Data Mining System for Biochemical Analysis in Experimental Physiology

Junior Altamiranda[1], Jose Aguilar[1], Luís Hernández[2]

Abstract— **We develop a Data Mining system to assist with the elucidation by graphical means of the biochemical changes in the brains of rodents. Manual analysis of such experiments is increasingly impractical because of the voluminous nature of the data that is generated, and the tedious nature of the analysis means that important information can be missed. For this purpose we are constructing an increasingly sophisticated Data Mining system which contains a number of pre-processing stages and classification via two steps of an Adaptive Resonance Theory Artificial Neural Network. In this paper we describe the system. The focus of our activity is the study of neurotransmitters: Glutamate and Aspartate and we present an example of how to utilize our Data Mining system for the automated classification of samples that are extracted from the brains of rodents. This methodology should prove equally valuable to other biochemical analysis problems in experimental Physiology.**

I. Problem Description

Experiments carried out in the department of Physiology of the Faculty of Medicine at the Universidad de Los Andes cause biochemical changes in the brains of rats, and the experimental analysis requires extracting chemical substances by means of capillary electrophoresis. These experiments are made with rodents to determine biochemical changes in their brain. The chemical substances that are acted upon in these experiments result in certain response chemical interactions that must be elucidated and understood by means of graphical interpretation.

Currently, the data that this provides is being stored in a database for its subsequent interpretation. The sheer volume of this data, its complexity, and its time consuming manual analysis will invariably cause some part of the information that resides in these experiments to be missed out of the analysis [1, 2]. This important problem has motivated us to develop an automatic method of data analysis to assist with this interpretative task.

A system of Data Mining is presented by the paper for the extraction of knowledge from great volumes of data that are obtained from experiments that are made to rodents to determine biochemical changes in their brain, with the purpose of understanding the interactions that happen in the brain when a rat undertakes as specified activity (walk, run, sleep, etc.).

[1] J. Altamiranda, and J. Aguilar are with the Universidad de los Andes, Facultad de Ingeniería, Departamento de Computación, CEMISID, Núcleo La Hechicera, Mérida 5101-Venezuela (phone: 00 58 274 2402914; e-mail: aguilar@ula.ve).

[2]L.Hernandez is with the Universidad de los Andes, Facultad de Medicina. Laboratorio de Fisiología, Mérida 5101, Venezuela (email: hernande@ula.ve)

II. Research Questions

The particular problem focus of our activity is the study of neurotransmitters: Glutamate and Aspartate. Can we in this context usefully and efficiently use our Data Mining system to help to analyze the extracted samples from the brain of rodents?

III. Data Mining Method

The system of Data Mining that we have developed facilitates the automatic classification and analysis of the peaks in the graphs, and considers the classification of the samples from rodents by the activity (walk, run, sleep, etc.) (see Fig. 1).

Data Mining has three components: (a) an intermediate representation; (b) a historical data repository; and (c) an algorithm that recognizes and defines patterns. Data Mining is implemented in two stages: pre-processing and post-processing (see Fig. 2).

A. Intermediate Representation

In pre-processing the results of the experiments and the

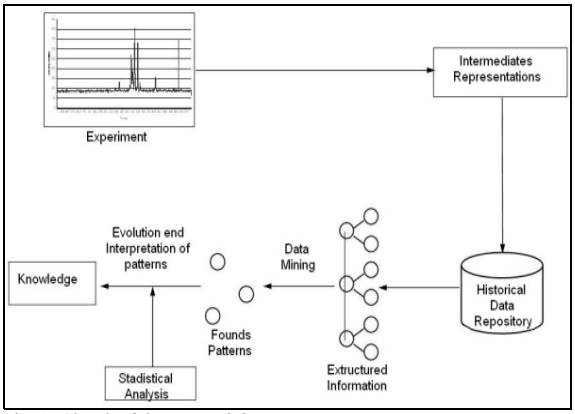


Fig. 1. Sketch of the Data Mining system.

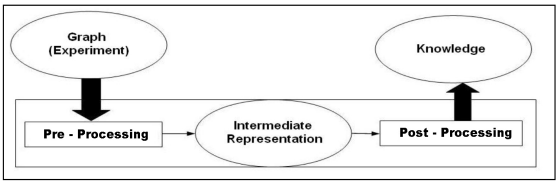


Fig. 2. Process of Data Mining.

978-1-4244-1821-3/08/$25.00©2008 IEEE

graphs are combined into a type of simple intermediate representation that facilitates further analysis. The methods that are used to arrive at this intermediate representation are the following:

1) Savitzky-Golay filter: The objective of this step is to replace the original data by a signal that represent a smoother version of the original signal. (see Fig. 3).

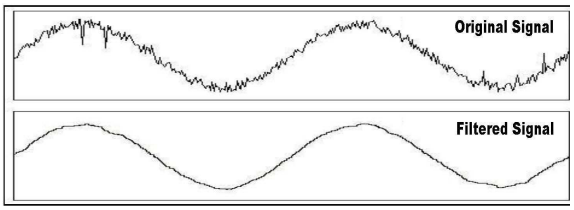


Fig. 3. Savitzky – Golay Filter.

2) Peak extraction algorithm: this divides the graphics that corresponds to the sample into sectors. It selects each sector to represents a peak. The sectors stand for the presence of a chemical substance (see Fig. 4).

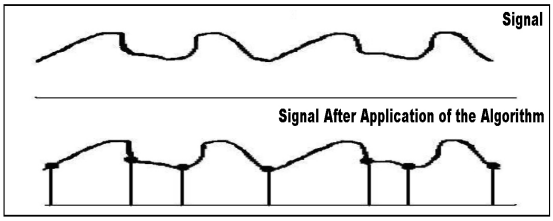


Fig. 4. Algorithm for the extraction of peaks.

When we finish these procedures we obtain the intermediate representation that contains: area, height, begin point, end point and width of peaks. This intermediate representation about peaks of samples is stored in the Historical Data Repository (see Fig. 5).

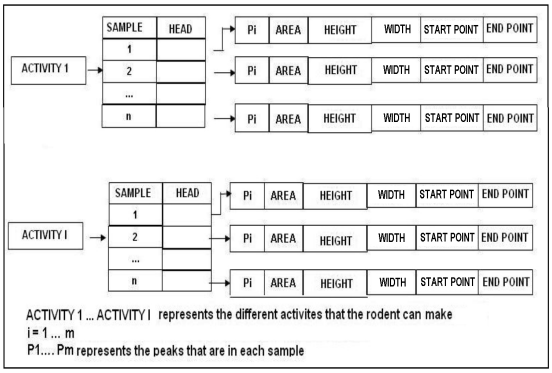


Fig. 5. Historical Data Repository structure.

This intermediate form qualifies the patterns in the graphs for classification into specific activities.

B. Adaptative Resonance Theory Classifier

In post-processing, the intermediate representations of the peaks are analyzed with the objective of classifying them into the different candidate chemical substances and as feed to overall statistical analysis.

Different types of patterns in a data set are arrived at in this analysis. As far as is possible we automate the process of identification of the peaks in the samples, and classification of samples by activity. We use the Artificial Neural Network (ANN) known as ART2 (Adaptative Resonance Theory) [3-7]. Our System of Data Mining comprises two ANN ART2 stages:

1) ANN ART2 to classify the peaks of a sample: This is for the recognition and classification of peaks. In this case, a pattern is a curve that represents a chemical substance that acts on the brain of the rodent. ANN ART2 must be able to recognize a chemical substance by the position that the peak occupies in the input graph. ANN ART2 inputs are begin point and end point. It returns the class to which the peak belongs (see Fig. 6). In this figure P1, P2, P3, P4, P5, P6, P7, P8, P9 are outputs of the ANN ART2 and each one represents a different chemical substance.

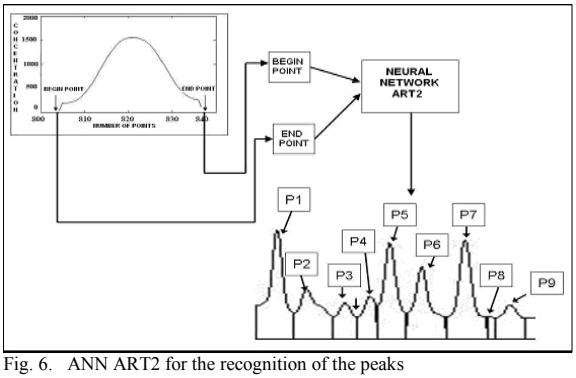


Fig. 6. ANN ART2 for the recognition of the peaks

2) ANN ART2 to classify a determined activity of a rodent: It classifies an activity as eating, sleeping, for example under the effects of a drug. For this task, ANN ART2 classifies the sample (see Fig. 7).

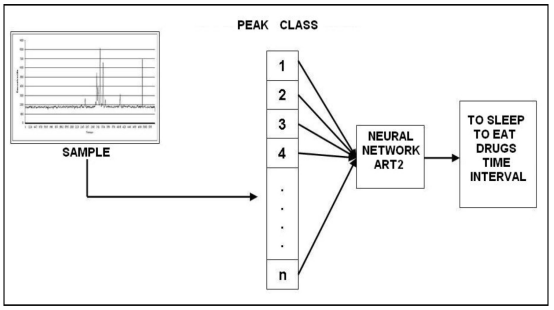


Fig. 7. ANN ART2 Recognition of the sample.

The classification of sample into activity classes therefore comprises these stages: (a) extracting a sample of the brain of a rodent; (b) classifying the peaks from the sample by means of the first ANN ART2; (c) using this to build a list of classes of peaks: these represent the substances present in the sample; (d) using this list as input to the second ANN ART2 stage; (e) the output of this stage is a class of behavior that is associated to the sample.

The number of peaks in the samples is not constant and this presents us a problem because it is necessary to have a fixed number of peaks for the analysis by the ANN ART2 since it is dynamically impossible to change the number of neurons. We set the second stage ANN ART2 a maximum number of n input neurons (peaks discovered in the first stage ANN ART2 to be classified by it) and several exit neurons that represent the classes of activities of the rodent (see Fig. 7). If in the sample the number of peaks is smaller than n, then the list is completed with zeros (absence of peaks). For the classification of the samples these have classes from 1 to k, that is, our system recognizes k different activities.

C. *Implementation*

The Data Mining system was built in Matlab 7.1 under the Linux operating system. For the automated classification were used 7 samples extracted from the brains of rodents. Fig. 8 illustrates an interface to the Data Mining system.

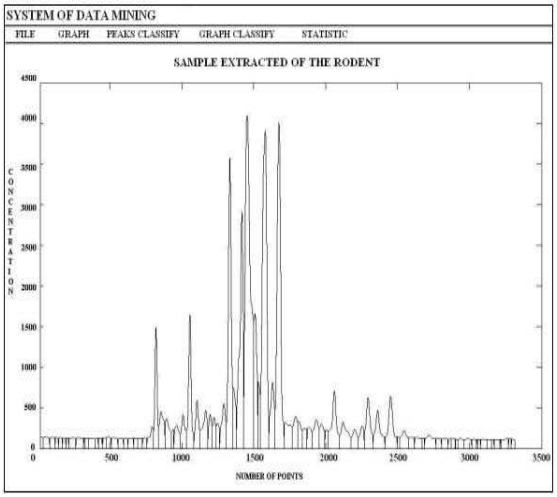
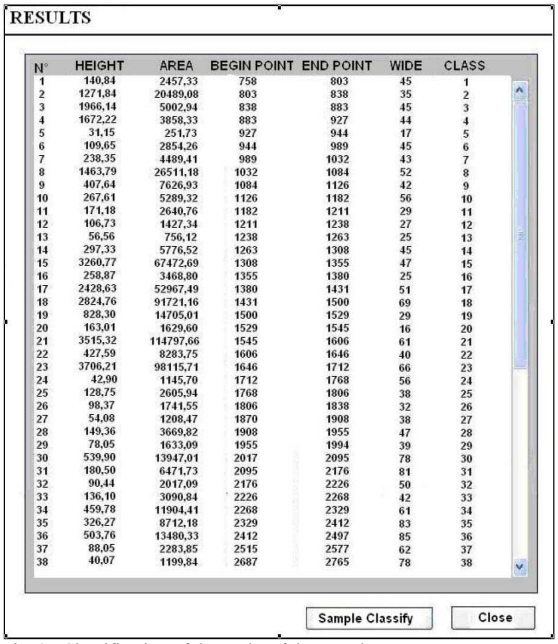


Fig. 8. The main panel of the system of Data Mining and filtrate of the first sample.

To analyze the sample is necessary to filter out the data. The Data Mining system classified the peaks in the corresponding class of the chemical substance that it represented and a typical output is illustrated in Fig. 9.

The system of Data Mining classifies the peaks as numbers from 1 to n, an easy way to represent the classification whereby each number represents a different chemical substance. However, our Data Mining system allows associating new activities to the samples. We use nomenclature 1 to m to classify to new classes of activities.



N°	HEIGHT	AREA	BEGIN POINT	END POINT	WIDE	CLASS
1	140,84	2457,33	758	803	45	1
2	1271,84	20489,08	803	838	35	2
3	1966,14	5002,94	838	883	45	3
4	1672,22	3858,33	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	67472,69	1308	1355	47	15
16	258,87	3468,80	1355	1380	25	16
17	2428,63	52967,49	1380	1431	51	17
18	2824,76	91721,16	1431	1500	69	18
19	828,30	14705,01	1500	1529	29	19
20	163,01	1629,60	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	3669,82	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. Classification of the peaks of the sample.

IV. EXPERIMENTAL RESULTS

A sample associated with class 1 is shown by Fig. 10, i.e. it belongs to the activity 1 since pre-defined classes do not exist.

However, observe in Fig. 11 the phantoms of the

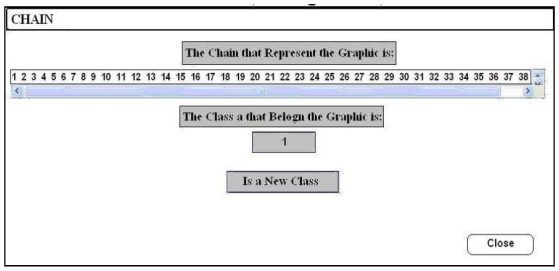


Fig. 10. Classification of the peaks of the sample.

neurotransmitters Glutamate and Aspartate. These appear to correspond very well with the judgement of experts. They corresponded to peaks 34 and 36 in the classification that was arrived at by our Data Mining system.

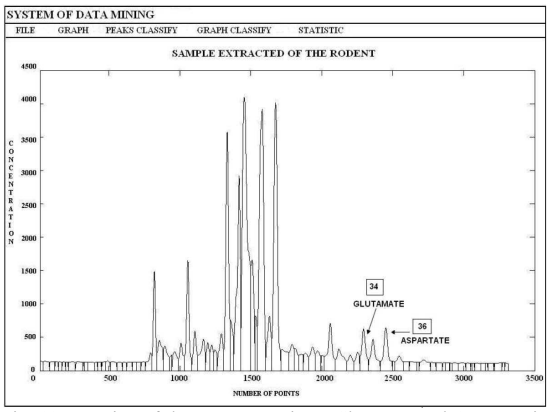


Fig. 11. Location of the neurotransmitters Glutamate and Aspartate in the first sample

For the peak of Glutamate and Aspartate in the first sample, our system of Data Mining gave the results as shown in Fig. 12. Note that the peaks of the neurotransmitters Glutamate and Aspartate appear in both samples. As our system recognizes them automatically, it gives in addition the characteristics for each peak, i.e., height, area, start point, end point, width and class to which it belongs. Additionally, our system of Data Mining can compare peaks from different samples.

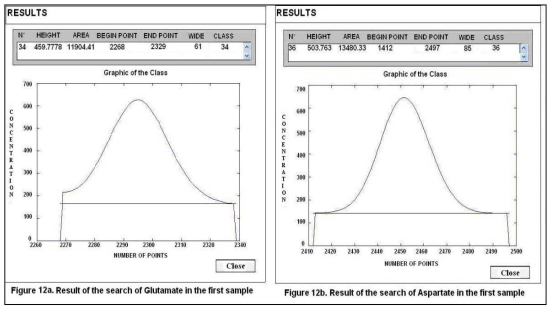


Fig. 12. Result research for Glutamate and Aspartate in the first sample

We plan to compare in the first and second sample the phantoms of the neurotransmitters Glutamate and Aspartate (see Fig. 13).

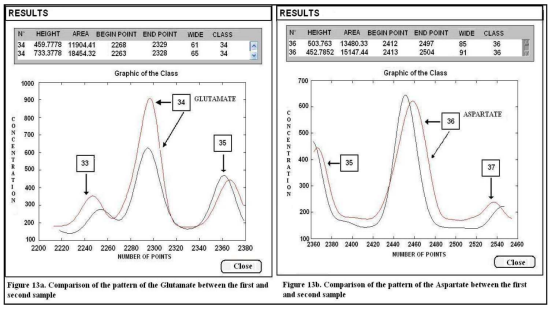


Fig. 13. Glutamate and Aspartate between the first and the second samples.

We also plan to develop the Data Mining system to make a statistical analysis of the peaks of the chemical substances, using variables such as height and area. We will use this to study the height and the area of the Glutamate neurotransmitter (Fig. 14). As this illustration is difficult to discern from this figure, it is presented as Table 1.

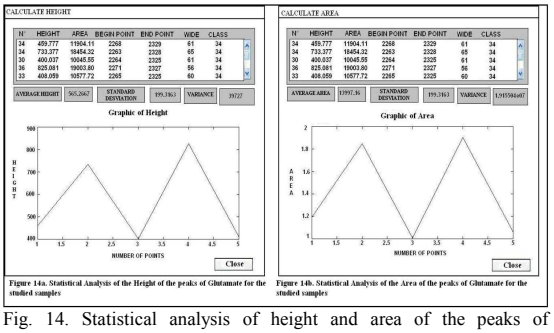


Fig. 14. Statistical analysis of height and area of the peaks of Glutamate for the studied

Thus, we can see the variation of the values of the Glutamate in the samples. This is studied by the experts to reach conclusions with respect to the operation of the brain of the rodent, which can be originated by the presence of substances administered by them (for example, a drug). In this particular case we have for the Glutamate: average weight =˙565, 2667, variance =˙39727, standard deviation =˙199˂3163. For the case of the area: average area =˙13997˂ψ 16, variance =˙19155040, standard deviation =˙4376˂ψ676.

TABLE I
CHARACTERISTICS OF PEAKS REPRESENTING GLUTAMATE

Peak No.	Height	Area	Star pt.	End pt.	Width	Class
1	459,7	11904,4	2268	2329	61	34
2	733,3	18454,3	2263	2328	65	34
3	400,0	10045, 5	2264	2325	61	34
4	825,0	19003,8	2271	2327	56	34
5	408,0	10577,5	2265	2325	60	34

V. CONCLUSION

The extraction of patterns and knowledge from sources of data is possible because there exist computational tools for this purpose. The Data Mining system we have developed comprises methods to extract knowledge that is represented in patterns.

It enables the representation of knowledge that emerges naturally in patterns that are created by chemical substances in the brain of rodents. The goals of our Data Mining system are simple: to discover the chemical substances in the samples from the brains of rodents, and to classify these samples into activities.

The area of attention for our analysis are the neurotransmitters Glutamate and Aspartate. The results can be considered to be very promising since, from the Data Mining point of view, the system discovers patterns of the

chemical substances and obtains consistent results.

Our results indicate that our Data Mining system is sufficiently useful and efficient to help to analyze the extracted samples from the brain of rodents, and its results exhibit a strong correlation with the judgement of human specialists.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the FONACIT 2005000170 project.

REFERENCES

[1] Hernández L., "Manual de CZE", Universidad de Los Andes, Facultad de Medicina, Departamento de Fisiología, Mérida - Venezuela, (2004). Technical Report

[2] León V., "Manual de usuario del ONICE", Universidad de los Andes, Facultad de Medicina, Departamento de Fisiología, Mérida - Venezuela, (1998) Technical Report

[3] Agrawal R., Shafer J., "Data Mining & Knowledge Discovery in Databases (KDD)." IEEE Transactions on Knowledge and Data Engineering, December (1996).

[4] Fallad U., Piatesky - Shapiro G., Smyth P. (eds.) "Advances in Knowledge Discovery and Data Mining", MIT Press, (1996) pp 37 – 57

[5] Aguilar J., Rivas F. (Ed.), "Introducción a la Computación Inteligente", MERITEC, Venezuela, (2001).

[6] Carpenter G., Grossberg S., "The ART of Adaptative Resonance Theory by a Self - Organizing Neural Network" IEEE Computer, 21(3) (1988) pp 77 – 88

[7] Higuera J., Matinez V. "Redes Neuronales Artificiales: Fundamentos, Modelos y Aplicaciones". Addison -Wesley. (1995)