Proyecto Final - Reconocimiento de Patrones

Salim Perchy
Maestría en Ingeniería
Énfasis en Computación
Pontificia Universidad Javeriana
Cali, Valle del Cauca
Email: ysperchy@cic.javerianacali.edu.co

Mario Mora
Maestría en Ingeniería
Énfasis en Computación
Pontificia Universidad Javeriana
Cali, Valle del Cauca
Email: mariomora@javerianacali.edu.co

Resumen

El siguiente escrito documenta el proceso de abstracción y experimentación para el reconocimiento de sitios de clivaje en poliproteinas con finalidad de automatizar dicho proceso usando técnicas vistas en la materia *Reconocimiento de Patrones*

I. Introduction

La predicción de puntos de clivaje en secuencias de poliproteinas es de amplia importancia para la comunidad científica, ya que de ellos depende el análisis del comportamiento de la familia viral *Potyviridae* que cuyos efectos negativos en plantaciones van desde infecciones hasta la muerte de estos mismos[1].

Este informe pretende documentar los diferentes experimentos llevados a cabo y técnicas usadas para poder automatizar la predicción de dichos puntos de clivaje para posterior uso en el estudio del funcionamiento del Potyvirus.

El documento está organizado de la siguiente manera: una sección dedicada a detallar las herramientas, técnicas y código usado para automatizar dicho proceso, la sección posterior es dedicada a mostrar los resultados(en términos de desempeño en error) de cada experimento y también qué parámetros al igual que su forma se tomaron para ejecutar cada iteración de los experimentos.

Finalmente se cierra con una sección de conclusiones sobre los resultados mostrados.

II. Herramientas, Técnicas y Código usado

La principal herramienta usada para automatizar el proceso de predicción de puntos de clivaje fue `MATLAB`, en ella se desarrollaron las rutinas necesarias para el pre-procesamiento a los datos de las secuencias de poliproteinas.

Se usaron algoritmos de reconocimiento de patrones para la predicción de los clivajes en la secuencias ya procesadas. En esta fase se usaron implementaciones ya existentes de estos algoritmos, la librería `PRTools`[2] los provee de manera óptima y sobre `MATLAB`.

También se usó la técnica de *Cross-Validation*[3] para encontrar el mejor estimado del error de clasificación y así evitar *sobreentrenamiento* y *Distorción en los errores* en cada clasificador usado.

Debido a que la librería `PRTools` solamente lee datos numéricos para entrenar sus clasificadores y también que las secuencias de poliprotenias son bastante grandes(orden de miles de datos en tipo caracter), fue también necesario hacer un procedimiento que seleccionara y transformara dichas secuencias para su uso en `MATLAB`.

Las secuencias entonces quedaban en el siguiente formato:

```
#potyvirus.dat

89 84 73 69 1
84 73 69 72 2
69 82 73 69 1
82 73 69 89 2
.
.
.
```

Donde 1 ó 2 significa si la muestra pertenece o no a una secuencia donde hay clivaje respectivamente. Los número restantes son el equivalente *ASCII* de cada letra de la secuencia. Se puede observar que no se tomó la secuencia en su totalidad, esto se explica con detalle en la sección de *Experimentos*

La siguiente es la función en MATLAB que calcula el error de cada clasificador y devuelve el mejor en cada caso:

```
function [ errores, classifiers ] =
  entrenar_clasificadores( datasets, folds, repeticiones, clasificadores )
    [errores stds] = crossval(datasets, clasificadores, folds, repeticiones, []);
    [err indices] = min(transpose(errores));

    for i=1:length(indices);
        classifiers{i} = datasets{i}*clasificadores{indices(i)};
    end;
end
```

Como se puede ver, la función *crossval* hace la mayoría del trabajo, también cabe notar que esta función permite recibir más de un clasificador para hacer *Cross-Validation* y al mismo tiempo puede procesar diferentes conjuntos de datos, lo cual también fue usado para validar los datos de los 9 puntos de clivaje documentados en las secuencias.

III. Experimentos

Los datos de entrenamiento y validación usados en el experimento consistieron de 445 secuencias de poliproteinas cada una con 9 puntos de clivaje, cada posición de la proteina fue representada por una letra del abecedario latino.

Se realizaron en total 3 experimentos, cada uno corresponde a una selección y extracción de características específica para los datos de las secuencias. Invariantemente cada experimento usó 2 grupos de clasificadores:

Grupo Lineal:

- *Fisher* (`fisherc`)
- *Nearest Mean* (`nmc`)
- *Linear Density Based Linear* (`ldc`)
- *Quadratic Density Based Linear* (`qdc`)

Grupo No Lineal:

- *Parzen* (`parzenc`)
- *K-Nearest Neighbors* (`knnc`)
- *Neural Network* (`neurc`)

Debido a que PRTools optimiza los parámetros de la mayoría de los clasificadores lineales, estos se usaron directamente sin especificar parámetros. En el grupo de los No-Lineales, con excepción del *parzen*(que también tiene un parámetro auto optimizable) se probaron diferentes valores:

- K-Nearest Neighbors:
  - 1 Vecino
  - 2 Vecinos
  - 4 Vecinos
- Neural Network:
  - 2 Unidades escondidas
  - 10 Unidades escondidas
  - 40 Unidades escondidas

Igualemente para los 3 experimentos se usaron los mismos parámetros para la validación cruzada; se crearon 10 folds para cada repetición, esto debido a que la práctica ha mostrado que este es el mejor parámetro en la mayoría de las ocasiones[4].

Se efectuaron 5 repeticiones de validación cruzada para cada clasificador.

El error recolectado es de índole frecuentista, esto quiere decir que sencillamente informa el porcentaje de desaciertos en las clasificaciones de secuencias de validación.

El proceso de validación cruzada en el grupo de clasificadores lineales tomaba aproximadamente 1 minuto en realizarse, en el grupo de clasificadores no lineales tomaba aproximadamente entre 4 y 6 horas.

III-A. Experimento No. 1

En el primer experimento la representación de los datos válidos (aquellos que contienen un punto de clivaje) con dimensión $\mathbb{N}^6$ se muestra en la figura 1.

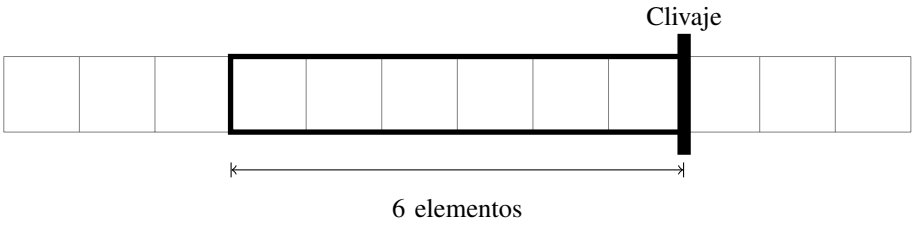

Figura 1. Representación No.1

Los errores del primer y segundo grupo de clasificadores se aprecian en los cuadros I y II.

	Fisher	**Nearest Mean**	**Normal Dist. Linear**	**Normal Dist. Quadratic**
Clivaje 1	11.53 %	21.03 %	11.53 %	9.21 %
Clivaje 2	1.42 %	1.30 %	1.42 %	1.55 %
Clivaje 3	6.47 %	6.85 %	6.47 %	3.12 %
Clivaje 4	4.97 %	5.89 %	4.97 %	1.98 %
Clivaje 5	5.89 %	6.27 %	5.89 %	3.91 %
Clivaje 6	4.25 %	6.97 %	4.25 %	4.83 %
Clivaje 7	8.63 %	7.66 %	8.63 %	5.33 %
Clivaje 8	13.60 %	13.80 %	13.60 %	2.61 %
Clivaje 9	7.01 %	8.65 %	7.01 %	5.01 %
Promedio	7.09 %	10.0 %	7.09 %	4.17 %

Cuadro I
ERRORES DE CLASIFICADORES LINEALES - EXPERIMENTO NO.1

	Parzen	**K-NN(1)**	**K-NN(2)**	**K-NN(4)**	**Neural(2)**	**Neural(10)**	**Neural(40)**
Clivaje 1	2.92 %	2.83 %	3.03 %	3.51 %	5.48 %	2.88 %	2.22 %
Clivaje 2	1.15 %	2.40 %	2.13 %	1.75 %	1.46 %	1.37 %	1.08 %
Clivaje 3	1.80 %	1.46 %	1.57 %	1.80 %	2.94 %	2.16 %	1.91 %
Clivaje 4	1.46 %	1.82 %	2.07 %	2.29 %	1.19 %	1.06 %	1.71 %
Clivaje 5	1.82 %	1.71 %	2.43 %	2.58 %	3.01 %	1.91 %	2.09 %
Clivaje 6	1.84 %	2.13 %	2.56 %	2.67 %	2.34 %	1.46 %	2.18 %
Clivaje 7	4.13 %	2.99 %	4.56 %	5.15 %	3.28 %	2.74 %	2.72 %
Clivaje 8	3.55 %	2.72 %	2.20 %	1.80 %	4.61 %	2.22 %	2.83 %
Clivaje 9	3.26 %	3.35 %	3.89 %	3.62 %	4.79 %	2.99 %	2.94 %
Promedio	2.44 %	2.38 %	2.72 %	2.80 %	3.23 %	2.09 %	2.19 %

Cuadro II
ERRORES DE CLASIFICADORES NO LINEALES - EXPERIMENTO NO.1

III-B. Experimento No. 2

En el segundo experimento la representación de los datos válidos con dimensión $\mathbb{N}^3$ se muestra en la figura 2.

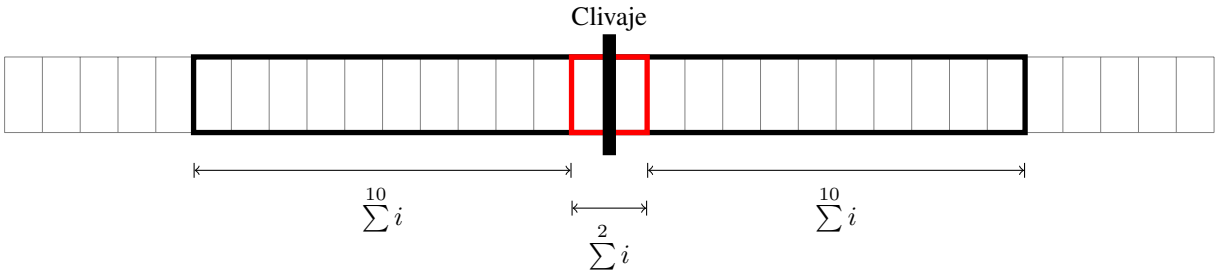

Figura 2. Representación No.2

Los errores del primer y segundo grupo de clasificadores se aprecian en los cuadros III y IV.

	Fisher	**Nearest Mean**	**Normal Dist. Linear**	**Normal Dist. Quadratic**
Clivaje 1	29.80 %	26.22 %	29.80 %	28.61 %
Clivaje 2	18.16 %	28.76 %	18.16 %	15.78 %
Clivaje 3	39.51 %	36.81 %	39.51 %	29.10 %
Clivaje 4	13.26 %	44.43 %	13.26 %	33.44 %
Clivaje 5	41.84 %	40.34 %	41.84 %	49.51 %
Clivaje 6	40.04 %	48.43 %	40.04 %	14.18 %
Clivaje 7	32.18 %	28.27 %	32.18 %	24.83 %
Clivaje 8	45.37 %	43.51 %	45.37 %	42.43 %
Clivaje 9	32.67 %	34.81 %	32.67 %	33.37 %
Promedio	32.54 %	36.84 %	32.54 %	30.14 %

Cuadro III
ERRORES DE CLASIFICADORES LINEALES - EXPERIMENTO NO.2

	Parzen	**K-NN(1)**	**K-NN(2)**	**K-NN(4)**	**Neural(2)**	**Neural(10)**	**Neural(40)**
Clivaje 1	10.29 %	7.51 %	8.40 %	9.46 %	16.16 %	4.47 %	4.67 %
Clivaje 2	8.99 %	4.58 %	6.63 %	7.73 %	3.06 %	2.20 %	2.36 %
Clivaje 3	9.33 %	6.65 %	8.65 %	9.39 %	13.33 %	8.65 %	6.52 %
Clivaje 4	5.44 %	4.11 %	5.21 %	5.46 %	6.70 %	2.56 %	2.11 %
Clivaje 5	9.15 %	8.56 %	9.26 %	9.82 %	21.57 %	6.25 %	6.16 %
Clivaje 6	7.06 %	6.70 %	8.45 %	8.74 %	9.46 %	5.87 %	5.87 %
Clivaje 7	6.88 %	6.79 %	7.03 %	7.37 %	11.66 %	5.66 %	5.08 %
Clivaje 8	16.29 %	8.88 %	9.17 %	10.27 %	35.55 %	8.11 %	8.36 %
Clivaje 9	11.69 %	8.58 %	10.74 %	11.57 %	22.97 %	4.13 %	5.26 %
Promedio	9.46 %	6.93 %	8.17 %	8.87 %	15.61 %	5.32 %	5.15 %

Cuadro IV
ERRORES DE CLASIFICADORES NO LINEALES - EXPERIMENTO NO.2

III-C. Experimento No. 3

En el tercer experimento la representación de los datos válidos con dimensión $\mathbb{N}^{10}$ se muestra en la figura 3.

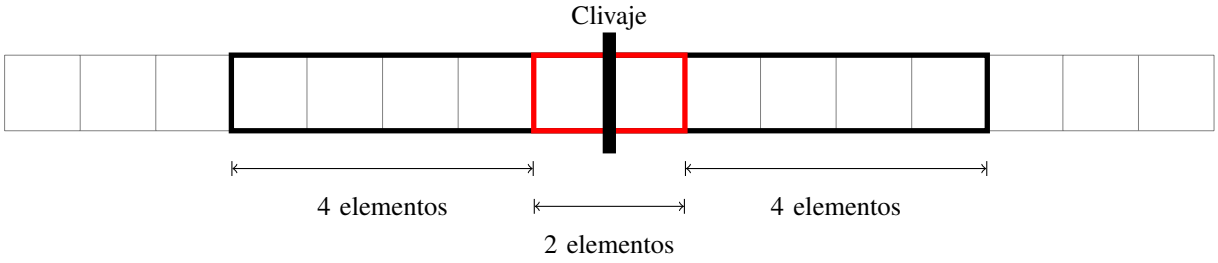

Figura 3. Representación No.3

Los errores del primer y segundo grupo de clasificadores se aprecian en los cuadros V y VI.

	Fisher	Nearest Mean	Normal Dist. Linear	Normal Dist. Quadratic
Clivaje 1	12.43 %	13.60 %	12.43 %	7.26 %
Clivaje 2	1.44 %	1.57 %	1.44 %	1.44 %
Clivaje 3	4.13 %	4.04 %	4.13 %	1.93 %
Clivaje 4	7.46 %	6.29 %	7.46 %	1.03 %
Clivaje 5	5.06 %	5.57 %	5.06 %	3.19 %
Clivaje 6	5.84 %	7.84 %	5.84 %	4.88 %
Clivaje 7	5.42 %	7.46 %	5.42 %	2.94 %
Clivaje 8	8.27 %	14.34 %	8.27 %	3.89 %
Clivaje 9	7.24 %	10.79 %	7.24 %	6.02 %
Promedio	6.37 %	7.94 %	6.37 %	3.62 %

Cuadro V
ERRORES DE CLASIFICADORES LINEALES - EXPERIMENTO NO.3

	Parzen	K-NN(1)	K-NN(2)	K-NN(4)	Neural(2)	Neural(10)	Neural(40)
Clivaje 1	3.44 %	3.75 %	3.98 %	4.04 %	5.64 %	4.04 %	3.57 %
Clivaje 2	1.19 %	1.33 %	1.51 %	1.12 %	1.46 %	1.48 %	1.66 %
Clivaje 3	1.55 %	2.07 %	2.13 %	2.00 %	2.38 %	2.25 %	2.25 %
Clivaje 4	1.24 %	1.66 %	2.00 %	1.33 %	1.82 %	1.51 %	1.71 %
Clivaje 5	2.20 %	2.76 %	2.72 %	2.83 %	4.16 %	2.70 %	3.10 %
Clivaje 6	1.28 %	1.53 %	1.35 %	1.12 %	2.09 %	1.87 %	1.84 %
Clivaje 7	3.21 %	2.83 %	2.94 %	3.39 %	4.16 %	2.99 %	2.49 %
Clivaje 8	3.37 %	2.67 %	3.01 %	2.94 %	3.08 %	2.58 %	2.92 %
Clivaje 9	3.60 %	3.26 %	4.31 %	4.76 %	3.80 %	3.98 %	3.80 %
Promedio	2.34 %	2.43 %	2.66 %	2.61 %	3.18 %	2.60 %	2.59 %

Cuadro VI
ERRORES DE CLASIFICADORES NO LINEALES - EXPERIMENTO NO.3

III-D. Medidas de Desempeño

Por medio de la matriz de confusión se analizaron las medidas de desempeño en cada uno de los experimentos. Para ello se tomaron los clasificadores que mejor se comportaron con respecto a los errores de clasificación, es decir los que tuvieron menor porcentaje de error clasificando los datos en promedio, para los nueve puntos de clivaje. Estos fueron:

- Experimento No.1: *Quadratic Density Based Linear* (`qdc`), entre los clasificadores lineales y el *Neural Network* (`neurc`) con 10 capas ocultas entre los no-lineales
- Experimento No.2: *Quadratic Density Based Linear* (`qdc`), entre los clasificadores lineales y el *Neural Network* (`neurc`) con 40 capas ocultas entre los no-lineales
- Experimento No.3: *Quadratic Density Based Linear* (`qdc`), entre los clasificadores lineales y el *Parzen* (`parzenc`) entre los no-lineales

III-D1. Medidas de Desempeño - Experimento No.1: Para el primer experimento, en las figuras 4 y 5 se muestran la matrices de confusión respectivas para los clasificadores *Quadratic Density Based Linear* (`qdc`) y *Neural Network* (`neurc`) con 10 capas ocultas.

Desde dichas matrices de confusión se calculan las medidas de desempeño, las cuales se encuentran en los cuadros VII y VIII.

Para el caso del clasificador lineal cuadrático (ver: VII) notamos que la **Sensibilidad** es de **0.8439**, lo que indica que el desempeño de este clasificador para clasificar como positivos las muestras que realmente son positivas es del 84.39 %; mientras que la **Especificidad** que es de **0.8501**, está indicando que el 85.01 % de las muestras clasificadas como negativas de verdad lo eran.

Finalmente el **Recall**, de **0.8802** indica que el 88.02 % de las muestras están clasificadas correctamente.

Para el caso del clasificador red neuronal con 10 capas ocultas (ver: VIII) se observa que la **Sensibilidad** es de **0.9788**, lo que indica que el desempeño de este clasificador para clasificar como positivos las muestras que realmente son positivas es del 97.88 %; mientras que la **Especificidad** que es de **0.9786**, está indicando que el 97.86 % de las muestras clasificadas como negativas de verdad lo eran.

Finalmente el **Recall**, de **0.9727** indica que el 97.27 % de las muestras están clasificadas correctamente.

III-D2. Medidas de Desempeño - Experimento No.2: Para el segundo experimento, en las figuras 6 y 7 se muestran la matrices de confusión respectivas para los clasificadores *Quadratic Density Based Linear* (`qdc`) y *Neural Network* (`neurc`) con 40 capas ocultas.

True Labels	Estimated Labels: Clivaj	Estimated Labels: No Cli	Totals
Clivaje	676	125	801
No Clivaje	92	709	801
Totals	768	834	1602

Figura 4. Matriz de confusión para el experimento No.1 y clasificador *Quadratic Density Based Linear*

True Labels	Estimated Labels: Clivaj	Estimated Labels: No Cli	Totals
Clivaje	784	17	801
No Clivaje	22	779	801
Totals	806	796	1602

Figura 5. Matriz de confusión para el experimento No.1 y clasificador *Neural Network* (`neurc`) con 10 capas ocultas

Sensitivity	**Specificity**	**Accuracy**	**Recall**
0,8439	0,8501	0,8645	0,8802

Cuadro VII
MEDIDAS DE DESEMPEÑO PARA *Quadratic Density Based Linear* - EXPERIMENTO NO.1

Sensitivity	**Specificity**	**Accuracy**	**Recall**
0,9788	0,9786	0,9757	0,9727

Cuadro VIII
MEDIDAS DE DESEMPEÑO PARA *Neural Network - 10 capas ocultas* - EXPERIMENTO NO.1

Desde dichas matrices de confusión se calculan las medidas de desempeño, las cuales se encuentran en los cuadros IX y X.

Para el caso del clasificador lineal cuadrático (ver: IX) notamos que la **Sensibilidad** es de **0.4382**, lo que indica que el desempeño de este clasificador para clasificar como positivos las muestras que realmente son positivas es del 43.82 %; mientras que la **Especificidad** que es de **0.5530**, está indicando que el 55.30 % de las muestras clasificadas como negativas de verdad lo eran.

Finalmente el **Recall**, de **0.5840** indica que el 58.40 % de las muestras están clasificadas correctamente.

Realmente, según estas medidas de desempeño el clasificador no es muy bueno.

Muy diferente es el caso del clasificador red neuronal con 40 capas ocultas (ver: X), en el que se observa que la **Sensibilidad** es de **0.9238**, lo que indica que el desempeño de este clasificador para clasificar como positivos las muestras que realmente son positivas es del 92.38 %; mientras que la **Especificidad** que es de **0.9257**, está indicando que el 92.57 % de las muestras clasificadas como negativas de verdad lo eran.

Finalmente el **Recall**, de **0.9475** indica que el 94.75 % de las muestras están clasificadas correctamente.

Este clasificador si obtuvo mejores resultados en este experimento.

Sensitivity	**Specificity**	**Accuracy**	**Recall**
0,4382	0,5504	0,5630	0,5840

Cuadro IX
MEDIDAS DE DESEMPEÑO PARA *Quadratic Density Based Linear* - EXPERIMENTO NO.2

True Labels	Estimated Labels: Clivaj	Estimated Labels: No Cli	Totals
Clivaje	351	450	801
No Clivaje	250	551	801
Totals	601	1001	1602

Figura 6. Matriz de confusión para el experimento No.2 y clasificador *Quadratic Density Based Linear*

True Labels	Estimated Labels: Clivaj	Estimated Labels: No Cli	Totals
Clivaje	740	61	801
No Clivaje	41	760	801
Totals	781	821	1602

Figura 7. Matriz de confusión para el experimento No.2 y clasificador *Neural Network* (`neurc`) con 40 capas ocultas

Sensitivity	**Specificity**	**Accuracy**	**Recall**
0,9238	0,9257	0,9363	0,9475

Cuadro X
MEDIDAS DE DESEMPEÑO PARA *Neural Network - 40 capas ocultas* - EXPERIMENTO NO.2

III-D3. Medidas de Desempeño - Experimento No.3: Para el tercer experimento, en las figuras 8 y 9 se muestran la matrices de confusión respectivas para los clasificadores *Quadratic Density Based Linear* (`qdc`) y *Parzen* (`parzenc`).

Desde dichas matrices de confusión se calculan las medidas de desempeño, las cuales se encuentran en los cuadros XI y XII.

Para el caso del clasificador lineal cuadrático (ver: XI) notamos que la **Sensibilidad** es de **0.8664**, lo que indica que el desempeño de este clasificador para clasificar como positivos las muestras que realmente son positivas es del 86.64 %; mientras que la **Especificidad** que es de **0.8700**, está indicando que el 87 % de las muestras clasificadas como negativas de verdad lo eran.

Finalmente el **Recall**, de **0.8909** indica que el 89.09 % de las muestras están clasificadas correctamente.

Para el caso del clasificador de Parzen (ver: XII) las medidas de sempelo mejoran sustancialmente; en él se observa que la **Sensibilidad** es de **0.9700**, lo que indica que el desempeño de este clasificador para clasificar como positivos las muestras que realmente son positivas es del 97 %; mientras que la **Especificidad** que es de **0.9701**, está indicando que el 97.01 % de las muestras clasificadas como negativas de verdad lo eran.

Finalmente el **Recall**, de **0.9725** indica que el 97.25 % de las muestras están clasificadas correctamente.

Este clasificador obtuvo muy buenos resultados en este experimento.

Sensitivity	**Specificity**	**Accuracy**	**Recall**
0,8664	0,8700	0,8801	0,8909

Cuadro XI
MEDIDAS DE DESEMPEÑO PARA *Quadratic Density Based Linear* - EXPERIMENTO NO.3

```
 True       | Estimated Labels
 Labels     | Clivaj No Cli| Totals
------------|--------------|-------
 Clivaje    |  694    107  |  801
 No Clivaje|   85    716  |  801
------------|--------------|-------
 Totals     |  779    823  | 1602
```

Figura 8. Matriz de confusión para el experimento No.3 y clasificador *Quadratic Density Based Linear*

```
 True       | Estimated Labels
 Labels     | Clivaj No Cli| Totals
------------|--------------|-------
 Clivaje    |  777     24  |  801
 No Clivaje|   22    779  |  801
------------|--------------|-------
 Totals     |  799    803  | 1602
```

Figura 9. Matriz de confusión para el experimento No.3 y clasificador *Parzen* (`parzenc`)

Sensitivity	**Specificity**	**Accuracy**	**Recall**
0,9700	0,9701	0,9755	0,9725

Cuadro XII
MEDIDAS DE DESEMPEÑO PARA *Parzen* - EXPERIMENTO NO.3

IV. CONCLUSIONES

- La gigante dimensionalidad de las cadenas a analizar hacían casi imposible esta tarea, por lo que hubo la necesidad de explorar unas ventanas que contuvieran los sitios de clivaje identificados previamente por expertos.
- El uso de una ventana óptima, establecida por expertos con anterioridad optimizó el comportamiento de los clasificadores.
- Hubo una ventana propuesta por este trabajo diferente a la sugerida por los expertos en la que el resultado dado por un clasificador en especial, fué muy buena, en ocasiones mejor que la sugerida.
- Queda en todo caso la inquietud con respecto a la validez de medir el desempeño de los clasificadores en el escenario de unión de todos los puntos de clivaje en un solo conjunto de datos y además de hacerlo con el clasificador que mejor se comportó en $promedio$ para los nueve puntos de clivaje.

REFERENCIAS

[1] Alvarez Gloria, *Especificación del proyecto - Reconocimiento de Patrones*. Universidad Javeriana, 2011.

[2] Delft Pattern Recognition Group, http://www.prtools.org/. Delft University of Technology, retrieved 2011.

[3] Bishop Christopher. *PatternRecognition and Machine Learning*. Springer, 2006.

[4] Alvarez Gloria, *Notas de Clase - Reconocimiento de Patrones(2011)*. Universidad Javeriana, 2011.