ISSN 1870-4069

Análisis exploratorio para la caracterización de la adicción a la cocaína a través del aprendizaje computacional

Arturo Téllez-Velázquez [1], Eduardo A. Garza-Villareal [2], Jorge J. González-Olvera [2], Raúl Cruz-Barbosa [3]

[1] CONACyT – Universidad Tecnológica de la Mixteca, Instituto de Computación, Huajuapan de León, Oaxaca, México

[2] Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Subdirección de Investigaciones Clínicas, Ciudad de México, México

[3] Universidad Tecnológica de la Mixteca, Instituto de Computación, Huajuapan de León, Oaxaca, México

atellezv@mixteco.utm.mx, {egarza, jjgonz}@imp.edu.mx, rcruz@mixteco.utm.mx

Resumen. Las adicciones son trastornos neuro-psiquiátricos con serias repercusiones en la salud. En México, la cocaína es la segunda droga más usada después de la marihuana. Sin embargo, su uso conlleva a mayor adicción, síndrome de abstinencia y psicopatología. Este estudio tiene como objetivo seleccionar, usando aprendizaje computacional, las variables más representativas de entre las variables demográficas, cognitivas y de impulsividad para predecir la adicción a la cocaína con nuevas muestras. Para lograrlo, se obtuvieron datos de 39 pacientes con adicción a la cocaína y 23 controles sanos. A partir de esto, se obtuvo un 88.24% de exactitud de clasificación, usando el método de selección Relief con tan sólo 28 predictores; mientras que usando todos los descriptores iniciales (61 características) el rendimiento obtenido es bajo. Nuestros resultados sugieren que la selección de un subconjunto características es muy importante, no sólo para reducir el cómputo durante el entrenamiento de los métodos, sino también para obtener resultados de clasificación de adicción a cocaína mejores que los obtenidos al utilizar el conjunto completo de características.

Palabras clave: aprendizaje computacional, adicción, cocaína, impulsividad, selección de características.

Exploratory Analysis for the Characterization of Cocaine Addiction through Machine Learning

Abstract. Addictions are neuro-psychiatric disorders with serious repercussions on health. In Mexico, cocaine is the second most commonly used drug after marijuana. However, its use leads to increased addiction, withdrawal syndrome and psychopathology. This study aims to select, using computational learning, the most representative variables among demographic, cognitive and impulsivity

to predict cocaine addiction with new samples. To achieve this, data were obtained from 39 patients with cocaine addiction and 23 healthy controls. From this, an accuracy of 88.24% of classification was obtained, using the Relief selection method with only 28 predictors; while using all the initial descriptors (61 characteristics) the accuracy obtained is low. Our results suggest that selecting a subset of characteristics is very important, not only to reduce computation during training methods, but also to obtain better results from cocaine addiction classification than those obtained when using the complete set of features.

Keywords: machine learning, feature selection, addiction classification, cocaine, impulsivity.

1. Introducción

En México, los trastornos mentales causados por las adicciones se han venido incrementando en los últimos años, por lo que el gobierno ha destinado más de 6 mil 600 millones de pesos (2007-2012) de los recursos públicos para atender a la gente con este tipo de problemas [16]. La necesidad de caracterizar las enfermedades asociadas a las adicciones es una tarea primordial para entender el problema y darle el seguimiento con terapias e intervenciones adecuadas.

En este sentido, existen diversos estudios que ayudan a los médicos a determinar qué variables intervienen en la caracterización de las adicciones. Inclusive existen problemas psiquiátricos que están correlacionados y que son indicadores que permiten entender de mejor forma el problema. Por ejemplo, la impulsividad es un problema psiquiátrico en la mayoría los trastornos del uso de substancias (SUD, por sus siglas en inglés *Substance Use Disorder*) y sobre todo en la adicción a la cocaína [9].

El presente artículo presenta una perspectiva desde el punto de vista de aprendizaje computacional, donde se trata el problema de clasificación binaria de la adicción a la cocaína a través de métodos de selección de características.

Los selectores de características, frecuentemente usados en aprendizaje computacional, ayudan a jerarquizar las variables más importantes para resolver problemas de clasificación. Para lograr este objetivo en nuestro estudio, se recolectó información demográfica, cognitiva y de impulsividad de pacientes con adicción a la cocaína y controles sanos en las instalaciones del *Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz* (INPRF). Esta información (que no está disponible públicamente) fue analizada mediante el uso de selectores de características para obtener un conjunto de características reducido y representativo, con el cual es posible caracterizar las adicciones desde el punto de vista de clasificación de patrones [5]. Con esto, también es posible reducir el costo computacional que conlleva el entrenamiento de los modelos de clasificación.

Los resultados obtenidos muestran que los selectores de características ayudan a incrementar el rendimiento de clasificación de adicción a la cocaína de manera sustancial y ayudan a eliminar información redundante contenida en las variables del registro de cada paciente.

Este artículo está organizado en las siguientes secciones. La sección 2 presenta un breve estado del arte del problema de clasificación y caracterización del problema de

adicción a la cocaína. La sección 3 presenta la metodología utilizada para afrontar el problema de clasificación mediante el uso de selectores de características. La sección 4 presenta los resultados obtenidos con base en la metodología implementada y finalmente la sección 5 destaca las conclusiones y el trabajo futuro.

2. Trabajo relacionado

Existen varios trabajos relacionados donde se utiliza el aprendizaje computacional para afrontar el problema de clasificación a las adicciones. Dada la complejidad de este problema debido a que existen problemas psiquiátricos relacionados, la caracterización de las adicciones se ha tenido que analizar de forma indirecta.

Una característica notable en los adictos a la cocaína es la impulsividad. La impulsividad se ha utilizado como una medida indirecta de la adicción a la cocaína (CA) [1]. Para analizar la impulsividad, en [11] se propuso la *Escala de Impulsividad de Barrat* (BIS) como un instrumento que mide el comportamiento impulsivo de las personas (bajo ciertas circunstancias) para entender su relación con otras comorbilidades clínicas. La escala más utilizada es la BIS-11 [7, 13].

Otra prueba relacionada que ha sido usada para caracterizar la adicción a la cocaína es la *Prueba Eriksen-Flanker* (FT), la cual es usada para probar la habilidad de las personas para suprimir respuestas inapropiadas en un contexto particular [6]. Principalmente, la prueba FT estimula al paciente con el objetivo de medir la cantidad de respuestas incongruentes. También es posible con esta prueba medir el tiempo entre el estímulo y la respuesta. Trabajos recientes, que usan aprendizaje computacional [12], utilizan los resultados de la prueba FT como predictores para determinar la relación entre los SUD y las respuestas incongruentes.

Por otro lado, también existe la *Prueba de Iowa-Gambling* (IGT) [4], que es un instrumento frecuentemente usado para determinar el daño que puede existir en la corteza prefrontal en pacientes con SUD, lo cual dificulta la toma de decisiones debido a la insensibilidad a las consecuencias futuras en sus acciones [3]. Los resultados de esta prueba también han sido usados como predictores en trabajos recientes; por ejemplo, en [2] se ha utilizado para caracterizar los trastornos debidos al consumo de heroína y anfetaminas usando técnicas de aprendizaje computacional.

Otro factor que se toma en cuenta, desde el campo de aprendizaje computacional, es que la selección de características, hasta cierto punto, ayuda a mejorar las tareas de clasificación al facilitar la explicación del problema, al usar un número inferior de variables [14]. Con lo que respecta a la selección de características en el ámbito de la caracterización de las adicciones, investigaciones destacadas [8, 10, 15, 17] obtienen una mejora en rendimiento de clasificación cuando se eliminan variables que no aportan un valor significativo a las predicciones.

3. Metodología

A diferencia de los trabajos relacionados descritos en la sección anterior, nuestra investigación realiza una selección de características para la clasificación de la adicción a la cocaína, misma que sirve para identificar las variables que son más representativas.

Para lograr esto, se recolectó información demográfica, de pruebas cognitivas y de impulsividad de 62 pacientes en las instalaciones del INPRF, donde 39 pacientes son adictos a la cocaína y 23 son controles. En total se utilizaron 61 descriptores / características: 8 demográficas, 4 de impulsividad y 49 cognitivas. Por otro lado, se utilizó una muestra representativa de 17 pacientes para prueba y 45 pacientes para el entrenamiento de los algoritmos de clasificación.

Diversas técnicas de clasificación fueron utilizadas en la configuración de los experimentos y los selectores de características utilizados (implementados en MATLAB) fueron los siguientes:

- Prueba Xi-cuadrada (CHI),
- Selección secuencial hacia adelante (SFS),
- Algoritmo Relief (RLFF) y
- Análisis de componentes principales (PCA).

Para fines de comparación, se utilizan todas las características (ALL) como entrada a los algoritmos de clasificación, con la finalidad de obtener un rendimiento de referencia, con respecto al rendimiento obtenido a partir de un subconjunto de estas.

Tabla 1. Rendimientos promedio obtenidos de los clasificadores usando todas las variables demográficas, cognitivas y de impulsividad.

#	NB			DT			SVM		
caract.	ACC.	SENS	SPEC	ACC.	SENS	SPEC	ACC.	SENS	SPEC
61	0.52941	0.45455	0.66667	0.41176	0.54545	0.16667	0.64706	0.63636	0.66667

Tabla 2. Matrices de confusión correspondientes a la Tabla 1 usando todas las características.

NB			DT			SVM		
C	0	1	C	0	1	C	0	1
0	4	2	0	1	5	0	4	2
1	6	5	1	5	6	1	4	7

Los tres métodos de selección enunciados arriba, y el método de extracción PCA permiten obtener un número finito de características menor que el número de variables originales. Para obtener el mínimo número adecuado de características se ha utilizado un algoritmo de clasificación para encontrar el desempeño máximo de cada algoritmo. La idea es simplemente obtener el mejor rendimiento con el menor número de variables posible.

Una vez obtenido el número adecuado de características, se procede a la evaluación del algoritmo de clasificación adecuado. En este experimento se utilizaron los siguientes algoritmos:

- Bayes ingenuo (NB),
- Árboles de decisión (DT) y
- Máquinas de vectores de soporte lineal (SVM).

Debido a la cantidad limitada de datos de entrenamiento, los tres modelos mencionados arriba utilizan validación cruzada *leave-one-out* para estimar su rendimiento.

Por otro lado, las medidas de rendimiento utilizadas para comparar los clasificadores usados en este estudio son los siguientes:

- *Exactitud* (ACC derivado de la abreviación en inglés Accuracy). Esta medida de rendimiento nos permite evaluar de manera general el desempeño del clasificador y es utilizado para determinar la capacidad de discriminación entre adictos y controles correctamente.
- *Sensibilidad* (SENS derivado de la abreviación en inglés Sensitivity). Esta medida permite conocer el desempeño del clasificador para clasificar correctamente a las personas adictas.
- *Especificidad* (SPEC derivado de la abreviación en inglés Specificity). Esta otra medida se usa en este estudio para conocer el desempeño del clasificador para clasificar correctamente a los controles.

Por último y para una mejor visualización del rendimiento de los algoritmos de clasificación, también se muestra la *matriz de confusión* de los resultados de cada método. Esta matriz permite observar claramente los errores de clasificación y confirmar los resultados de las medidas anteriores (exactitud, sensibilidad y especificidad).

Tabla 3. Rendimiento promedio de los clasificadores usando selección y extracción de características para la clasificación de adicción a la cocaína.

Mét.	#	NB			DT			SVM		
	car.	ACC.	SENS	SPEC	ACC.	SENS	SPEC	ACC.	SENS	SPEC
CHI	29	0.52941	0.54545	0.50000	0.58824	0.72727	0.33333	0.88235	0.90909	0.83333
SFS	18	0.58824	0.72727	0.33333	0.58824	0.72727	0.33333	0.64706	0.81818	0.33333
RLFF	28	0.70588	0.72727	0.66667	0.47059	0.63636	0.16667	0.88235	0.81818	1.00000
PCA	54	0.41176	0.54545	0.16667	0.35294	0.45455	0.16667	0.82353	0.81818	0.83333

Tabla 4. Matrices de confusión usando las variables seleccionadas de la Tabla 3.

Método	NB			DT			SVM		
	C	0	1	C	0	1	C	0	1
CHI	0	3	3	0	2	4	0	5	1
	1	5	6	1	3	8	1	1	10
	C	0	1	C	0	1	C	0	1
SFS	0	2	4	0	2	4	0	2	4
	1	3	8	1	3	8	1	2	9
	C	0	1	C	0	1	C	0	1
RLFF	0	4	2	0	1	5	0	6	0
	1	3	8	1	4	7	1	2	9
	C	0	1	C	0	1	C	0	1
PCA	0	1	5	0	1	5	0	5	1
	1	5	6	1	6	5	1	2	9

4. Resultados

De acuerdo con la configuración de los experimentos, primero se utilizaron las 61 características (de las que consta nuestra base de datos) con los clasificadores NB, DT y SVM. Esto con el objetivo de obtener una referencia de rendimiento inicial.

El porcentaje promedio de exactitud de clasificación obtenido fue precario en los tres clasificadores (ver Tabla 1) siendo el mejor el obtenido con SVM, alcanzando un porcentaje de 64.71%. Como es de esperarse, en la misma tabla también se observa que se obtuvo un porcentaje bajo tanto para predecir los adictos (sensibilidad) como los controles (especificidad). Para mayor detalle de los resultados de la Tabla 1, se presenta la matriz de confusión correspondiente en la Tabla 2, donde se observa que el desempeño de SVM es similar al de NB, pero este último tiene mayor número de falsos negativos y por ello su sensibilidad es menor que la sensibilidad obtenida con SVM. Como notación, en las Tablas 2 y 4, los encabezados y primera columna de cada matriz de confusión referidos como 0 y 1 representan a las clases *control* y *adicto*, respectivamente.

Con el objetivo de mejorar el desempeño de los clasificadores y reducir el número de características utilizadas durante el entrenamiento, se utilizaron tres selectores y un extractor de características (todos ellos mencionados en la sección anterior).

Tabla 5. Características seleccionadas con RLFF, SFS y CHI, separados por categoría.

RLFF		SFS		CHI	
DEMOGRÁFICA					
SEX		SEX		SEX	
DEGREE		AGE		SCORE	
LATERALITY				DEGREE	
IMPULSIVIDAD					
BIS.INoPl		BIS.INoPl		BIS.INoPl	
BIS.Total		BIS.ICog		BIS.Total	
BIS.IMo				BIS.IMo	
BIS.ICog				BIS.ICog	
COGNITIVAS					
F-AI4	IGT-B1	F-TREC	IGT-B5	F-AI4	IGT-B1
F-TREC	IGT-B3	F-TREI		F-DTRC1	IGT-B3
F-TREI	IGT-B2	F-EI		F-AC4	IGT-B2
F-EI4	IGT-TRD	F-AC4		F-AI1	IGT-DTRD
F-EI	IGT-T	F-TRI4		F-TRC	IGT-T
F-DTRI4	IGT-I	F-AC		F-DTRI	IGT-I
F-DTRC1	IGT-RV	F-NRC		F-EC	IGT-RV
F-NRI1	IGT-RD	F-EC4		F-ANG4	IGT-RD
F-AC4	IGT-TRV	F-AI1		F-NRC1	IGT-B5
	IGT-B5	F-NRI		F-ANG	IGT-B4
	IGT-B4	F-DTREI		F-NCI4	IGT-DTRV
	IGT-DTRD	F-EI1			
		F-TRC			

En las filas de la Tabla 3 se presentan los resultados de los cuatro algoritmos de selección y extracción utilizados. La segunda columna de esta tabla muestra el número mínimo de características obtenidas con cada método y las demás columnas muestran los resultados de rendimiento máximo (de exactitud, sensibilidad y especificidad) obtenidos con los tres clasificadores utilizados.

De la misma manera que en la Tabla 1, el desempeño de NB y DT fue muy pobre con cualquiera de los métodos de selección. Como puede observarse también, casi todos los algoritmos de selección y extracción mejoraron el desempeño de clasificación con SVM, en comparación con el uso de las 61 variables de la Tabla 1. Los métodos CHI y RLFF, mismos que han sido usados exitosamente para selección de características en otros trabajos [17, 18], maximizaron el desempeño utilizando solamente 29 y 28 características, respectivamente.

Las matrices de confusión de los resultados de todos los métodos de selección y extracción, usando el clasificador SVM, se presentan en la Tabla 4. A pesar de que los métodos CHI y RLFF obtienen la misma exactitud de clasificación (ver Tabla 3), el método CHI tiene la capacidad de reconocer mejor a los adictos (concretamente para el conjunto de datos analizado), lo cual puede ser decisivo al momento de escoger entre estos dos métodos de selección.

Por otro lado, el extractor PCA, que genera 54 variables nuevas a partir de las 61, obtiene resultados modestos y porcentajes de sensibilidad y especificidad similares. Sin embargo, su desempeño no supera a RLFF o CHI y las nuevas variables no son interpretables directamente, ya que son generadas a partir de la combinación lineal de las originales.

Por último, la Tabla 5 presenta las variables que fueron seleccionadas por los tres métodos de selección de características analizados. Observe que tanto RLFF como CHI toman en cuenta la mayoría de las variables obtenidas con las pruebas de impulsividad y cognitivas, así como aquéllas que miden la capacidad de toma de decisión de las personas (BIS e IGT, respectivamente).

En todos los métodos de selección analizados, las variables demográficas resultan ser poco relevantes para la caracterización de la adicción a la cocaína. De acuerdo con los resultados del método de selección de ranking RLFF, las variables de impulsividad resultaron ser las más importantes. De esto se desprende que, los pacientes adictos a la cocaína presentan un déficit en la inhibición de las acciones motivadas por las emociones del momento (impulsividad motora, BIS.IMo); en la inhibición al pensamiento apresurado y sin planificación (impulsividad no planeada, BIS.NoPl) y; en la inhibición de la realización de múltiples conjeturas en corto tiempo (impulsividad cognitiva, BIS.ICog). Estos resultados confirman también que algunos factores de impulsividad primarios y secundarios de la escala de Barrat (BIS-11) son predominantes en pacientes adictos a la cocaína.

5. Conclusiones

Los resultados obtenidos de los experimentos en este trabajo muestran que la selección de características es particularmente útil para obtener aquéllas variables psiquiátricas que están fuertemente relacionadas con la adicción a la cocaína y para mejorar el rendimiento en las tareas de clasificación.

Las variables seleccionadas indican que los trastornos que se manifiestan con el incremento de la impulsividad y la incapacidad de la toma de decisiones son relevantes para la caracterización de la adicción a la cocaína.

De manera general, los métodos de selección o extracción de características ayudan a reducir el número de características del registro de cada paciente. Como consecuencia, se puede decrementar el tiempo de cómputo y el uso de memoria asociados al entrenamiento de los modelos de clasificación usados en aprendizaje computacional.

Como trabajo futuro, se buscará enriquecer esta investigación utilizando datos adicionales obtenidos a partir de imágenes de resonancia magnética.

Agradecimientos. Parte de este proyecto fue financiado por CONACYT-Cátedras proyecto No. 1170 y No. 2358948, por CONACYT-FOSISS proyecto No. 0201493 y por la Srita. Ana Teresa Martínez Alanís. Por otra parte, agradecemos a las personas que hicieron posible la parte de adicción de este estudio: Francisco J. Pellicer-Graham, Margarita López-Titla, Aline Leduc, Erik Morelos-Santana, Diego Ángeles, Alely Valencia, Daniela Casillas, Sarael Alcauter, Luis Concha y Bernd Foerster. Agradecemos también a Rocío Estrada-Ordoñez e Isabel Lizarindari Espinosa-Luna de la Unidad de Atención Toxicológica Xochimilco por su apoyo. Finalmente, agradecemos a los participantes por su cooperación y paciencia.

Referencias

1. Ahn, W. Y., Ramesh, D., Moeller, F. G., Vassileva, J.: Utility of Machine-Learning Approaches to Identify Behavioral Markers for Substance Use Disorders: Impulsivity Dimensions as Predictors of Current Cocaine Dependence. Front. Psychiatry, 7(34) (2016)
2. Ahn, W. Y., Vassileva, J.: Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. Drug Alcohol Depend, 161, pp. 247–257 (2016)
3. Bechara, A., Damasio, A. R., Damasio, H., Anderson, S. W.: Insensitivity to future consequences following damage to human prefrontal cortex. Cognition, 50(7), pp. 289–202 (1994)
4. Bechara, A., Tranel, D., Damasio, H.: Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. Brain, 132(7), pp. 289–202 (2000)
5. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification. Wiley Interscience Publication, U.S., 2nd edn. (2001)
6. Eriksen, B. A., Eriksen, C. W.: Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception and Psychophysics, 16(1), pp. 143–149 (1974)
7. Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., Gorban, A. N.: The Five Factor Model of personality and evaluation of drug consumption risk. Tech. rep., Cornell University Library (2017)
8. Mete, M., Sakoglu, U., Spence, J. S., Devous, M. D., Harris, T. S., Adino, B.: Successful classification of cocaine dependence using brain imaging: a generalizable machine learning approach. In: 13th Annual MCBIOS conference, Vol. 17, BioMed Central (2016)
9. Mitchell, M. R., Potenza, M. N.: Addictions and Personality Traits: Impulsivity and Related Constructs. Curr. Behav. Neurosci. Rep., 1(1), pp. 1–12 (2014)
10. Pariyadath, V., Stein, E. A., Ross, T. J.: Machine learning classification of resting state functional connectivity predicts smoking status. Front. Hum. Neurosci., 8(425) (2014)

11. Patton, J. H., Stanford, M. S., Barrat, E. S.: Factor structure of the Barratt impulsiveness scale. J. Clin. Psychol., 51(6), pp. 768–774 (1995)
12. Plewan, T., Wascher, E., Falkenstein, M., Hoffmann, S.: Classifying Response Correctness across Different Task Sets: A Machine Learning Approach. PLoS One, 11(3) (2016)
13. Reise, S. P., Moore, T. M., Sabb, F. W., Brown, A. K., London, E. D.: The Barratt Impulsiveness Scale - 11: Reassessment of its Structure in a Community Sample. Psychol. Assess, 25(2), pp. 631–642 (2013)
14. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, pp. 1157–1182 (2003)
15. Rish, I., Bashivan, P., Cecchi, G. A., Goldstein, R. Z.: Evaluating effects of methylphenidate on brain activity in cocaine addiction: a machine-learning approach. In: Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging, SPIE (2016)
16. Villatoro-Velázquez, J., Medina-Mora, M., Fleiz-Bautista, C., Téllez-Rojo, M. M., Mendoza-Alvarado, L. R., Romero-Martínez, M., Gutiérrez-Reyes, J. P., Castro-Tinoco, M., Hernández-Ávila, M., Tena-Tamayo, C., Alvear-Sevilla, C., Guisa-Cruz, V.: Encuesta Nacional de Adicciones 2011: Reporte de Drogas. INPRF, Tlalpan, Ciudad de México (2011)
17. Jin, X., Xu, A., Bie, R., Guo, P.: Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles. In: Proceedings of Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore, pp. 106–115 (2006)
18. Durgabai, R.: Feature Selection using ReliefF Algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 3(10), pp. 8215–8218 (2014)