ALGORITMOS DE CLASIFICACION LINEAL PARA LA IDENTIFICACIÓN DE ZONAS CEREBRALES

ANDRÉS FELIPE LONDOÑO

PROGRAMA DE INGENIERÍA ELÉCTRICA
FACULTAD DE INGENIERÍAS
UNIVERSIDAD TECNOLÓGICA DE PEREIRA
PEREIRA
2016

ALGORITMOS DE CLASIFICACION LINEAL PARA LA IDENTIFICACIÓN DE ZONAS CEREBRALES

ANDRÉS FELIPE LONDOÑO

Trabajo de grado presentado como requisito para optar al título de Ingeniero Electricista

Director:
PhD. Mauricio Álvarez

PROGRAMA DE INGENIERÍA ELÉCTRICA
FACULTAD DE INGENIERÍAS
UNIVERSIDAD TECNOLÓGICA DE PEREIRA
PEREIRA
201

Índice general

Tabla de contenido

Índice general III

Agradecimientos V

Resumen VI

1. Introducción 7

1.1. Definición del problema 8

1.2. Objetivos 9

1.2.1. Objetivo general 9

1.2.2. Objetivos específicos 9

1.3. Trabajos anteriores 10

2. Marco Teórico. 12

2.2. Método de Mínimos Cuadrados para Múltiples Clases. 13

2.3. Clasificador Bayesiano basado en múltiples clases. 14

2.4. Regresión Logística para Múltiples Clases. 16

2.4.1. Formulación del modelo. 16

2.5. Validación cruzada. 18

3. Metodología. 20

3.1. Base de datos. 20

3.2 Algoritmos de clasificación. 22

4. Resultados 23

4.1. Método de Mínimos Cuadrados para Múltiples Clases. 23
3. Trabajos futuros 28
4. Conclusiones 29
5. Bibliografía 30

Agradecimientos

A la Universidad Tecnológica de Pereira, de la cual nos sentimos orgullosos de pertenecer; A sus profesores, a quienes desarrollan su actividad docente comprometidos con la calidad educativa y reconocen que su tarea se engloba en la formación de la persona, del ciudadano y del profesional.

. ……

Resumen

La estimulación cerebral profunda es el procedimiento quirúrgico más ampliamente usado en pacientes con avanzados síntomas en la enfermedad de parkinson , en los que el tratamiento farmacológico mediante L-dopa se torna insuficiente , la principal dificultad del procedimiento es la identificación de la zona cerebral en la cual se debe implantar el micro estimulador .

En este documento se plantean algoritmos de clasificación lineal basados en regresión logística para la identificación de zonas cerebrales a partir de vectores de características obtenidos de registros de micro electrodos.

1. Introducción

La DBS consiste en la implantación de micro electrodos que proveen una diminuta corriente eléctrica continua sobre estructuras profundas del cerebro. En las personas con enfermedad de Parkinson (EP), a medida que las células cerebrales degeneran, los ciclos de retroalimentación eléctrica funcionan anormalmente. Algunas partes tienen actividad excesiva mientras otras tienen menor actividad que la normal. Como resultado, los movimientos físicos normales son reemplazados por temblores, rigidez y lentitud. Pero al usar un micro electrodo en la zona profunda del cerebro que provee una corriente eléctrica, es posible forzar las señales anormales entre las estructuras cerebrales e impulsar la actividad eléctrica del sistema hacia la función normal **[12].**

La estimulación cerebral profunda (DBS) se constituye entonces como una solución eficaz dentro del tratamiento de la enfermedad de Parkinson en casos donde el tratamiento farmacológico no ha reportado mejorías, en cuanto permite una reducción de los síntomas y desórdenes del movimiento asociados a ella. Este procedimiento se logra mediante la inserción de microelectrodos de estimulación en las regiones cerebrales que presentan actividad neuronal incontrolada. Normalmente el subtálamo es la región objetivo en este tipo de procedimientos. La localización de esta zona se logra mediante programas de planeación pre-quirúrgica y mediante el registro eléctrico de la actividad neuronal **[22].**

La identificación de la zona celebrar se lleva a cabo mediante la inserción de un electrodo de registro y estimulación. Mediante esta inserción el electrodo avanza recogiendo los diferentes registros de las zonas del cerebro que va atravesando. El patrón de actividad eléctrica en el núcleo subtalamico se compone de picos asimétricos de alta frecuencia, y típicamente exhiben patrones de ruptura en la enfermedad de Parkinson. Este registro responde a los movimientos pasivos de las articulaciones de las extremidades y propioceptivos, tales como la presión muscular, así como que exhiben actividad sincrónica cuando el paciente tiene temblor en el lado contralateral al hemisferio que se está registrando.)[20]. Cuando el electrodo entra en la SNr(sustancia nigra reticulada), se registran picos simétricos de gran amplitud y actividad regular y que no responde a los estímulos externos en general .

Estos registro constituyen la base para la identificación de la zona definitiva en la cual implantar el microeletrodo ,y realizar la estimulación de la zona cerebral en cuestión; en este trabajo se realiza dicha clasificación usando algoritmos de clasificación lineal para identificar los zonas cerebrales, los métodos usados son [5]-[6] :

Clasificador Bayesiano basado en múltiples clases.

Método de Mínimos Cuadrados para Múltiples Clases.

Regresión Logística para Múltiples Clases.

1.1. **Definición del problema**

En medicina uno de los trastornos del movimiento más estudiados tiene relación con la **enfermedad de Parkinson**. A través de diferentes estudios se ha podido establecer que la estimulación adecuada de determinadas zonas cerebrales del núcleo subtalámico (STN) han aportado mejorías significativas a pacientes que padecen esta enfermedad [12][13], debido a que esta zona controla la parte motora, estos problemas motrices suelen ser lentos en el inicio y en la ejecución (bradicinesia) también se manifiestan con temblor aún cuando la persona se encuentra en reposo y rigidez muscular del cuerpo humano, por lo tanto un mal funcionamiento de las neuronas en esta región del cerebro genera un recrudecimiento de la enfermedad.

Con el fin de reducir los síntomas inherentes a la enfermedad de Parkinson, suele recurrirse a un procedimiento quirúrgico conocido como estimulación cerebral profunda (DBS en inglés), el cual consiste en la implantación de micro estimuladores en el núcleo del sub tálamo (STN) o en el Glóbulo pálido interno (GPi).Una de las tareas más difíciles en esta cirugía consiste en identificar la región del cerebro en la cual debe ubicarse el micro estimulador [14][15].

Muchas de las cirugías de este tipo emplean grabaciones de micro electrodo (MER)[9][11] en la cual se toman lecturas de estática rítmica producida por los potenciales de acción de las neuronas que se encuentran cerca al electrodo instalado en el paciente, a juicio de estas lecturas se determina la zona del subtálamo en la cual se encuentra ubicado el micro electrodo, lastimosamente este procedimiento es ambiguo y depende de la pericia y experiencia del cirujano además carece de repetitividad y por consiguiente no es posible entrenar nuevos profesionales.

Una de las alternativas para asistir o ayudar en la cirugía para identificar la región cerebral en la cual se debe ubicar el electrodo, es formularlo como un problema de

clasificación[13] o reconocimiento de patrones, mediante el conocimiento previo de cirugías exitosas y las señales obtenidas de los micro estimuladores instalados en los pacientes.

En este trabajo de grado se pretende explorar la posibilidad de emplear otro tipo de clasificadores conocidos como paramétricos, basados en modelos probabilísticos lineales, este tipo de clasificadores a pesar de su simplicidad algorítmica han sido menos explorados en la literatura referida.

Debido a la complejidad que presenta este intervención quirúrgica llamada estimulación cerebral profunda (DBS) se pretende desarrollar técnicas automatizadas que brinden apoyo a las decisiones que pueda tomar el cirujano para una correcta selección de la zona del núcleo subtalamico (STN) en la cual se pretende estimular con el micro electrodo[17].

En este contexto y aplicándolo a la teoría de clasificación de patrones, cada zona cerebral del sub tálamo representa una clase o categoría particular, mientras las características de entrada al sistema son registro de vectores de características obtenidos de las grabaciones de micro electrodos (MER).

De esta manera se inicia el desarrollo de una metodología de clasificación de patrones basada en regresión logística bayesiana para la identificación de zonas cerebrales [18].

1.2. **Objetivos**

1.2.1. **Objetivo general**

Desarrollar algoritmos de clasificación lineal basados en regresión logística para la identificación de zonas cerebrales a partir de vectores de características obtenidos de registros de micro electrodos.

1.2.2. **Objetivos específicos**

- Aplicar un clasificador Bayesiano lineal a una base de registro de micro electrodos.
- Desarrollar un clasificador lineal basado en regresión logística y aplicarlo en una base de datos de registro de micro electrodos.

- Comparar el desempeño de clasificadores lineales en el problema del reconocimiento de zonas cerebrales.

1.3. **Trabajos anteriores**

En el programa de ingeniería eléctrica, en trabajos recientes tanto en nivel de pregrado como de maestría se han implementado diferentes técnicas de clasificación de patrones para la identificación de zonas cerebrales; estas técnicas incluyen: Los K-vecinos más cercanos [15]-[7], modelos ocultos de Markov [7] y máquinas de soporte vectorial [16]. Estos diferentes tipos de clasificadores pertenecen a la categoría de clasificadores “No paramétricos”, que emplean un conjunto de las observaciones para realizar la clasificación de estas observaciones de una manera adecuada.

A continuación se ilustran los proyectos y trabajos propuestos en clasificación para la identificación de zonas cerebral.

En 2005 [16] A. Orozco *et al.* presentan una metodología para la identificación de zonas del sub tálamo que incluye transformada de Wavelets a las señales generadas en las grabaciones de micro electrodo (MER) para la correcta caracterización de la no estacionariedad del tren de pico de las señales y modelos ocultos de Markov (HMM) que permiten el modelado del comportamiento dinámico de secuencias en el tiempo. Una forma diferente para la validación de los modelos consiste en medir una distancia estocástica dada entre los modelos y la señal que pretende ser reconocida y el uso de K- vecinos más cercanos como la regla de clasificación.

En 2006 [20] A. Orozco *et al.* Presentan el análisis de proximidad usando modelos ocultos de Markov (HMM) para la identificación de las fuentes de pico (tálamo y Sub tálamo). La idea general detrás de un análisis basado en la proximidad consiste en que dado un conjunto de valores de disimilitud por parejas, un nuevo espacio de representación pueda ser construido, en la que cada uno de los objetos se describe por estos valores. Se utiliza el enfoque de proximidad para la clasificación de las diferentes fuentes de pico representados como los parámetros de los modelos ocultos de Markov (HMM).

En 2010 [21] R.D. Pinzón *et al*. Proponen una metodología de clasificación partiendo de un método llamado extracción optima de características wavelet (OWFE) que se construye por los esquemas de elevación (LS), que son una aplicación flexible y rápida de la transformada wavelet (WT). Los operadores en el (LS) se optimizan mediante Algoritmos Genéticos y multiplicadores de LaGrange basándose en la información contenida en las señales de (MER) Entonces un clasificador bayesiano se utiliza para identificar zonas del (STN). El método es

dependiente de la señal y no es necesaria información a priori para descomponer la señal (MER).

En 2012[7] H. Vargas *et al*. Proponen una mejora de la exactitud aplicando aprendizaje supervisado mediante la inclusión de correlaciones entre los pacientes que han sido sometidos a cirugía, llamado aprendizaje multi-tarea. En este contexto, se pretende que cada paciente sometido a (DBS) es una tarea diferente seguidamente se intenta aumentar la precisión en la orientación al Núcleo subtalámico (STN) de ese paciente en particular mediante el aprendizaje en varios pacientes. En esta configuración, el aprendizaje multi-tarea se convierte en aprendizaje multi- paciente. Para obtener los vectores de entrada X hacen uso de wavelets adaptativas. Este método tiene varias etapas, primero de divide el conjunto total de datos en conjunto de entrenamiento y conjunto de validación, primero se toma el conjunto de entrenamiento y se aplican diferentes alternativas de clasificación para procesos Gaussianos multi salida y finalmente se mide la precisión de los clasificadores en el conjunto de validación

2. Marco Teórico.

2.1 MARCO CONCEPTUAL

- **Clase:** Se define las clases como el tipo de zona cerebral del subtálamo la cual se pretende clasificar adecuadamente (Zona Incerta-ZI, Thalamus-TH, Substantia nigra reticulata-SNR.)

- **Patrón:** Se define un patrón como un conjunto previamente establecido de todos los posibles objetos a los que se les pudiera dar un nombre, es una entidad vagamente definido Por ejemplo, un patrón puede ser una imagen de la huella, una palabra en cursiva escrito a mano, un rostro humano, o una señal de voz. Dado un patrón, el reconocimiento/clasificación puede consistir en una de las dos tareas siguientes [8]: 1) la clasificación supervisada (por ejemplo, el análisis discriminante) en el que se identifica el patrón de entrada como miembro de una clase predefinida, 2) clasificación no supervisada (por ejemplo, la agrupación) en el que el patrón se le asigna a una clase hasta ahora desconocida. [5]

- **Reconocimiento de Patrones:** Reconocimiento de patrones es el área del conocimiento que se encarga de la descripción, clasificación y agrupamiento de objetos, personas, señales, representaciones etc., En problemas importantes de ingeniería y científicas como la biología, la psicología, la medicina, el marketing, la visión e inteligencia artificial, y la teledetección. [5]

- **Verosimilitud:** Es la apariencia de verdad en las cosas aunque en la realidad no la tengan; bastante como para formar un juicio prudente [21].

- **Estimulación cerebral profunda con microelectrodo**: La estimulación cerebral profunda utiliza un dispositivo llamado neuroestimulador para transmitir señales eléctricas a las áreas del cerebro que controlan el movimiento, el dolor, el peso y el estado de alerta [12].

2.2. **Método de Mínimos Cuadrados para Múltiples Clases.**

Mínimos cuadrados es una técnica de análisis numérico enmarcada dentro de la optimización matemática, en la que, dados un conjunto de pares ordenados —variable independiente, variable dependiente— y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos (un "mejor ajuste"), de acuerdo con el criterio de mínimo error cuadrático [1].

En su forma más simple, intenta minimizar la suma de cuadrados de las diferencias en las ordenadas (llamadas residuos) entre los puntos generados por la función elegida y los correspondientes valores en los datos. Específicamente, se llama mínimos cuadrados promedio (LMS) cuando el número de datos medidos es 1 y se usa el método de descenso por gradiente para minimizar el residuo cuadrado. Se puede demostrar que LMS minimiza el residuo cuadrado esperado, con el mínimo de operaciones (por iteración), pero requiere un gran número de iteraciones para converger [5] .

Considerando un problema de clasificación general con K clases. Una justificación para el uso de los mínimos cuadrados en este contexto es que se aproxima a la esperanza $E(\mathrm{T}\,|\,\mathrm{X})$ de los valores objetivo (T) dado el vector de entrada (X) . Desafortunadamente, estas probabilidades suelen aproximarse bastante mal, de hecho, las aproximaciones pueden tener valores fuera del rango (0,1)[6].

Cada clase C_k es descrita por su propio modelo lineal de manera que cumplan la relación [18]:

$$Y_k(X) = w_k^T X + w_{ko} \qquad (2.1)$$

Cada clase C_k es descrita por su propio modelo lineal de manera que cumplan la relación:

Donde k=1,…, K. convenientemente puede agruparse esta notación de manera vectorial tal que [9]:

$$y(x) = \widehat{W}^T \hat{x} \qquad (2.2)$$

Donde, $\widehat{W}$ es una matriz cuya K-ésimo columna contiene el elemento D+1 del vector dimensión $\widehat{w} = (w_{ko}, {w_k}^T)^T$ y $\hat{x}$ corresponde al vector de entradas aumentado $(1, x^T)^T$ con entrada de prueba x=1.

2.3. **Clasificador Bayesiano basado en múltiples clases.**

El objetivo de un sistema Bayesiano es saber cuál es la hipótesis más probable entre varios conjuntos de datos. Si $P(D)$ es la probabilidad a priori de los datos, $P(D|h)$ su probabilidad dada una hipótesis y se desea estimar $P(h|D)$, la probabilidad posterior de h dados los datos. Se puede plantear el siguiente teorema [1]:

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} \tag{2.3}$$

La hipótesis más probable o MAP (maximun a posteriori hipótesis):

$$h_{MAP} = argmax_{h\in H}(P(h|D))$$

$$h_{MAP} = argmax_{h\in H}(\frac{p(D|h)p(h)}{p(D)}) \tag{2.4}$$

Un algoritmo Bayesiano puede ser fácilmente implantado si se calculan todas las posibles hipótesis en la ecuación (2.3) y se selecciona la hipótesis de mayor probabilidad. En general si tiene un sistema de aprendizaje de lo general a lo específico (o al revés) que busca especializaciones más generales (o generalizaciones más específicas) se puede caracterizar asumiendo que las hipótesis más generales (o específicas) son más probables que otras [5] [18].

El principio básico, ilustrado en el párrafo anterior, puede ser empleado para determinar de una forma a priori la clase en la cual puede ser clasificado un dato, según las probabilidades generadas por una serie de funciones de clasificación bayesiana, previamente definidas [2].

Sea x_i , la matriz de hiperpuntos que contiene cada una de las n clases, de tamaño N_c (muestras por clase) $*$ D (Número de características) $*$ C (Clases). Se procede del siguiente modo [6]:

a. Se calcula el vector de medias u_i de x_i .

b. Se calcula la matriz de covarianza Σ_i de x_i.

c. Se calculan los coeficientes de las funciones discriminantes para cada una de las **n** clases [3]-[11]:

$$W_i = \frac{1}{2}\sum_i^{-1} \quad (2.5)$$

$$W_i = \sum_i^{-1} u_i \quad (2.6)$$

$$W_{i0} = \frac{1}{2} u_i^t \sum_i^{-1} u_i - \frac{1}{2} ln|\sum_i \quad | + lnP(\overline{w_i}) \quad (2.7)$$

Los coeficientes de las funciones discriminantes, son calculados para cada una de las **n** clases, con una base de datos de entrenamiento, previamente definidas [4].

d. Se definen las funciones discriminantes para cada una de las **n** clases:

$$g_i = x^t W_i x + w_i^t x + w_{i0} \quad (2.8)$$

e. Los valores de características seleccionadas y pertenecientes a la base de datos de validación son evaluados sobre cada una de las **n** funciones discriminantes de probabilidad generadas (Ecuación (2.8)). Se supone que la muestra evaluada, pertenece a aquella clase cuya función de probabilidad genera el máximo valor.
f. Los resultados obtenidos en el numeral anterior son comparados con una etiqueta previamente establecida, la cual indica el valor real de la clase, a la que pertenece la muestra [4] [21].

Como resultado de la anterior comparación se define el porcentaje de aciertos y errores que el sistema proporciona en la identificación, de las clases a las cuales pertenece la muestra.

2.4. **Regresión Logística para Múltiples Clases.**

La regresión logística multinomial (HOSMER & LEMESHOW, 1989) es utilizada en modelos con variable dependiente de tipo nominal con más de dos categorías (politómica) y es una extensión multivariante de la regresión logística binaria clásica [21].

Las variables independientes pueden ser tanto continuas (regresores) como categóricas (factores). Tradicionalmente las variables dependientes politómicas han sido modeladas mediante análisis discriminante pero, gracias al creciente desarrollo de las técnicas de cálculo, cada vez es más habitual el uso de modelos de regresión logística multinomial, ya implementados en paquetes estadísticos como S.A.S. (PROC CATMOD) o S.P.S.S. (NOMREG), debido a la mejor interpretabilidad de los resultados que proporciona.

2.4.1. **Formulación del modelo.**

Consideramos una variable aleatoria dependiente Y categórica nominal politómica con Soporte $(Y) = \{1,2,3\}$ y con probabilidades $p1 = p(Y = 1), p2 = p(Y = 2)$ y $p3 = p(Y = 3) = 1 - p1 - p2$.
Supongamos que queremos analizar el efecto que ejercen dos variables explicativas continuas $X1, X2$ sobre las probabilidades $p1\ y\ p2$ que caracterizan a la variable Y. Podemos redefinir a la variable Y mediante un vector $(Y1, Y2)$ construido de la siguiente forma [21] [8]:

$$(Y1, Y2) = \begin{cases} (1.0)\ si\ Y = 1 \\ (0,1)\ si\ Y = 2 \\ (0,0)\ si\ Y = 3 \end{cases} \qquad (2.9)$$

Las variables $Y1$ e $Y2$ tienen una distribución de Bernouilli con $E(Y1) = p1$ y $E(Y2) = p2$, al igual que la variable dependiente en una regresión logística binaria clásica. Obviamente estas dos variables no son independientes ya que $Cov(Y1, Y2) = -p1p2$.

Se formula el modelo multivariante definido por las siguientes ecuaciones [9]:

$$p1(X1, X2) = p1 = E(Y1) = \frac{\exp(Z1)}{1+\exp(Z1)+\exp(Z2)} \qquad (2.10)$$

$$p2(X1,X2) = p2 = E(Y2) = \frac{\exp(Z2)}{1+\exp(Z1)+\exp(Z2)} \qquad (2.11)$$

Donde $Z1 = \beta01 + \beta11 \cdot X1 + \beta21 \cdot X2$ y $Z2 = \beta02 + \beta12 \cdot X1 + \beta22 \cdot X2$, siendo $\beta01, \beta11, \beta21, \beta02, \beta12, \beta22$, parámetros que deseamos estimar.

Con el propósito de interpretar mejor los parámetros que aparecen en el modelo, podríamos reescribir éste de la siguiente forma [8]:

$$\frac{p1}{p3} = \exp(Z1) = \exp(\beta_{01}) \,.\, (\exp(\beta_{11}))^{x_1} \,.\, (exp(\beta_{21}))^{x_2} \qquad (2.12)$$

$$\frac{p2}{p3} = exp(Z2) = exp(\beta_{02}) \,.\, (exp(\beta_{12}))^{x_1} \,.\, (exp(\beta_{22}))^{x_2} \qquad (2.13)$$

Al cociente $p1/p3$ se le denomina '*odds*' de la categoría 1 respecto de la categoría 3 y se le representa por $O1\,(X1,X2) = O1$ (idem. para $O2$). De este modo puede observarse fácilmente que la razón de cambio en $O1$ cuando $X1$ se incrementa en una unidad manteniéndose constante $X2$ viene dado por[6]:

$$\exp(\beta_{11}) = \frac{O_1(X1+1,X2)}{O_1(X1,X2)} \qquad (2.14)$$

que recibe el nombre de '*odds-ratio*' de la categoría 1 respecto de la variable $X1$ y se representa por $OR1\,(X1)$ (idem. para $OR1\,(X2), OR2\,(X1)\; y\; OR2\,(X2)$.

Es interesante observar que estas 'odds-ratio' dependen de las unidades en que vengan medidas las variables regresoras (si multiplicamos $X1$ por 10, $OR1\,(X1)$ pasaría a ser $\sqrt[10]{\exp(\beta_{11})}$). Por tanto la importancia de cada variable regresora en el modelo debería medirse por el valor de la odds-ratio suponiendo estandarizada dicha variable [21].

Este es el motivo por el que se habla de las '*odds-ratio*' *estandarizadas en las variables regresoras*. Por ejemplo $OR1\,(X*1) = exp(\beta11 \cdot Sx1)$ siendo $Sx1$ la desviación típica muestral de la variable $X1$ (idem. para $OR1\,(X*2), OR2\,(X*1)\; y\; OR2\,(X*2)$). Cuanto más grande sea este valor más relevante es la variable dentro del modelo.

También interesa definir las *proporciones de cambio en las* '*odds*' con respecto a cada variable regresora que, por ejemplo, para $O1$ con respecto a $X1$, viene dada por:

$$\exp(\beta_{11}) - 1 = OR_1(X1) - 1 = \frac{O_1(X1+1,X2) - O_1(X1,X2)}{O_1(X1,X2)} \quad (2.15)$$

y que representaremos por: $OC1\ (X1)$ (idem. Para $OC1\ (X2), OC2\ (X1)\ y\ OC2\ (X2)$).

Otra formulación alternativa, y quizás más conocida, se obtiene tomando logaritmos en ambas ecuaciones del modelo:

$$\ln\left(\frac{p1}{p3}\right) = Z1 = \beta_{01} + \beta_{11}.X1 + \beta_{21}.X2 \quad (2.16)$$

$$ln\left(\frac{p2}{p3}\right) = Z2 = \beta_{02} + \beta_{12}.X1 + \beta_{22}.X2 \quad (2.17)$$

donde las expresiones del miembro izquierdo se denominan '*logits*' (al igual que en la regresión logística binaria) y los parámetros representan las *tasas de cambio en los* '*logits*' cuando una de las variables explicativas se incrementa en una unidad manteniéndose constante la otra [9] [5].

2.5. **Validación cruzada.**

La validación cruzada o cross-validation es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica.[1] Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados [20].

La validación cruzada proviene de la mejora del método de retención o *holdout method*. Este consiste en dividir en dos conjuntos complementarios los datos de muestra, realizar el análisis de un subconjunto (denominado datos de entrenamiento o *training set*), y validar el análisis en el otro subconjunto (denominado datos de prueba o *test set*), de forma que la función de aproximación sólo se ajusta con el conjunto de datos de entrenamiento y a partir de aquí calcula los valores de salida para el conjunto de datos de prueba (valores que no ha analizado antes)[6]. La ventaja de este método es que es muy rápido a la hora de computar. Sin embargo, este método no es demasiado preciso debido a la variación de los resultados

obtenidos para diferentes datos de entrenamiento. La evaluación puede depender en gran medida de cómo es la división entre datos de entrenamiento y de prueba, y por lo tanto puede ser significativamente diferente en función de cómo se realice esta división. Debido a estas carencias aparece el concepto de validación cruzada [21].

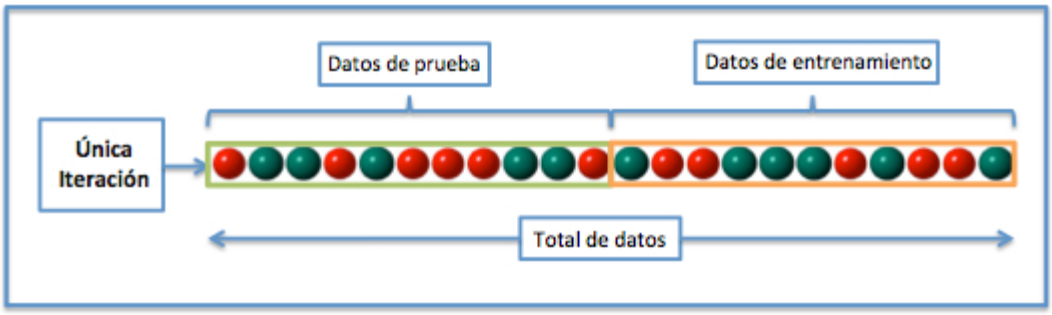

Figura 2.1. División de datos para una única iteración [22].

Se supone que se tiene un modelo con uno o más parámetros de ajuste desconocidos y unos datos de entrenamiento que queremos analizar. El proceso de ajuste optimiza los parámetros del modelo para que éste se ajuste a los datos de entrenamiento tan bien como pueda. Si se escoge una muestra independiente como dato de prueba (validación), del mismo grupo que los datos de entrenamiento, normalmente el modelo no se ajustará a los datos de prueba igual de bien que a los datos de entrenamiento. Esto se denomina sobreajuste y acostumbra a pasar cuando el tamaño de los datos de entrenamiento es pequeño o cuando el número de parámetros del modelo es grande. La validación cruzada es una manera de predecir el ajuste de un modelo a un hipotético conjunto de datos de prueba cuando no disponemos del conjunto explícito de datos de prueba [6].

En la validación cruzada de K iteraciones o *K-fold cross-validation* los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, pero aun así tiene una desventaja, y es que, a diferencia del método de retención, es lento desde el punto de vista computacional. En la práctica, la elección del número de iteraciones depende de la medida del conjunto de datos. Lo más común es utlizar la validación cruzada de 10 iteraciones (10-fold cross-validation)[22].

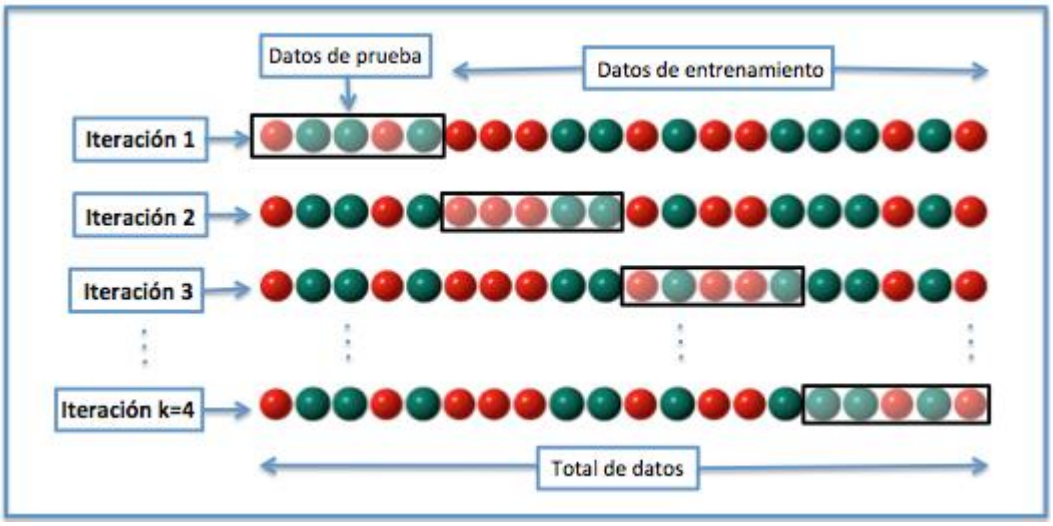

Figura 2.2. Validación cruzada para k iteraciones [22].

3. Metodología.

3.1. Base de datos.

La base de datos es proporcionada por la Universidad Tecnologica de Pereira y consiste en grabaciones de procedimientos quirúrgicos para seis pacientes con enfermedad de Parkinson avanzado cuyas edades estaban en el rango de 55 ± 6. Los pacientes firmaron un formulario de consentimiento informado. Las grabaciones de micro electrodos se obtuvieron utilizando el sistema ISIS MER (Inomed Medical GmbH)[15]. Las señales (MER) fueron etiquetadas por especialistas en neurocirugía y neurofisiología del Instituto de Parkinson y Epilepsia del Eje Cafetero, ubicados en la ciudad de Pereira. En total, hay 600 grabaciones de un segundo de duración, muestreado a 25 KHz con una resolución de 16 bits. Consideramos dos clases: 300 grabaciones pertenecen al Núcleo Subtalámico, y 300 grabaciones pertenecen a otras regiones del cerebro (Zona Incerta-ZI, Thalamus-TH, Substantia nigra reticulata-SNR.)

Mapa de los ganglios basales mostrando los objetivos de la cirugía: el globo pálido, el tálamo y el núcleo subtalámico

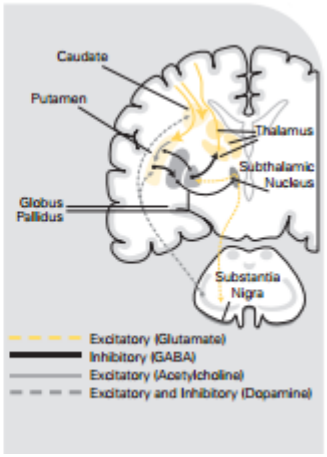

Figura 3.1. Zonas estimuladas mediante el procedimiento de estimulación cerebral profunda con microelectrodo.

Esta base de datos fue entregada luego de un proceso de extracción de características, usando dos métodos diferentes:

a. Transformada wavelet.

   Se aplicó la Transformada Wavelet (WT) para las señales MER utilizando la función de Daubechies 3 (3 db) con 2 niveles de descomposición en ventanas de 80 ms con traslape de 50%. A partir de los coeficientes de aproximación calculamos la media normalizada, el máximo absoluto, la kurtosis y la energía, obteniendo un total de ocho características por cada muestra (cuatro para cada nivel de descomposición).
   Al final del proceso de extracción de características sobre el total de muestras se generó una matriz de 8*600.
   Esta matriz se entregó en la forma de un arreglo tipo celda en formato .mat (dbsUTPWave.mat). En la primera componente de la celda, Data_MTL{1} se encuentra la matriz de 8 por 600(8 características por las 600 muestras obtenidas con MER), y en la segunda componente, Data_MTL{2}, hay un vector de 600 por 1 (correspondientes a cada una de las clases).

b. Intervalo *Inter Spike (*ISI*).*

   El ISI determina el tiempo de ocurrencia del potencial de acción para cada región del cerebro. Este método busca organizar la actividad eléctrica del cerebro de acuerdo a patrones comunes entre los potenciales de diferentes áreas.

Con este procedimiento se extrajeron 13 características del vector ISI que fueron:

Longitud promedio, desviación promedio, longitud máxima, longitud mínima, frecuencia instantánea media, desviación estándar de MIF,relación de contenido de baja frecuencia, relación de contenido de alta frecuencia, dispersión de ISI,índice de dispersión, índice de Burst, índice de asimetría, e índice de pausa.

Esta matriz se entregó en la forma de un arreglo tipo celda en formato .mat (dbsUTPSpikes.mat).

3.2 Algoritmos de clasificación.

Los métodos de clasificación aplicados sobre las bases de datos fueron el Método de Mínimos Cuadrados para Múltiples Clases, Clasificador Bayesiano basado en múltiples clases y Regresión Logística para Múltiples Clases, (algoritmos presentados en las secciones 2.1 2.2 y 2.3 respectivamente); para la aplicación de cada algoritmo de clasificación se hizo una partición de la base de datos con base en el criterio de validación cruzada expuesto en la sección 2.4.

Se realizó validación cruzada para k=5, k=10,k=15 y k=20 para cada uno de los tres algoritmos de clasificación lineal, mostrando el porcentaje de acierto en la predicción en cada caso.

4. Resultados

4.1. Método de Mínimos Cuadrados para Múltiples Clases.

Con base a lo expuesto en la sección 2.1 se implementó en simulación , usando el software Matlab, el esquema de clasificación basado en el método de mínimos cuadrados, tomando las base de datos ‘’dbsUTPSpikes.mat’’ y ‘’dbsUTPWave.mat’’ explicadas en la sección 3.1.

El esquema de validación cruzada se realizó para diferentes valores de k (sección 3.2).

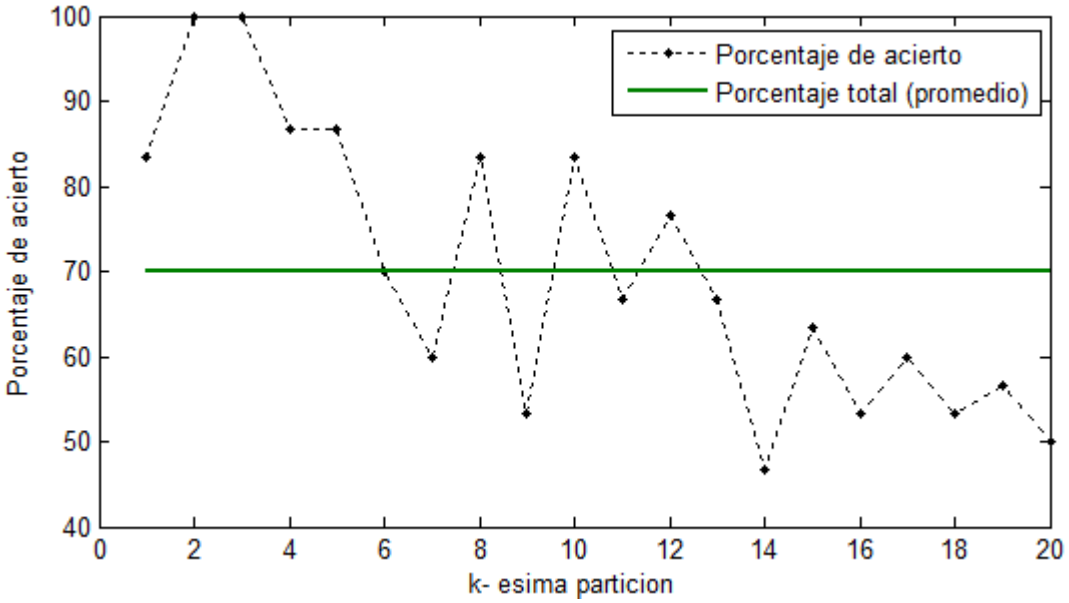

Figura 4.1. Porcentaje de acierto para k=20 usando la base de datos ‘’dbsUTPWave.mat’’.

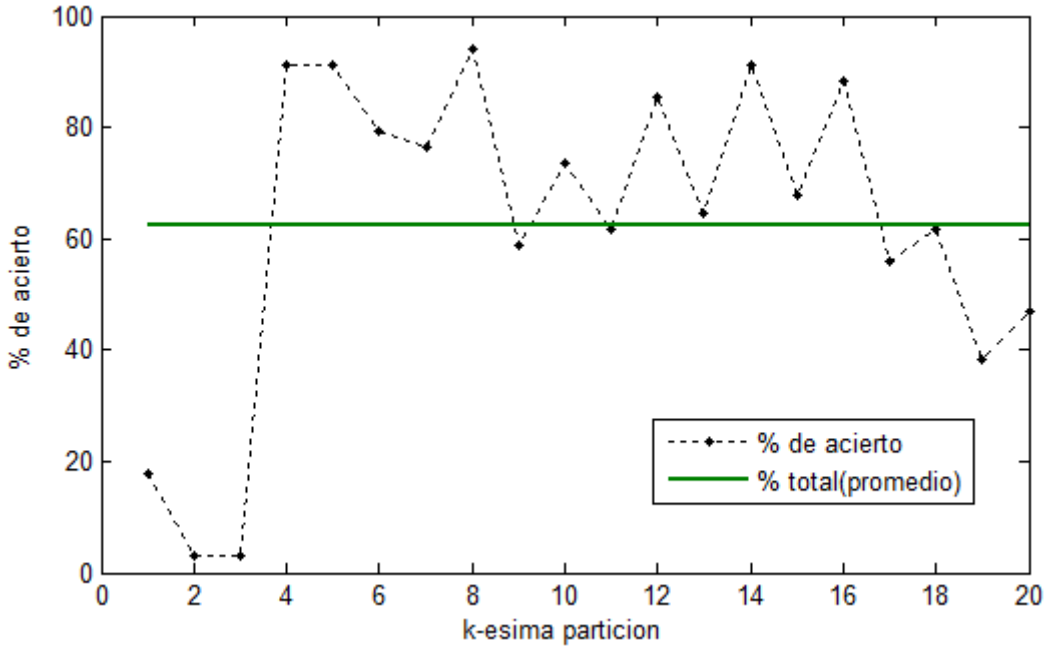

Figura 4.2. Porcentaje de acierto para k=20 usando la base de datos ‘’dbsUTPSpikes.mat’’.

Los resultados mostrados en las gráficas 4.1 y 4.2 muestran los porcentajes de aciertos usando validación cruzada con k=20 para ambas bases de datos.

4.2 Clasificador Bayesiano.

Con base a lo expuesto en la sección 2.2 se implementó en simulación , usando el software Matlab, el esquema de clasificación basado en el método de inferencia Bayesiana, tomando las base de datos ‘’dbsUTPSpikes.mat’’ y ‘’dbsUTPWave.mat’’ explicadas en la sección 3.1.

El esquema de validación cruzada se realizó para diferentes valores de k (sección 3.2).

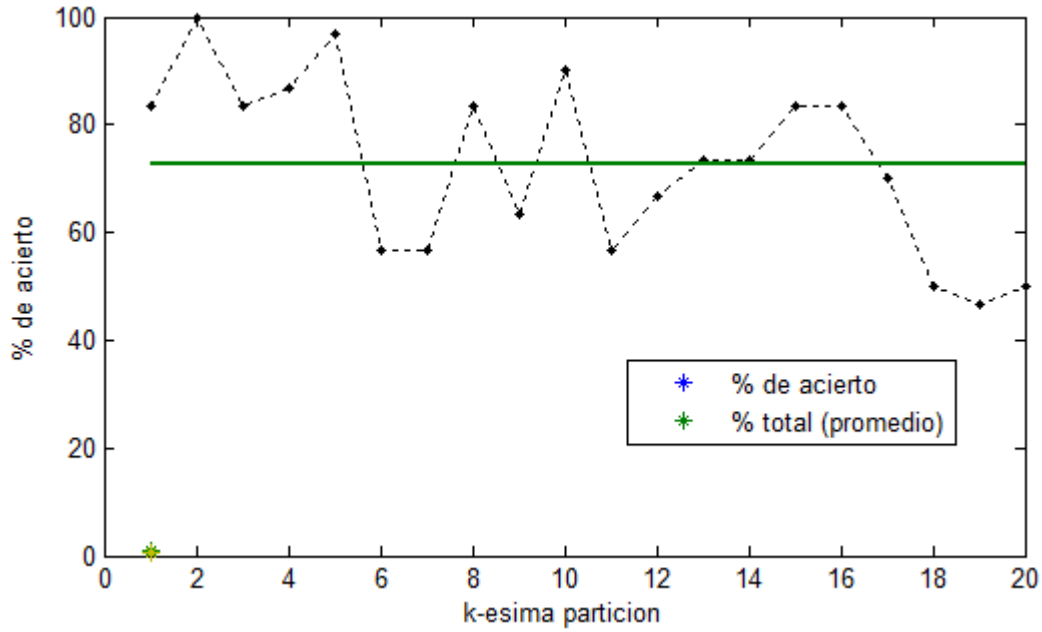

Figura 4.3. Porcentaje de acierto para k=20 usando la base de datos ‘’dbsUTPWave.mat’’.

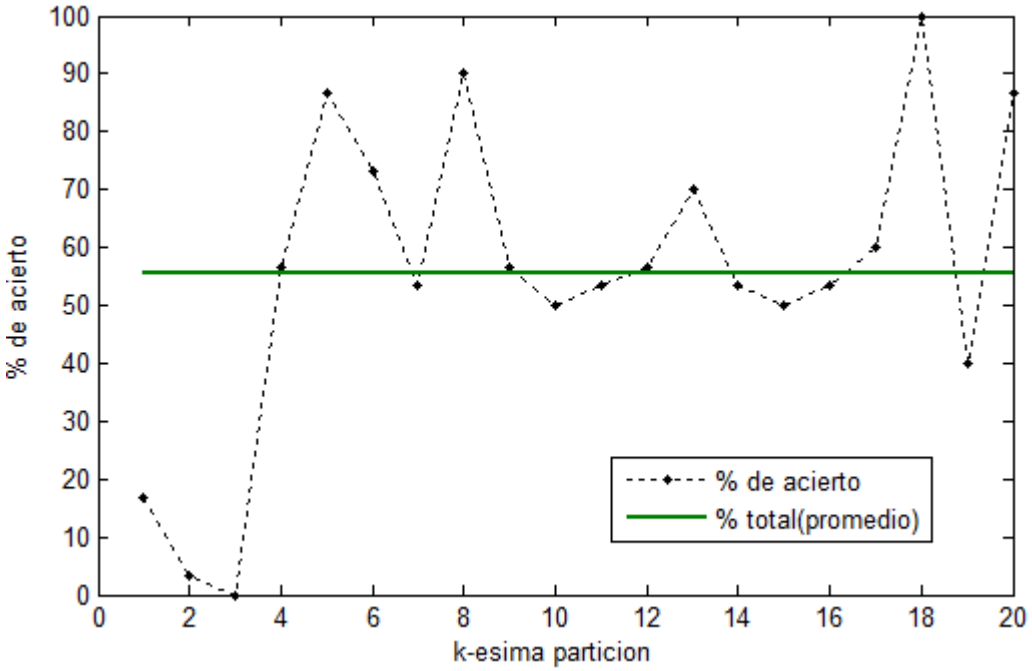

Figura 4.4. Porcentaje de acierto para k=20 usando la base de datos ‘’dbsUTPSpikes.mat’’.

4.2. Regresión Logística.

Con base a lo expuesto en la sección 2.3 se implementó en simulación , usando el software Matlab, el esquema de clasificación basado en el método de Regresión Logística, tomando las base de datos ‘’dbsUTPSpikes.mat’’ y ‘’dbsUTPWave.mat’’ explicadas en la sección 3.1.

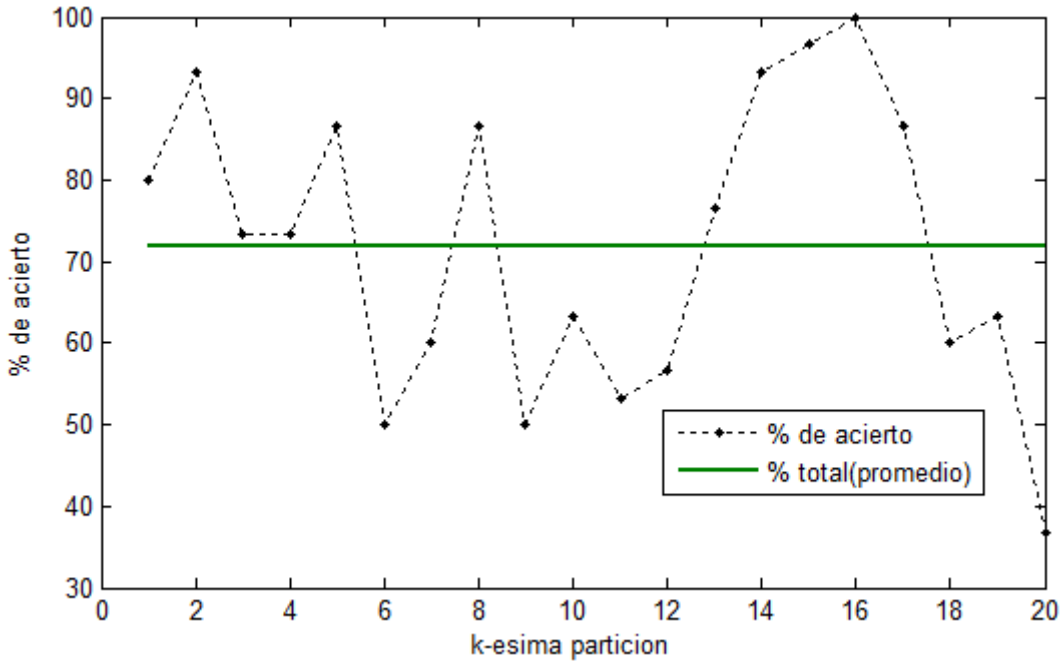

Figura 4.5. Porcentaje de acierto para k=20 usando la base de datos ‘’dbsUTPWave.mat’’

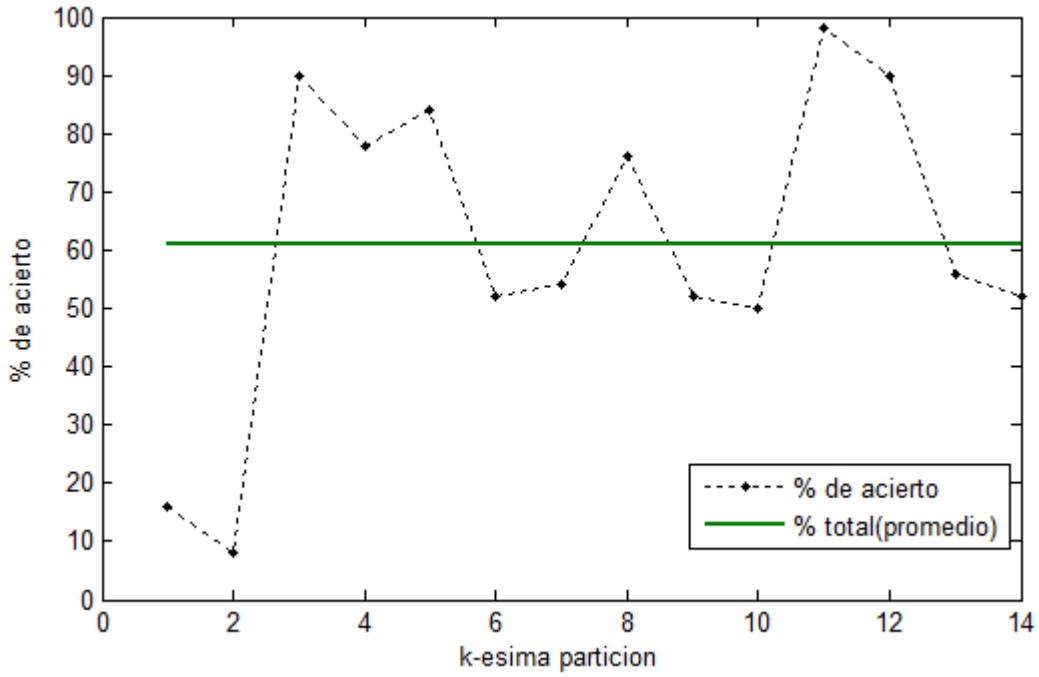

Figura 4.6. Porcentaje de acierto para k=14 usando la base de datos ‘’dbsUTPSpikes.mat’’.

Método	K	ISI	WAVELET TRANSFORM
		$\%\ de\ acierto$	$\%\ de\ acierto$
$Bayes\ lineal$	5	44	69.17
	10	46.67	71.0
	15	52.17	71.83
	20	55.55	72.67
$Mínimos\ cuadrados$	5	59	65.83
	10	60.86	65.50
	15	60.91	69.82
	20	62.50	70.0
$Regresión\ logística$	5	64.0	69.0
	10	65.14	74.33
	15	64.7	72.0
	20	**65.0**	**74.67**

Tabla 4.1. Porcentaje de aciertos para los tres métodos de clasificación lineal, aplicado a la base de datos DB-UTP caracterizada con transformada Wavelet y con ‘Inter Spike Interval’ (ISI).Se realiza validación cruzada de K particiones.

3. **Trabajos futuros**

Como trabajo futuro se podría plantear desarrollar un modelo de clasificación no paramétrico, por ejemplo uno basado en modelos ocultos de Markov, donde se plantearía un modelo oculto de Markov por cada clase, de tal forma que en la etapa de validación de los modelos se calcularía la probabilidad de cada modelo dada una secuencia de observación, y el valor de la probabilidad máxima determinaría la predicción.

De igual forma que en los modelos de clasificación lineal se tiene la etapa de entrenamiento, en la cual cada observación la constituye un vector de n características, donde dicha etapa se llevaría a cabo por el algoritmo de Baum-Welch, que ajustaría los parámetros del modelo de Markov para maximizar la probabilidad de la secuencia de observación.

Cada modelo oculto de Markov se podría generar con base en modelos continuos, donde las observaciones serian modeladas como funciones de densidad de probabilidad continuas, con una mezcla de n gaussianas por estado y N estados por modelo.

A priori un modelo tal presenta la ventaja de estimar la probabilidad de cada modelo basado solo en las observación , lo que permitiría dado un entrenamiento adecuado una alta tasa de efectividad , mayor que la de los modelos de regresión lineal.

4. **Conclusiones**

Los modelos de regresión lineal presentan un porcentaje de aciertos muy bajo debido a la distribución de los datos que se alejan de distribuciones lineales; por ejemplo en el caso del modelo de clasificación basado en mínimos cuadrados se presenta como un sistema poco robusto, ya que los puntos lejanos a la frontera de decisión presentan demasiado peso relativo con respecto a los demás puntos con lo cual el margen de error en la predicción se hace mayor (puntos lejanos a la frontera tienen demasiado peso relativo), en otras palabras un clasificador lineal no puede clasificar sin errores un conjunto no linealmente separable.

De forma general el método de clasificación que presenta mayor número de aciertos es el método de regresión logística, efectividad que de forma generalizada aumenta al incrementar el número de particiones en la validación cruzada, llegando a estabilizarse para un valor de k=20, donde a falta de un mayor número de datos se hace imposible el aumento de las particiones sin perder generalización el entrenamiento de los clasificadores.

5. Bibliografía

[1] R. O. Duda, Pattern Classification, New York: Wiley Interscience. Pub., 2002.

[2] Q. Cazorla. Miguel, Un enfoque bayesiano para la extracción de características y agrupamiento en visión artificial, Alicante: Universidad de alicante. Pub., 2000.

[3] E. Morales, Descubrimiento de conocimiento en bases de datos, Moterrey: Instituto Tecnológico de Monterrey.Pub., 2001.

[4] J.H. Hong, J.K. Min, U.G. Cho, Sung-Bae Cho, "Fingerprint classification using one-vs-all support vector machines dynamically ordered with naïve Bayes classifiers", Pattern Recognition, Vol. 41, No. 4, 2008, pp. 662 – 671.

[5] C.M. Giorgio, M. Zaffalon, "JNCC2: An extension of naive Bayes classifier suited for small and incomplete data sets", Environmental Modelling & Software, Vol. 23, No. 1, 2008, pp. 960 –961.

[6] MENARD, S.; 2000. Coefficients of Determination for Multiple Logistic Regression Analysis. The American Statistician 54: 17-24.

[7] H. Vargas , A .Orozco and M. A. Alvarez. Multi-patient learning increases accuracy for Subthalamic nucleus identification in deep brain stimulation, 34th Annual International Conference of the IEEE EMBS, 2012

[8] NORTH, M.P. & REYNOLDS, J.H.; 1996. Microhabitat analysis using radiotelemetry
locations and polytomous logistic regression. J. Wild. Manage. 60(3): 639-653.

[9] Pizarro J, Guerrero E and Galindo P (2002). Multiple comparison procedures applied to model selection. Neurocomputing, 48:155– 173.

[10] Rodriguez A, Delgado E, Orozco A, Castellanos G and Guijarro E (2008). Nonlinear dynamics techniques for the detection of the brain areas using MER signals. In IEEE International Conference on BioMedical Engineering and Informatics, 2:198–202.

[11] J. Chen, H. Huang, et All, “A Selective Bayes Classifier for Classifying Incomplete Data Based on Gain Ratio”, Knowledge

[12] 11. Gemmar P, Gronz O, Henrichs T and Hertel F (2008). Advanced methods for target navigation using microelectrode recordings in stereotactic neurosurgery for deep brain stimulation. In CBMS 08: Proceedings of the 21st IEEE International Symposium on Computer-BasedMedical Systems, IEEE Computer Society, 21:99–104.

[13] R. A. Santiago,J McNames,K. Burchiel,G. Lenderis. Developments in Understanding neuronal spike trains and fuctional specializations in brain regions. Neural Networks 16,2003 pag 601-607.

[14] J. Guridi, M .Rodríguez-Oroz, and M. Manrique,“Tratamiento quirúrgico de la enfermedad de parkinson,” Neurocirugía, no. 15, pp. 5–16, 2004..

[15] Z. Israel and K. J. Burchiel, Microelectrode Recording in Movement Disorder Surgery, S. Liu, Ed.Thieme, 2004.

[16] A. Orozco, M. Alvarez, E. Guijarro, G. Castellanos. “Identification of Spike Sources using Proximity Analysis through Hidden Markov Models”. Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, 2006

[17] Roberto A. Santiago, James McNames, Kim Burchiel, George G. Lendaris. An Automated Method for Neuronal Spike Source Identification.

[18] C. Bishop.“Patter Recognition and Machine Learning.

[19] E. Pekalska and R. Duin. Dissimilarity representations allow for building good classifiers. Pattern Recognition Letters, vol. 23, 2002, pp 943-956.

[20] A. Orozco, M. Alvarez, E. Guijarro, G. Castellanos. “Identification of Spike Sources using Proximity Analysis through Hidden Markov Models”. Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, 2006

[21] R.D.Pinzon, A. Orozco, H. Carmona-Villada, C.G.Castellanos, “Towards High Accuracy Classification of MER Signals for Target Localization in Parkinson’s Disease” 32nd Annual International Conference of the IEEE EMBS,2010

[22]Wikipedia, La enciclopedia libre. (Sitio Web). Disponible: