Sistema de Reconocimiento de Patrones en Bioinformática

J. Altamiranda[1], J. Aguilar[2] and L. Hernández[3]

[1] Universidad de Los Andes, Facultad de Ingeniería, Postgrado Computación, CEMISID, Mérida, Venezuela
[2] Universidad de Los Andes, Facultad de Ingeniería, Departamento de Computación, CEMISID, Mérida, Venezuela
[3] Universidad de Los Andes, Facultad de Medicina, Laboratorio de Fisiología de la Conducta, Mérida, Venezuela

Abstract— **Data Mining is defined like a set of methods for the extraction of knowledge from large databases. In this work we propose the construction of a System of Data Mining for Systems Biology, whose objective is to identify the patterns of the chemical substances present in the brain of a rodent during the development of a given activity (to sleep, to eat, etc.) The system identifies the classes that represent the chemical substances, and the classes that represent the activities made by the rodents. The performance of the system of Data Mining was tested using an example in which the neurotransmitters Glutamate and Aspartate are studied and the samples obtained are classified.**

Palabras claves—**Systems Biology, Bioinformatics, Data Mining, Artificial Intelligence, Artificial Neural Network.**

I. Introducción

Las investigaciones en ciencias biomédicas están generando un enorme volumen de información biológica cada vez más compleja. Por ello, las herramientas computacionales y la inteligencia artificial son cruciales para almacenar e interpretar estos datos de un modo eficiente y robusto. Así, se origina una nueva disciplina llamada "Biología de Sistemas". Los sistemas inteligentes constituyen el campo de la computación donde se estudian y desarrollan algoritmos que implementan los modelos basados en aprendizaje. Estos están orientados hacia el descubrimiento de patrones o regularidades en estructuras de información. Una técnica desarrollada para tal fin es la Minería de Datos y la Búsqueda de Conocimientos en Base de Datos [1, 2, 3]. En este trabajo se plantea construir un Sistema de Minería de Datos para la extracción de conocimiento a partir de grandes volúmenes de datos que se obtienen de experimentos que se realizan a roedores, para determinar las sustancias químicas que actúan en el cerebro y entender las interacciones que suceden en él cuando un roedor realiza una actividad específica. En nuestro caso vamos a estudiar los neurotransmisores Glutámato y Aspártato. Actualmente todos los datos obtenidos en los experimentos son almacenados en una base de datos para ser analizados en forma manual y poder así generar conocimiento. De allí viene la necesidad de la minería de datos para el análisis e interpretación de los datos.

II. Diseño del sistema de mineria de datos para iologia e sistemas

El sistema de Minería de Datos permite la automatización de los procesos de clasificación y análisis de los picos de la gráfica que representan las sustancias químicas presentes en el cerebro de un roedor, y la clasificación de estas muestras (ver figura 1).

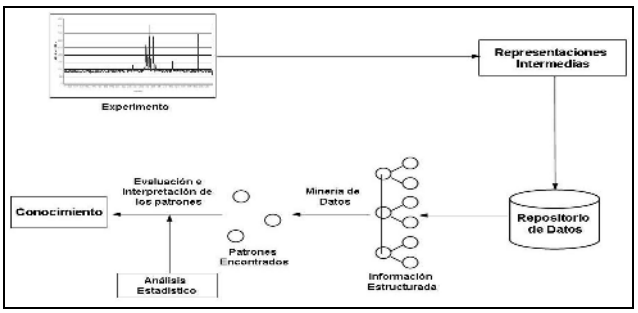


Fig. 1 Diseño del sistema de Minería de Datos Propuesto.

Nuestro sistema de Minería de Datos está dividido en dos etapas: una etapa de pre –procesamiento y una etapa de post – procesamiento (ver figura 2).

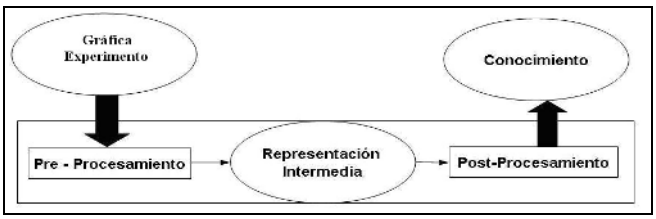


Fig. 2 Proceso de Minería de Datos en el sistema propuesto.

En la etapa de pre – procesamiento los datos se transforman en una representación intermedia que facilite su posterior análisis. Esta representación esta compuesta para cada pico de la gráfica por: área, altura, ancho, punto de inicio pico y punto de fin. En la etapa de post – procesamiento las representaciones intermedias se analizan para clasificarlas en sustancias químicas y realizarles

Carmen Mueller-Karger, Sara Wong, Alexandra La Cruz (Eds.): CLAIB 2007, IFMBE Proceedings 18, pp. 573–577, 2007
www.springerlink.com

análisis estadísticos. Además, se toma el patrón de la gráfica y se clasifica en una actividad específica. El sistema de Minería de Datos esta formado por tres componentes: Una representación intermedia, un repositorio de datos, un algoritmo que realice el reconocimiento y construcción de patrones.

A. Representación Intermedia

Los métodos usados para realizar esta tarea son:

- Filtro Savitzky-Golay: El objetivo del filtro es sustituir los datos originales para suavizar la señal original. Este producirá una señal más suave que la señal original, pero con el mismo número de puntos (ver figura 3).

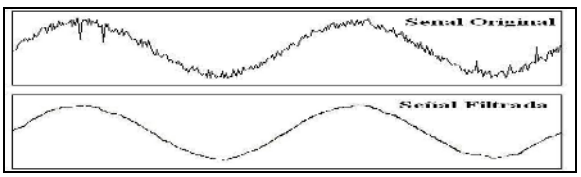


Fig. 3 Filtro de Savitzky – Golay aplicado a una señal.

- Algoritmo para la extracción de los picos de la muestra: éste permite dividir la gráfica en sectores por medio de una barra vertical, donde cada uno de ellos representa el comienzo y fin de un pico. Estos representan una sustancia química en ese instante de tiempo (ver figura 4).

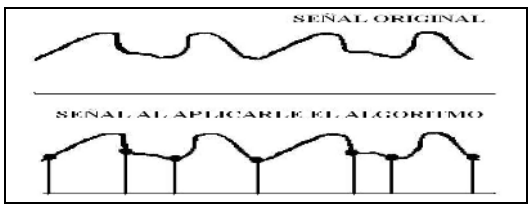


Fig. 4 Algoritmo de extracción de picos aplicado a una señal.

Después de aplicar estos métodos se obtiene la representación intermedia de los picos. Esta debe contener las características: punto inicio, punto fin, ancho, altura, área para cada uno de los picos de las muestras analizadas.

B. Repositorio de datos

Las representaciones intermedias obtenidas de los picos de las muestras son almacenadas en un repositorio de datos. La estructura de datos es mostrada en la figura 5.

Fig. 5 Estructura del repositorio de datos.

C. Minería de datos

Es la etapa central del sistema. En ella se realizan varias tareas que permiten identificar distintos tipos de patrones en un conjunto de datos para poder extraer conocimiento. Así, se requiere automatizar el proceso de identificación de los picos en las muestras y la clasificación de éstas. Nosotros vamos a utilizar técnicas de inteligencia artificial para esto. Nuestro Sistema de Minería de Datos consta de dos redes neuronales ART2 (Adaptive Resonance Theory) [3, 4, 5].

Red ART2 para clasificar los picos de una muestra según su posición en la gráfica: esta permitirá la clasificación y reconocimiento de los picos en la muestra. Así, un patrón es una curva que representa una sustancia química y sus características (área, altura, punto inicio, punto de fin, ancho). La red neuronal ART2 utilizará como entrada: punto de inicio y punto de fin de cada pico, y se obtiene como salida las clases a las que pertenece el pico (ver figura 6).

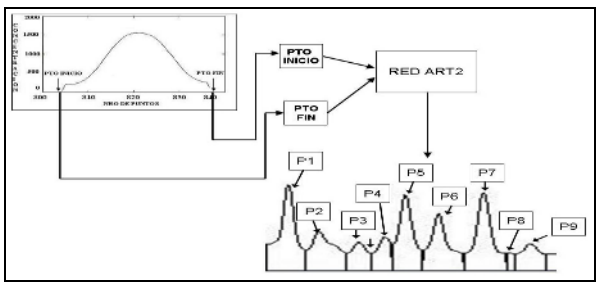


Fig. 6 Diagrama de la red neuronal ART2
para el reconocimiento de los picos de la muestra.

En la Figura 6, P1, P2, P3, P4, P5, P6, P7, P8, P9 representan la salida de la red neuronal ART2 y, por lo tanto, las clases a los que pertenecen los picos. Cada una de ellas representa una sustancia química diferente presente en la muestra.

Red ART2 para clasificar una actividad determinada de un roedor en un patrón: esto permite al sistema el descubrimiento de patrones de actividades del cerebro del roedor de manera automatizada. Se utilizará una red neuronal ART2 que usará como entrada una cadena de clases (sustancias químicas) que van a representar el conjunto de picos presentes en la muestra a ser clasificada, y como salida va a estar la actividad a la que pertenece (ver figura 7). La clasificación de una muestra en una clase de actividad sigue los siguientes pasos: se tiene una muestra extraída del cerebro de un roedor, se clasifican los picos de la muestra utilizando la red neuronal ART2 del punto anterior, se construye una cadena de clases de picos, esta cadena es la entrada a la red neuronal ART2 de esta sección, la salida de la red neuronal ART2 es una clase de comportamiento de la muestra. Aunque el número de picos en las muestras no son constantes, se necesita tener un número de entradas fijas para la red neuronal ART2. Es imposible cambiar dinámicamente el número de neuronas, la red neuronal ART2 esta formada por un número n máximo de neuronas de entrada y neuronas de salida que representan las clases de actividades del roedor (ver figura 7). Si en la muestra la cantidad de picos es menor de n, entonces la cadena se completa con cero (ausencia de clases) en el lugar de la clase. Para la clasificación de las muestras se van a tener clases desde 1 hasta k, nuestro sistema podrá reconocer k actividades distintas.

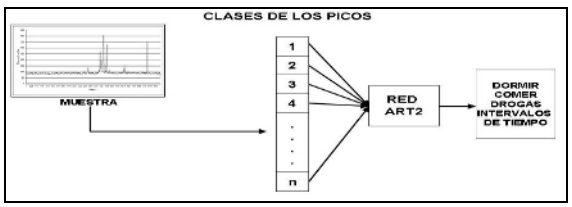


Fig. 7 Diagrama de la red ART2 para el reconocimiento de la muestra.

III. Caso de estudio

En el Departamento de Fisiología, en la Facultad de Medicina de la Universidad de los Andes se realizan experimentos para determinar cambios bioquímicos en el cerebro de roedores, con el fin de entender las interacciones que suceden en él a través de electroforesis capilar [6, 7]. En nuestro caso vamos a estudiar los neurotransmisores Glutámato y Aspártato, analizando las muestras extraídas del cerebro de roedores. En la figura 8 podemos observar el panel principal del sistema de Minería de Datos. La muestra es filtrada para luego poder obtener la representación intermedia de ella.

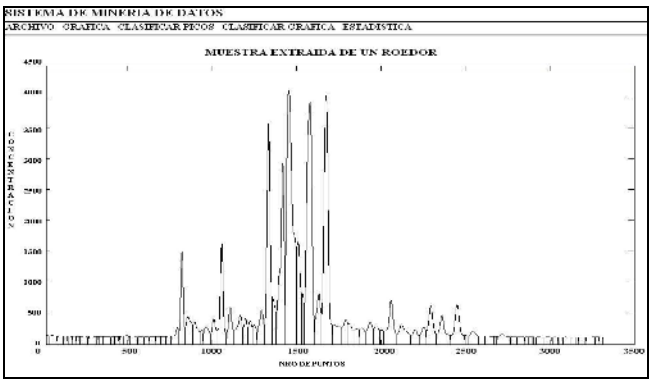
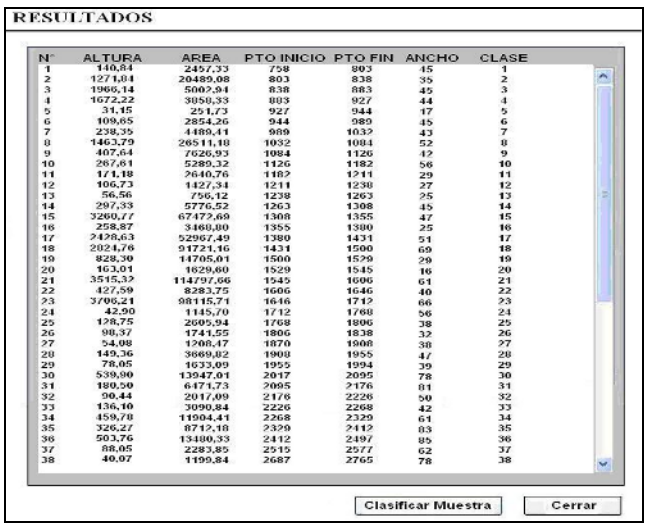


Fig. 8 Panel Principal y filtrado de una muestra.

El sistema de Minería de Datos clasificará los picos a la clase correspondiente (ver figura 9).



Nº	ALTURA	AREA	PTO INICIO	PTO FIN	ANCHO	CLASE
1	110,84	2457,33	758	803	45	1
2	1271,04	20489,08	803	838	35	2
3	1966,14	5002,04	838	883	45	3
4	1672,22	3058,33	883	927	44	4
5	31,15	251,73	927	944	17	5
6	109,65	2854,26	944	989	45	6
7	238,35	4489,41	989	1032	43	7
8	1463,79	26511,18	1032	1084	52	8
9	407,64	7626,93	1084	1126	42	9
10	267,61	5289,32	1126	1182	56	10
11	171,18	2640,76	1182	1211	29	11
12	106,73	1427,34	1211	1238	27	12
13	56,56	756,12	1238	1263	25	13
14	297,33	5776,52	1263	1308	45	14
15	3260,77	67477,69	1308	1355	47	15
16	258,87	3460,80	1355	1380	25	16
17	2428,63	52967,49	1380	1431	51	17
18	2024,76	91721,16	1431	1500	69	18
19	828,30	14705,01	1500	1529	29	19
20	163,01	1829,80	1529	1545	16	20
21	3515,32	114797,66	1545	1606	61	21
22	427,59	8283,75	1606	1646	40	22
23	3706,21	98115,71	1646	1712	66	23
24	42,90	1145,70	1712	1768	56	24
25	128,75	2605,94	1768	1806	38	25
26	98,37	1741,55	1806	1838	32	26
27	54,08	1208,47	1870	1908	38	27
28	149,36	3869,82	1908	1955	47	28
29	78,05	1633,09	1955	1994	39	29
30	539,90	13947,01	2017	2095	78	30
31	180,50	6471,73	2095	2176	81	31
32	90,44	2017,09	2176	2226	50	32
33	136,10	3090,84	2226	2268	42	33
34	459,78	11904,41	2268	2329	61	34
35	326,27	8712,18	2329	2412	83	35
36	503,76	13480,33	2412	2497	85	36
37	88,05	2283,85	2515	2577	62	37
38	40,07	1199,84	2687	2765	78	38

Fig. 9 Resultados de la clasificación de los picos de la primera muestra.

El sistema de Minería de Datos va a clasificar los picos por números desde 1 hasta n. Esta nomenclatura se utiliza porque los picos están asociados a sustancias químicas que han sido reconocidas y es una manera fácil de representar la clasificación. Cada número representa una sustancia química diferente. Por otro lado, nuestro sistema de Minería de Datos permite asociar a las muestras una actividad. Al igual que los picos utilizaremos la nomenclatura 1 hasta m para clasificar a dichas actividades (ver figura 9).

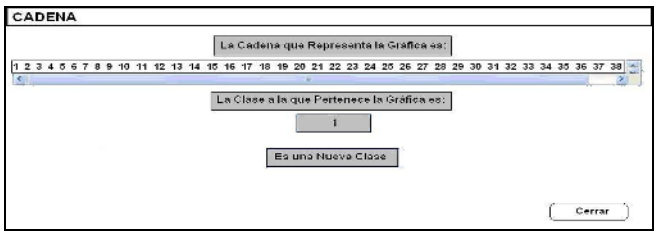


Fig. 9 Clasificación de la primera muestra.

Al observar la figura 10 los neurotransmisores Glutámato y Aspártato aparecen según la apreciación de los expertos. Estos representan los picos 34 y 36 de la clasificación realizada por nuestro Sistema de Minería de Datos para la muestra.

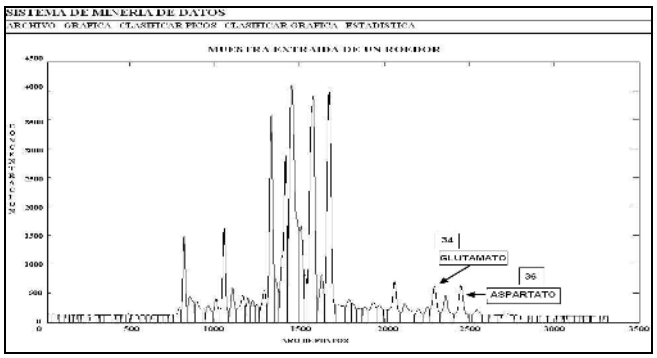


Fig. 10 Localización del Glutámato y Aspártato en la muestra.

Podemos buscar los picos de Glutámato y Aspartato. Nuestro sistema de Minería de Datos dará los siguientes resultados (Ver Figura 11).

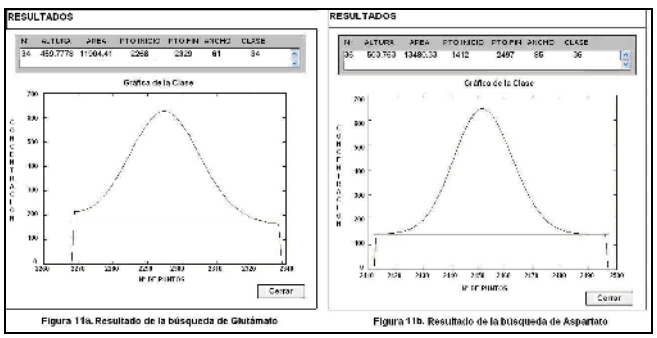


Fig. 11 Localización del Glutámato y Aspártato en la muestra.

Nuestro sistema de Minería de Datos compara picos de muestras diferentes. En nuestro caso, vamos a comparar en la primera y segunda muestra los neurotransmisores Glutámato y Aspártato (ver figura 12).

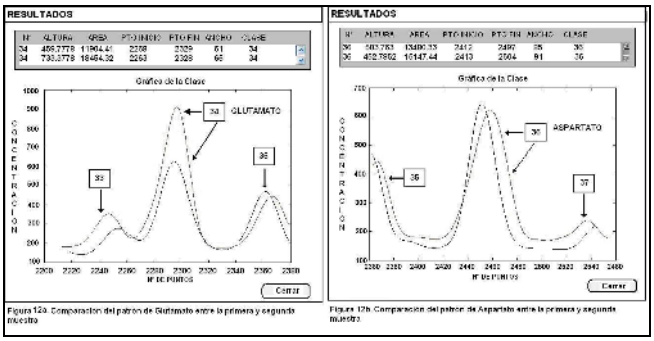


Fig. 12 Comparación del Glutámato y Aspártato en 2 muestras.

Nuestro sistema de Minería de Datos realiza un análisis estadístico para observar las tendencias de los picos de las sustancias químicas en diferentes muestras, usando como variables la altura y el área. Para el Glutámato tenemos: (ver figura 13).

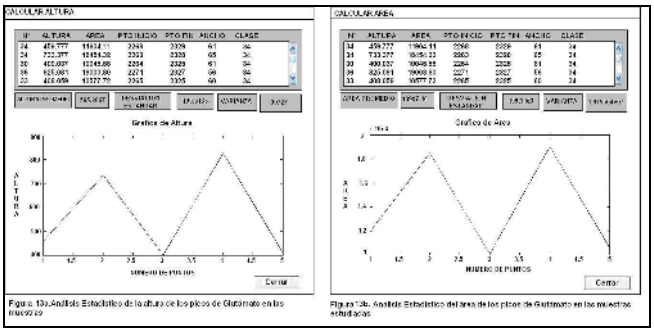


Fig. 13 Análisis Estadístico
de la altura y el área de los picos de Glutámato.

Así, los expertos pueden llegar a conclusiones del funcionamiento del cerebro del roedor. Por ejemplo, según estos resultados los valores de la altura y del área del Glutámato varían de una muestra a otra, lo cual puede ser originado por la presencia de drogas u otra sustancia administrada por los expertos.

IV. CONCLUSIONES

La extracción de patrones y conocimiento desde fuentes de información es posible porque existen herramientas computacionales para lograrlo. El proceso del sistema de Minería de Datos consiste en usar métodos para extraer

conocimiento de forma natural por medio de patrones que representan sustancias químicas extraídas del cerebro de roedores y luego clasificar la muestra en una actividad. En el caso de estudio, se analizaron los neurotransmisores Glutámato y Aspártato lográndose su reconocimiento en las muestras estudiadas. Desde el punto de vista de la Minería de Datos el sistema logra resultados consistentes, sin ningún conocimiento de fondo. El objetivo de este trabajo es incorporar el conocimiento obtenido por medio del Sistema de Minería de Datos a los experimentos realizados en la Facultad de Medicina de la Universidad de Los Andes.

Referencias

1. Agrawal R., Shafer J., (1996) Data Mining & Knowledge Discovery in Databases (KDD). IEEE Transactions on Knowledge and Data Engineering.
2. Febles J., González A., (2002) Aplicación de la minería de datos en la bioinformática, ACIMED, v.10 n.2, pp. 69 – 76
3. Aguilar J., Rivas F. (Ed.), (2001) Introducción a la Computación Inteligente, MERITEC, Venezuela, (2001).
4. Higuera J., Matinez V. (1995) Redes Neuronales Artificiales: Fundamentos, Modelos y Aplicaciones. Addison – Wesley.
5. Carpenter G., Grossberg S., (1988) The ART of Adaptative Resonance Theory by a Self - Organizing Neural Network IEEE Computer, 21(3) pp 77 – 88
6. Hernández L., (2004) Manual de CZE, Universidad de Los Andes, Facultad de Medicina, Departamento de Fisiología, Mérida – Venezuela, (2004). Technical Report
7. León V., (1998) Manual de usuario del ONICE, Universidad de los Andes, Facultad de Medicina, Departamento de Fisiología, Mérida – Venezuela, Technical Report

Autor: Junior Amilcar Altamiranda Pérez
Instituto: Universidad de Los Andes, Postgrado Computación, CEMISID
Calle: Núcleo la Hechicera Facultad de Ingeniería 3er piso Ala sur
Ciudad: Mérida
País: Venezuela
E-mail: altamira@ula.ve