

Evaluación del clasificador Naïve Bayes como herramienta de diagnóstico en Unidades de Cuidado Intensivo

Evaluation of the Naïve Bayes classifier as a tool of diagnosis in the intensive care units

Javier A. Chaparro, Beatriz Giraldo, Susana Rodón

Resumen

En el proceso de diagnóstico médico se busca identificar el estado de salud de un paciente. Sin embargo, la complejidad fisiológica del ser humano genera una amplia gama de condiciones difíciles de establecer por el personal médico en pacientes patológicos. En las unidades de cuidado intensivo, donde habitualmente se utiliza la ventilación asistida, se presenta un alto grado de incertidumbre a la hora de identificar el momento adecuado de retirar el ventilador mecánico. Teniendo en cuenta esta problemática, presentamos el diseño y evaluación de una herramienta basada en inferencia bayesiana que brinda información al médico sobre la viabilidad de tener una extubación exitosa. Para su diseño y validación se utilizó una base de datos con señales de flujo respiratorio compuesta por 98 pacientes exitosamente extubados, y 38 que fracasaron. Las señales fueron caracterizadas inicialmente con un grupo de series temporales, y posteriormente, el patrón respiratorio en los dos grupos de estudio fue analizado con técnicas estadísticas y de modelado autorregresivo. De este último proceso se derivó un grupo de variables con las cuales se diseñó un clasificador tipo Naive Bayes. Los resultados fueron medidos en función de la exactitud, sensibilidad y especificidad del clasificador, logrando un 78% en la primera medida y un 75% y 74% en las dos segundas.

Palabras claves: flujo respiratorio, diagnóstico clínico, extubación de pacientes, clasificador bayesiano.

Abstract

In the process of medical diagnosis is to identify the health status of a patient. However, the complexity of the human physiological generates a variety of difficult conditions set by the medical staff in each patient. A high degree of uncertainty comes to identifying the right time to begin the withdrawal of mechanical ventilation in Intensive Care Units. Considering this problem, we present the design and evaluation of a Bayesian inference based tool that provides information to the physician about the feasibility of having a successful extubation. For design and validation database with respiratory flow signals composed successfully extubated 98 patients, 38 who failed was used. The signals were initially characterized with a group of time series, and then the breathing pattern in the two study groups was analyzed with statistical techniques and autoregressive modeling. This latter process a group of variables with which a Naive Bayes classifier type design was derived. Results were measured in terms of exactitude, sensitivity and specificity of the classifier, achieving 78% in the first step and 75% and 74% in the second two.

Key words: Airflow, clinical diagnosis, patient extubation, Bayesian classifier.

Recibido: Enero 27 de 2013 **Aprobado:** Diciembre 20 de 2013

Tipo de artículo: Investigación científica y tecnológica terminada.

Afiliación Institucional de los autores: Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, Colombia e Instituto de Bioingeniería de Cataluña (IBEC), España.

Los autores declaran que no tienen conflicto de interés.

Introducción

El diagnóstico médico es el proceso de identificar las enfermedades o condiciones que tiene el paciente, y descartar otras que no tiene, mediante evaluación de los signos, síntomas, y resultados de varias pruebas diagnósticas [1]. La precisión de estas pruebas es la capacidad de discriminar entre las diferentes alternativas de estados de salud [2]. Una buena calidad de la prueba, medida en términos de especificidad (probabilidad que un sujeto sano tenga un resultado negativo) y sensibilidad (probabilidad que un sujeto enfermo tenga un resultado positivo), permite reducir la incertidumbre que se origina, no solamente por el resultado de la prueba, sino también por experiencia del personal médico y la veracidad de la información suministrada por el paciente, entre otros factores [3].

La incertidumbre en el diagnóstico en la práctica médica causa estrés en pacientes y médicos. Los factores que contribuyen a esta incluyen la variabilidad biológica de los pacientes, error en la interpretación de la prueba, diferentes valores y opiniones de pacientes y médicos, y la incertidumbre en torno a la toma de decisiones. A pesar de que la incertidumbre en la práctica no puede ser completamente eliminada, numerosas estrategias que pueden adoptarse para reducirla y aumentar la confianza de los pacientes. Estas incluyen la aplicación de la mejor información disponible basada en la evidencia, junto con la observancia de las prácticas clínicas básicas, incluyendo la historia minuciosa, tomando con exclusión de diagnósticos preocupantes, y la participación en la toma de decisiones compartida [4].

La inferencia bayesiana, basada en la aplicación sistémica del teorema de Bayes, permite desarrollar sistemas de apoyo al diagnóstico clínico que brindan información cuantitativa al personal médico y mitigan la incertidumbre en este proceso. Actualmente, la teoría bayesiana está demostrando ser de gran utilidad en ciertas estimaciones basadas en el conocimiento subjetivo a priori, además, el hecho de permitir hacer una revisión de esas estimaciones en función de la evidencia empírica, está abriendo nuevas formas de hacer conocimiento. Una aplicación de esto son los clasificadores bayesianos que son frecuentemente usados en medicina [5].

En este trabajo se presenta el diseño y evaluación de un clasificador bayesiano, tipo Naive. El propósito de este clasificador es analizar y clasificar la información de señales respiratorias de pacientes ingresados en Unidades de Cuidado Intensivo (UCI), y estimar la viabilidad de una extubación exitosa. La problemática que se aborda es el alto riesgo para la salud del paciente que se presenta a la hora de iniciar la extubación; se ha reportado que hasta un 25% de los pacientes requieren reintubación y por lo tanto un aumento en el probabilidad de muerte [6][7][8]. Para llegar a este diseño se utilizó una base de datos formada por señales de flujo respiratorio de 94 pacientes satisfactoriamente extubados y 38 que no.

Las señales de flujo respiratorio fueron caracterizadas con un grupo de series temporales, y posteriormente estas series fueron analizadas con métodos estadísticos y de modelado autorregresivo AR y ARMA. Dicho análisis permitió identificar un grupo de variables utilizadas para diseñar el clasificador tipo Naive Bayes. El resultado mostro un clasificador con una exactitud del 78%, una sensibilidad y especificidad 75% y 74% respectivamente, con el hecho particular de utilizar únicamente la información relacionada con el tiempo de espiración.

Materiales y Métodos

Pacientes

El estudio se realizó con la base de datos denominada WEANDB, formada por señales respiratorias de pacientes conectados al ventilador mecánico mediante un tubo endotraqueal. A partir del resultado del test de tubo en T, los pacientes fueron clasificados en dos grupos: grupo éxito con 94 pacientes que superaron satisfactoriamente la prueba y recuperaron la respiración espontánea, y grupo fracaso, con 38 pacientes que no superaron el test y fueron reconectados al ventilador. Estos pacientes fueron registrados en los Departamentos de Cuidados Intensivos del Hospital de la Santa Creu i Sant Pau, Barcelona, y del Hospital de Getafe, España, de acuerdo con los protocolos aprobados por sus comités éticos.

Con esta base de datos se elaboraron las siguientes series temporales: tiempo de inspiración (T_i), tiempo de espiración (T_e), duración del ciclo respiratorio (T_{tot}), volumen

tidal (V_T), fracción inspiratoria (T_V/T_{tot}), flujo inspirado medio (V_T/T_T), y el índice de respiración superficial (f/V_T), siendo f la frecuencia respiratoria. Una descripción más detallada de estas series se encuentra en [8].

Caracterización del patrón respiratorio

A continuación se exponen los fundamentos de las técnicas de modelado utilizadas para analizar las series temporales en este trabajo:

Modelo Autorregresivo (AR). Este modelo de la señal permite establecer la muestra $x(n)$ en función del historial de la misma más un componente de error $e(n)$ de la forma descrita en (1).

$$x(n) = a_1 x(n-1) + \dots + a_p x(n-p) + e(n) \quad (1)$$

donde p es el orden del modelo, a_1, \dots, a_p son sus coeficientes reales, y $e(n)$ es un proceso aleatorio gaussiano con media cero y varianza σ^2 . Se asume que la secuencia $x(n)$ es de media y varianza ergódica. En este trabajo se utilizó el orden del modelo, el primer coeficiente y el error final de predicción del mismo (FPE) para caracterizar las series temporales [9].

Modelo Autorregresivo de Promedio Móvil (ARMA). El modelo AR con un componente móvil se convierte en un modelo ARMA que se pueden describir de acuerdo con la ecuación (2).

$$\begin{aligned} x(n) + a_1 x(n-1) + \dots + a_p x(n-p) \\ e(n) + b_1 e(n-1) + \dots + b_q e(n-q) \end{aligned} \quad (2)$$

siendo p y q son los órdenes del proceso estimado mediante el criterio de Akaike [10], a_1, \dots, a_p y b_1, \dots, b_q son los coeficientes reales del proceso [9]. En este trabajo se utilizó los órdenes p y q , además del error final de predicción, para caracterizar las series temporales.

Inferencia bayesiana

Los métodos de clasificación bayesiana se basan principalmente en el Teorema de Bayes, a continuación se expone este teorema y el clasificador de naïve bayes.

Teorema de Bayes. Este se refiere al cálculo de la probabilidad condicional del evento A dado que ha ocurrido el evento B, su forma general es: Si A_1, A_2, \dots, A_n son eventos exhaustivos y exclusivos tales que $P(A_i) > 0, \forall i = 1, 2, \dots, n$, sea B un evento cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$, la probabilidad $P(A_i|B)$ viene dada por la expresión (3) [11].

$$\begin{aligned} p(A_i|B) &= \frac{p(A_i, B)}{p(B)} = \frac{p(A_i)p(B|A_i)}{p(B)} \\ &= \frac{p(A_i)p(B|A_i)}{\sum_{k=1}^n P(B|A_k)P(A_k)} \end{aligned} \quad 3$$

Clasificador Naïve Bayes. Los clasificadores bayesianos son modelos particulares de las redes bayesianas que se aplican en la tarea de clasificación supervisada con datos discretos. Se han propuesto diferentes tipos dependiendo de las relaciones de dependencia de las variables contempladas entre variables predictoras, entre ellos se tienen: naïve bayes, naïve bayes selectivo, semi naïve Bayes, k-dependiente, y naïve bayes aumentado a árbol (TAN) [12].

El clasificador de naïve bayes es el modelo de clasificación construido asumiendo que todas las variables predictoras (variables de los síntomas) son independientes dada la variable de clase C (variable a diagnosticar). La búsqueda del diagnóstico más probable a posteriori, c^* , una vez conocidos los síntomas de un paciente $X = (X_1, \dots, X_n)$ puede plantearse como la búsqueda del estado de la variable C con la mayor probabilidad a posteriori como se muestra en (4).

$$c^* = \arg \max_c p(C = c | X_1, \dots, X_n) \quad (4)$$

Este cálculo puede hacerse con el Teorema de Bayes de la forma que aparece en (5).

$$p(C = c | X_1, \dots, X_n) = \frac{p(C = c)p(X_1, \dots, X_n | C = c)}{p(X_1, \dots, X_n)} \quad (5)$$

En la práctica el denominador es constante, por este motivo se puede despreciar y así estudiar el numerador como una distribución de la forma descrita en (6).

$$p(C = c | X_1, \dots, X_n) = p(C = c)p(X_1, \dots, X_n | C = c) \quad (6)$$

El factor $p(X_1, \dots, X_n / C=c)$ se puede simplificar utilizando la regla de la cadena (7).

$$\begin{aligned} p(X_1 | X_2, \dots, X_n, C=c) \\ p(X_2 | X_3, \dots, X_n, C=c) \quad (6) \\ p(X_n | C=c) \end{aligned}$$

Si las variables X_i son independientes para todo $i \neq j$, se puede llegar a (8).

$$p(X_i | C=c, X_j) = p(X_i | C=c) \quad (8)$$

Por lo tanto, la probabilidad compuesta $p(C=c / X_1, \dots, X_n)$ puede expresarse finalmente como aparece en (9).

$$\begin{aligned} p(C=c) p(X_1 | C=c) p(X_2 | C=c), \dots, p(X_n | C=c) \\ p(C=c) \prod_{i=1}^n p(X_i | C=c) \quad (9) \end{aligned}$$

La expresión $p(X_i / C=c)$ denota la función densidad de la variable X_i condicionada a que el valor del diagnóstico sea c . Esta función puede ser ajustada a una normal, gamma, lognormal entre otras. Lo más acostumbrado es utilizar una variable aleatoria normal para modelar el comportamiento de la variable X_i para cada valor de C . De esta forma finalmente el clasificador se reduce a la siguiente expresión [13]:

$$c^* = \arg \max_c \left\{ p(C=c) \prod_{i=1}^n p(X_i | C=c) \right\} \quad (10)$$

Para efectuar el entrenamiento y la validación se utilizó la técnica de validación cruzada. La calidad del clasificador bayesiano se evaluó con el grupo de entrenamiento en función de tres medidas: la exactitud (Acc) (11), la sensibilidad (Sn) (12) y la especificidad (Sp) (13). Estas medidas se construyen a partir de los siguientes resultados de clasificación: Vp verdaderos positivos, Fp falsos positivos, Vn verdaderos negativos, y Fn falsos negativos [14].

$$Acc = \frac{Vp+Vn}{Vp+Fp+Vn+Fn} \quad (11)$$

$$Sn = \frac{Vp}{Vp+Fn} \quad (12)$$

$$Sp = \frac{Vn}{Vn+Vp} \quad (13)$$

Resultados

El objetivo del clasificador es establecer, si dadas las condiciones de un grupo de variables predictoras en un instante determinado, es posible tener un éxito o un fracaso en el proceso de retirada del ventilador mecánico de los pacientes en UCI. De acuerdo a esto, se definieron las siguientes variables:

Variables predictoras (X). Después de evaluar el desempeño de las variables obtenidas de los modelos AR y ARMA, en el diseño de clasificadores tipo Naive Bayes, se identificó un grupo que permitió obtener la mejor exactitud sensibilidad y especificidad:

- X_1 = Promedio de la serie T_E
- X_2 = Rango intercuartil serie T_E
- X_3 = Orden del modelo AR serie T_E
- X_4 = Primer coeficiente del modelo AR serie T_E
- X_5 = Error final de predicción modelo AR serie T_E
- X_6 = Error final de predicción modelo ARMA serie T_E

Variable de la clase de salida (C). El clasificador debe predecir, a partir del grupo de variables X_i , un éxito o un fracaso en la extubación: es decir, $C = \{\text{éxito, fracaso}\}$.

Diseño del clasificador naïve bayes. Las funciones de densidad de probabilidad de $P(X_i / C=c)$, para $i=1, \dots, 6$ y $C = \{\text{éxito, fracaso}\}$, fueron ajustadas a variables aleatorias normales con media y varianza. Las características de cada una de estas funciones se presentan en la Tabla I.

Tabla I. Funciones de probabilidad $P(X_i | C=c)$

	$P(X_i C=\text{éxito})$		$P(X_i C=\text{fracaso})$	
X_1				
X_1	1.698	0.527	1.260	0.355
X_2	0.380	0.236	0.264	0.145
X_3	40.085	38.531	23.947	39.634
X_4	0.360	0.162	0.632	0.289
X_5	1.016	7.016	0.276	0.491
X_6	0.994	6.738	0.2122	0.368

A continuación se presentan, a modo de ejemplo, las gráficas de las funciones de distribuciones para X_1 (Figura 1) y X_2 (Figura 2).

Figura 1. Función de distribución de $P(X_1|C=éxito)$

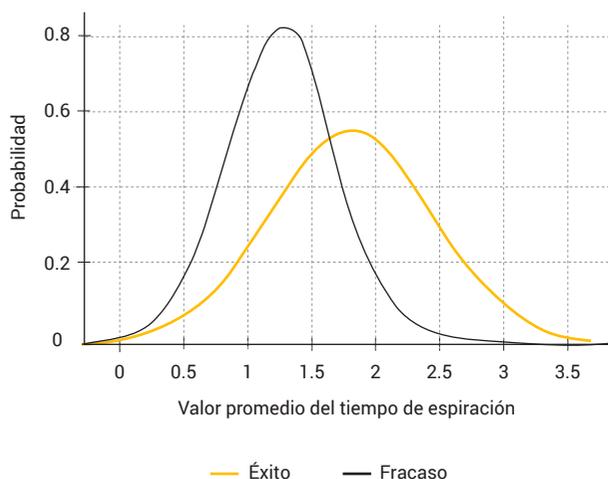
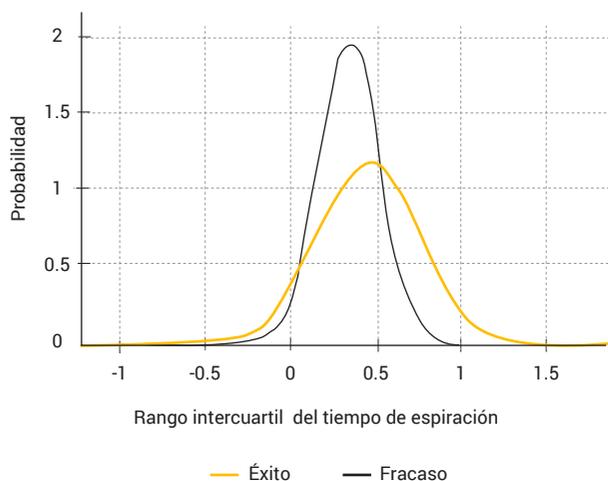


Figura 2. Función de distribución de $P(X_2|C=éxito)$



Los valores de $P(C=éxito)$ y $P(C=fracaso)$ son de $94/132 = 0.71$ y $32/132 = 0.29$ respectivamente. Teniendo completos los elementos de la ecuación (10), que corresponden al clasificador de naïve bayes, se procede a su validación. Los resultados mostraron una exactitud (Acc) del 78%, una sensibilidad (sn) de 75% y una especificidad (sp) de 74%. A continuación se hace la prueba del clasificador con los datos de la Tabla II para verificar su funcionamiento.

Tabla II. Ejemplo de aplicación del clasificador naïve Bayes

Variable (X_i)	Valor de X_i	$P(X_i c=éxito)$	$P(X_i c=fracaso)$
X_1	2	0.457	0.037
X_2	0.5	1.06	0.214
X_3	50	0.007	0.002
X_4	0.5	1.212	0.357
X_5	0.1	0.034	0.001
X_6	0.1	0.035	0.001

Según la ecuación (10), la probabilidad que $C=éxito$, es de

$$= 0.71 * 0.457 * 1.06 * 0.007 * 1.212 * 0.034 * 0.035 = 3.472 * 10^{-6} \tag{14}$$

Y el resultado para el caso que $C=fracaso$ es de:

$$= 0.29 * 0.37 * 0.214 * 0.002 * 0.357 * 0.001 * 0.001 = 1.634 * 10^{-6} \tag{15}$$

De acuerdo a estos resultados, como la probabilidad de un éxito es mayor a la de un fracaso, la clasificación del paciente según los valores utilizados en la Tabla II, será de éxito en al terminar el proceso de extubación.

Conclusiones

La etapa de caracterización de las señales de flujo respiratorio permitió tener un grupo de series temporales utilizado en la fase posterior de diseño del clasificador naïve bayes. A pesar de contar con un grupo de siete series, únicamente las variables correspondiente a la serie tiempo de espiración fueron utilizadas en el clasificador.

El análisis de las series temporales, con técnicas estadísticas y de modelado autorregresivo, permitió obtener un grupo de variables para analizar los grupos de éxito y fracaso en la extubación. Desde el punto de vista estadístico, se obtuvieron los valores promedio y el rango intercuartil. Para el modelado autorregresivo se estableció el orden, el primer coeficiente y el error final de predicción de cada modelo. Se estimaron siete variables por cada serie temporal, para un total de cuarenta y nueve variables que fueron reducidas a seis (X_1, \dots, X_6), mediante la identificación de

aquellas que mostraron las mejores diferencias estadísticas entre los grupos de éxito y fracaso.

Las funciones de probabilidad para $P(X_i/C=\text{éxito})$ y $P(X_i/C=\text{fracaso})$, para $i=1,\dots,6$, fueron estimadas aproximándolas a una distribución normal. Durante el trabajo se evaluaron el ajuste a distribuciones de probabilidad tipo Gamma, Lognormal y Weibull sin obtener mejores resultados.

Los modelos autorregresivos puros (AR) y de media móvil (ARMA) permitieron analizar las series temporales que caracterizaron los grupos de pacientes de éxito y fracaso en el proceso de retirada del ventilador mecánico. Esta información, complementada con estadísticos de tipo promedio y rango intercuartil, facilitó la elaboración de un grupo de variables utilizadas en el diseño del clasificador. El clasificador naïve bayes presentó una exactitud del 78%, una sensibilidad de 75% y una especificidad de 74%, estos resultados son aceptables pero se pueden evaluar los otros tipos de clasificadores bayesianos buscando mejorar estos resultados. También se pudo demostrar, al final de este trabajo, la sencillez de diseñar e implementar un clasificador naïve bayes para aplicaciones biomédicas. La tarea de clasificación se reduce simplemente a seleccionar la clase con la probabilidad más alta de ocurrencia.

Bibliografía

- [1] Xiao-Hua Zhou, Nancy A. Obuchowski, Donna K. McClish. "Statistical Methods in Diagnostic Medicine". Wiley, 2011
- [2] Zweing, M. H., Campbell, "Receiver-Operating Characteristic Plots: a fundamental evaluation tool in clinical medicine". 1993, Clinical Chemistry, pp. 1272-1276.
- [3] Bermejo Fraile B. "Epidemiología clínica aplicada a la toma de decisiones en medicina". Departamento de Salud. Gobierno de Navarra; 2001
- [4] Ghosh A.K. "Understanding Medical Uncertainty: A Primer for Physicians". Septiembre 2004, Journal of the Association of Physicians of India, Vol. 52.
- [5] Silva L., Muñoz A. "Debate sobre métodos frecuentistas vs bayesianos". Santiago de Compostela : s.n., 2000. XVII Reunión Científica SEE. pp. 482-494.
- [6] M.J. Tobin, "Advances in mechanical ventilator", N. Engl. J. Med., Vol. 344, no. 26, pp.1986-1996, 2001.
- [7] Bruce, E. Bruce, Measures of Respiratory Pattern Variability, Bioengineering Approaches to Pulmonary Physiology and Medicine, pp. 149-159. 1996.
- [8] Chaparro, Javier y Giraldo, Beatriz, Análisis y clasificación del patrón respiratorio de pacientes en proceso de retirada del ventilador mecánico, Revista de Ingeniería Biomédica, Vol. 10, pp. 42-48, 2010.
- [9] Box G., Jenkins G., Reinsel G. Time Series Analysis, Forecasting and Control. s.l. : Prentice Hall International Inc., 1994.
- [10] Daniel Peña y Gonzalo Arnáiz, "Criterios de selección de modelos ARIMA", Trabajos de Estadística y de Investigación Operativa, Vol. 32, no. 1, pp 70 – 93, 1981.
- [11] Fernández, E "Análisis de clasificadores bayesianos". Trabajo Final de Especialidad en Ingeniería de Sistemas Expertos. Escuela de Postgrado. Instituto Tecnológico de Buenos Aires. 2004.
- [12] Vega, D. A., "Clasificadores Bayesianos en la Selección Embrionaria en Tratamientos de Reproducción". San Sebastián, España: Universidad del País Vasco. 2008.
- [13] Barrientos R., Cruz N., Acosta H., et al, "Evaluación del Potencial de Redes Bayesianas en la Clasificación de Datos Médicos", Revista Médica de la Universidad de Veracruzana. Vol. 1 no. 8. 2008.
- [14] Chaparro J., Giraldo B., "Performance of Respiratory Pattern in Classifiers for Predict Weaning Process", Memorias 34th Conferencia Annual de la Sociedad EMB-IEEE. San Diego California. 2012.

Autor



Javier A. Chaparro Ph.D

Profesor Asociado de la Escuela Colombiana de Ingeniería Julio Garavito en el Decanatura de Ingeniería Electrónica. Ingeniero Electrónico y Especialista en Automatización Industrial de la Universidad Pedagógica y Tecnológica de Colombia UPTC, Magister en Ingeniería Electrónica de la Universidad de los Andes y Doctor en Ingeniería Biomédica de la Universidad Politécnica de Catalunya - España. Ha sido profesor de la Universidad de la Sabana, Nacional y Distrital. Ha desarrollado diversas investigaciones en el área de procesamiento de señales biomédicas relacionadas principalmente con el sistema respiratorio de pacientes en Cuidado Intensivo.

Email: javier.chaparro@escuelaing.edu.co, AK.45 No.205-59 (Autopista Norte)



Dr. Beatriz F. Giraldo

Profesora Agregada de la Universidad Politécnica de Cataluña (UPC), Departamento de Ingeniería de Sistemas, Automática e Informática Industrial (ESAII), Barcelona, España. Ingeniero Eléctrico, Universidad Tecnológica de Pereira (UTP), Risaralda, Colombia (1983), Postgrado en Ingeniería Biomédica, UPC (1989), Máster en Ingeniería Biomédica, UPC (1990), Doctora en Ingeniería Biomédica por la UPC (1996). Actualmente es miembro del grupo de investigación de procesado de señales biomédicas e interpretación, del Instituto de Bioingeniería de Cataluña (IBEC) y del Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN). Desde 2006 hasta 2011 fue Directora del programa de Bioingeniería de la Fundación Politécnica de Cataluña. Sus principales áreas de investigación son el procesado y modelado de señales biomédicas y análisis estadístico, aplicado a señales cardíacas y respiratorias, procesado de señal multimodal en interacción cardiorrespiratoria, métodos para el análisis de la insuficiencia cardíaca, y del destete de la ventilación mecánica.



Susana Rondón

Profesor Asistente de la Escuela Colombiana de Ingeniería Julio Garavito en el Departamento de Matemáticas. Licenciada en matemáticas y física Universidad del Tolima. Especialista en estadística (UNAL) y Maestría en Tecnología Educativa del Instituto tecnológico de Monterrey – México

Email: susana.rondon@escuelaing.edu.co, AK.45 No.205-59 (Autopista Norte)