VerdadNB	20	0	0	0	0
MentiraNB	60	80	100	100	80
PromedioKNN	40	70	50	100	50
PromedioNB	40	40	50	50	40

Figura 3: Resultados intercambiando los sujetos.Segundo conjunto de pruebas

Se observa en la figura 3, los resultados intercambiando sujetos de prueba y sujetos de entrenamiento.Se observa algunos cuadritos rellenos de rojo, estos quieren decir que su tasa de aciertos fue menor que el 50 %.

• Los Cuadros E, F, G y H son iguales a los cuadros A, B, C y D respectivamente.

• Prueba con 36 características de cada banda VS Prueba con 36 características de cada banda + 3 características de la señal completa.

• Se observa que al normalizar los valores, el rendimiento del detector incrementa notablemente. Esto se debe a que los valores normalizados fluctúan menos, variando solo entre 0 y 1, lo que proporciona mejores datos para que el clasificador realice las comparaciones. Al comparar los resultados con 36 y 39 características, no hay una gran diferencia en los resultados; sin embargo, el análisis con 36 características muestra un ligero incremento en el rendimiento. La tasa de rendimiento con 36 características es de un 79.332 %, mientras que con 39 características es de un 74.668 %. Sin normalizar, ambas tasas 62,664 % y 50,67 % respectivamente

• Rotando los sujetos: Aquí como en el punto anterior vemos como la normalización fue de una ayuda notable. Utilizando tanto 36 como 39 las características no se ven modificadas. También se ve una alta fluctuación en la tasa de aciertos en algunos sujetos, y esto es lógico, ya que no todos los sujetos actuaran igual ni reaccionara igual su mente a la hora de mentir, por lo que esta parte del estudio (rotando sujetos) es quizás mas difícil de analizar. Como en la base de datos directamente el sujeto de prueba no esta, se hace mas complicada la predicción. Utilizando quizás mas señales de otros electrodos y teniendo una base mas amplia a la hora del entrenamiento, seguramente la la tasa de acierto de lo desarrollado tendera a incrementarse. a tasa de rendimiento con 36 características es de un 62 %, mientras que con 39 características es de un 62 %. Sin normalizar, ambas tasas 36 % y 36 % respectivamente

Rot.Sesiones	Sin normalizar	Normalizado
Acierto con 36 caract.	36%	62%
Acierto con 39 caract.	36%	62%

Rot.Sujetos	Sin normalizar	Normalizado
Acierto con 36 caract.	62,66%	79,33%
Acierto con 39 caract.	50,67%	74,67%

Figura 4: Tasa de acierto.

Algo importante que es valioso recalcar es que los porcentajes mencionados en los párrafos anteriores fueron calculados sin utilizar los valores que calculamos para el clasificador Naive Bayes, ya que los mismos son bajos y distorcionarían los valores totales. Creemos que no es buen clasificador o el mismo no se implemento correctamente.

4. Conclusiones

Este trabajo ha explorado el desarrollo y la implementación de un detector de mentiras utilizando señales EEG, centrado en la comparación entre dos clasificadores: k vecinos más cercanos (KNN) y Naive Bayes. Se ha realizado el procesamiento de señales para extraer características relevantes en nueve bandas en el espectro fecruencial. Los resultados mostraron que la normalización de los datos mejoró significativamente la precisión de los clasificadores, alcanzando un rendimiento del 79.33 con 36 características. Sin embargo, al añadir tres características adicionales, la precisión disminuyó ligeramente. En pruebas rotando sujetos o sesiones en los entrenamiento se observó una alta variabilidad en la precisión, lo cual refleja la complejidad inherente de la detección de mentiras basada en respuestas neuronales individuales.

5. Agradecimientos

Agradecemos L. Di Persia por las constantes revisiones, recomendaciones y correcciones a lo largo de todo el desarrollo de este trabajo. Fueron de mucho ayuda para enriquecer y refinar dicho informe.

Referencias

[1] Ying-Fang Lai, Mu-Yen Chen, and Hsiu-Sen Chiang. Constructing the lie detection system with fuzzy reasoning approach. *Granular Computing*, 3:169–176, 2018.

[2] Junfeng Gao, Hongjun Tian, Yong Yang, Xiaolin Yu, Chenhong Li, and Nini Rao. A novel algorithm to enhance p300 in single trials: Application to lie detection using f-score and svm. *Plos one*, 9(11):e109700, 2014.

[3] Ayahito Ito, Nobuhito Abe, Toshikatsu Fujii, Aya Ueno, Yuta Koseki, Ryusaku Hashimoto, Shunji Mugikura, Shoki Takahashi, and Etsuro Mori. The role of the dorsolateral prefrontal cortex in deception when remembering neutral and emotional events. *Neuroscience research*, 69(2):121–128, 2011.