Clasificación de Pacientes con Diabetes Mellitus Tipo 1 Mediante Técnicas de Árbol de Decisión

Lucas Griva[1] y Marta Basualdo[1,2]

[1]FCEIA-UNR, Univ. Nac. de Rosario, Pellegrini 250 (S2000EZP), Rosario, Argentina.
[2]UTN-FRRo, Univ. Tecnológica Nacional, Zeballos 1341 (S2000BQA), Rosario, Argentina. griva@cifasis-conicet.gov.ar
mbasualdo@frro.utn.edu.ar

Resumen— El presente trabajo propone analizar los datos provenientes de 50 pacientes con Diabetes Mellitus Tipo1 (DMT1) con el objetivo de cuantificar su capacidad de regular adecuadamente sus niveles de glucosa en sangre (glucemia) mediante las dosis de insulina, calculadas empleando un algoritmo de control predictivo funcional (PFC). La clasificación de los pacientes se realiza empleando técnicas de árbol de decisión teniendo en cuenta las características de las variaciones temporales de la glucemia que se capturan mediante la modelización matemática de las mismas. Para ello se obtienen modelos tipo ARX y funciones de transferencia. Los rangos de los parámetros de dichas funciones constituyen la principal fuente de información empleada en los árboles de decisión. Como resultado se mostrará el estudio de las variables que presentan mayor incidencia para la determinación a priori de la capacidad de regulación que presentan los pacientes diabéticos. Por lo tanto la contribución de este trabajo radica en la definición de las variables que hemos considerado más relevantes y su tratamiento mediante árboles decisión. Los resultados de las pruebas realizadas sobre datos de 50 pacientes dan soporte a las conclusiones que se presentan al final del trabajo.

1 Introducción

La Diabetes Mellitus Tipo 1 (DMT1) es una enfermedad crónica caracterizada principalmente por la incapacidad del páncreas de producir insulina o de hacerlo de forma insuficiente para regular la concentración de glucosa en sangre (glucemia) y mantenerla en su rango saludable. Si la DMT1 no es debidamente diagnosticada y tratada conlleva a generar hiperglucemia, que puede producir daños en vasos sanguíneos e hipoglucemia, que puede conducir a la muerte. En el caso de la DM Tipo 2 (DMT2) el páncreas no está totalmente inhabilitado para producir insulina pero no presenta la adecuada capacidad de regular la glucemia de los pacientes.

Existe un creciente interés por contribuir a mejorar la calidad de vida de pacientes diabéticos. A través de aplicaciones bio-ingenieriles se logra el desarrollo de nuevas tecnologías, entre ellas el páncreas artificial (PA) aún en etapa de experimentación. El PA está conformado por un sensor continuo de glucosa en sangre, el suministro subcutáneo de insulina mediante el uso de una bomba y un algoritmo de control para realizar el cálculo de la insulina a inyectar, semejante al funcionamiento fisiológico

de un individuo no diabético. En principio el PA estaría recomendado para aquellos pacientes lábiles o muy sensitivos a los estímulos donde la regulación de glucemia les resulta muy compleja. Sin embargo, una gran parte de los pacientes pueden convivir bien con su enfermedad mediante un adecuado sistema de aprendizaje del funcionamiento de su sistema endocrino y metodologías informáticas y de control que sir-van de soporte para el cálculo de las dosis de insulina que deben inyectarse. En este contexto, el presente trabajo propone analizar los datos provenientes de 50 pacientes diabéticos con el objetivo de caracterizar las respuestas frente a ingestas y dosis de insulina. El principal objetivo es el de cuantificar la real capacidad de una persona con DMT1de responder al tratamiento con insulina para permanecer dentro del rango de glicemia saludable.

Típicamente se encuentra en los datos recopilados de los pacientes que las dosis de insulina se inyectan en forma simultánea a la ingesta. Esta situación produce cierta dificultad para separar adecuadamente los efectos de ambas excitaciones. Se probaron distintos tipos de modelos matemáticos para capturar la dinámica de la glicemia de los pacientes afectada por ingestas y dosis de insulina. Luego de rigurosas evaluaciones en [1] se determinó que los modelos auto-regresivos con entradas externas (ARX) eran capaces de copiar eficientemente la verdadera dinámica de los 50 pacientes estudiados. Una descripción más detallada del estado del arte en identificación de modelos del sistema endocrino y predicción para cortos plazos de tiempo se pueden consultar en [2], [3] y [4]. A partir de los modelos ARX se obtuvieron aproximaciones matemáticas simplificadas de los efectos de ambos estímulos, ingestas y dosis de insulina, en forma separada sobre la glucemia. Estos modelos simplificados, denominados funciones de transferencia, forman parte del diseño del algoritmo de control predictivo basado en modelos que se emplea en este trabajo. Particularmente aquí trabajamos con la tecnología del control predictivo funcional (PFC), [5], para el cálculo de las dosis de insulina.

En [6], se determinó un índice de la controlabilidad del paciente estudiando 30 pacientes con cuyos parámetros se corría una versión previa del simulador conocido como UVA-Padova. El algoritmo PFC fue probado en los 30 pacientes y los resultados se evaluaron mediante la técnica de grillas de análisis de la variabilidad del control, o bajo denominación anglosajona Control Variability Grid Analysis (CVGA), [7]. El CVGA es una representación gráfica donde se determinan las zonas en que se sitúan los pacientes, acorde a los valores mínimo/máximo de glucemia que presentan y permite establecer distintas jerarquías de riesgo. Provee una asistencia visual y numérica de la calidad de la regulación de la glucemia en dicha población empleando diferentes algoritmos de control. Por tanto esta técnica permitía evaluar rápidamente si el PFC había realizado adecuadamente los cálculos de insulina para cada paciente del grupo estudiado. El índice de controlabilidad, obtenido en [6], se basaba en el cálculo previo de un índice de severidad (“Sev”) en relación al peso corporal del paciente (“BW”). “**Sev”** se estimaba en base al cociente entre la máxima variación de glucemia producida por la ingesta respecto de la variación de glucemia por efecto de la dosis de insulina. La gráfica entre Sev y BW conjuntamente con los resultados del CVGA permitía delimitar dos zonas separadas por una recta y la distancia a esa recta de cada uno de los puntos correspondientes a cada paciente permitía inferir qué grupo de pacientes respondía mejor, en términos de *controlabilidad*, a las dosis calculadas

de insulina por el PFC. Este trabajo, si bien está inspirado en el análisis preliminar de [6] ha avanzado en el replanteo de Sev pero ahora analizado en función de los tiempos de respuestas glucémicas a ingestas e insulina. Es decir que nuestra mayor contribución consiste en tener en cuenta la característica dinámica de la respuesta del paciente como variable y la aplicación del método de árboles de decisión. Se pueden mencionar algunos trabajos que utilizaron algoritmos de árboles de decisión para ciertas problemáticas de diabetes. El artículo presentado en [8] se demuestra que se pudo predecir diabetes en diferentes pacientes con riesgo a partir de un cierto número de tests mostrándose un 73% de efectividad. En [9] aplica árboles de decisión CART C4.5 a determinada información de 2064 pacientes con DMT2, e identifica los cinco factores más importantes que influyen en control de glucosa. En [10] se utiliza C4.5 en pacientes con DMT2 para determinar cuál es el principal factor que influye en mayores variaciones en hemoglobina glicosilada, resultando ser la educación en diabetes del paciente.

En este contexto, cabe señalar que no se ha encontrado en la literatura que esta técnica de clasificación haya sido utilizada para definir aspectos de controlabilidad de pacientes diabéticos. El objetivo del presente trabajo es lograr una clasificación de pacientes que defina a priori cuáles presentan mayor tendencia a la *controlabilidad* de su glucemia. Para demostrarlo se analiza el grado de coincidencia que existe con las evaluaciones que proporciona un diagrama de CVGA. Se presentan los resultados y discusiones de aplicar esta metodología sobre 50 pacientes lo cual brinda el soporte necesario para indicar si las variables seleccionadas son adecuadas para caracterizar a los pacientes diabéticos. Finalmente en la sección de conclusiones y trabajos futuros se presenta un análisis riguroso del alcance del presente trabajo y se delinean las direcciones a seguir para profundizar esta temática.

2 Desarrollo del Estudio

2.1 Recopilación de los datos experimentales de pacientes con DMT1

Los datos utilizados para la identificación de modelos fueron cedidos por el Centro de Tecnologías para Diabetes de la Universidad de Virginia de Estados Unidos. Estos corresponden a un trabajo experimental con pacientes que participan como voluntarios para el proyecto internacional de páncreas artificial (los datos precisamente corresponden a la Fase 1 del Proyecto NIH/NIDDK RO1 DK 085623). Se dispone de datos históricos que corresponden a entre 14 y 32 días (promedio de 24 días) de 50 pacientes con DMT1 (peso promedio 79.2 kg, variando entre 52.6 kg y 160.1 kg) donde cada uno utiliza un sensor continuo de glucemia (marca Dexcom) que graba los valores medidos en el intersticio celular cada 5 minutos. Además, se dispone de los datos de las dosis de insulina inyectadas a través de una bomba de insulina en la cual se registran además los valores basales de dicha hormona que son suministrados cada 5 min.. Por otra parte, los pacientes deben registrar el horario y las cantidades de carbohidratos que ingieren en cada comida y realizar mediciones de control con el sensor por dígito punción del nivel de glucosa en sangre, procedimiento este último

necesario para la calibración del sensor continuo el cual por si solo presenta diferentes grados de inexactitudes. Las mediciones mediante dígito punción se realizaron en un promedio de 8 veces diarias en especial antes y después de cada comida. Los datos fueron recolectados en condiciones de vida libre sin restricciones en las cantidades de carbohidratos a ingerir, así como tampoco en la cantidad diaria de ingestas ni en el instante en que se ingieren.

2.2 Generación de los modelos a partir de los datos históricos

2.2.1 Modelos ARX

Se propone en este trabajo construir modelos ARX teniendo en cuenta las variaciones de glucosa en el intersticio (CGM) dada por ingestas y dosis de insulina y valores históricos de la variable de salida si se tiene en cuenta la predicción de los modelos para horizontes de 30- a 120 min.. Los principales propósitos para la identificación de los modelos es para uso como modelos simplificados en la estructura de un MPC y como paciente *in silico* para ayudar en el procedimiento de ajuste del controlador en una aplicación para Páncreas Artificial.

Los modelos fueron obtenidos con la función *arx()* de la *"Toolbox"* de identificación de MatLab [11]. El modelo es descripto por la ecuación

$$y(k+1) = -a_1 y(k) - \cdots - a_{n_1} y(k-n) + b_{11} u_1(k-1) + \ldots + b_{1,m1} u_1(k-m1) + b_{21} u_2(k-1) + \cdots + b_{1,m2} u_2(k-m2) \quad (1)$$

donde a_i, b_{ji} son los parámetros estimados por la función para ajustar la salida *y(k+1)* con los datos históricos. Los primeros dos tercios de los datos fueron utilizados para la estimación del modelo y el tercio restante para la validación. Las entradas u_j provienen de las dosis de insulina guardadas por la bomba y a las anotaciones en carbohidratos estimados de los pacientes, y fueron filtradas previas a utilizarse en la estimación. Tales filtros permiten aproximar la insulina en plasma desde las dosis de insulina, in mU/min, y la tasa de aparición de glucosa en plasma desde las ingestas en (mg/min) permitiendo que el modelo estimado se asocie a entradas con relaciones fisiológicas.

Los filtros utilizados están basados en las ecuaciones descriptas en [12],

$$\dot{I}_{sc1}(t) = -k_{d} * I_{sc1}(t) + I(t) \quad (2)$$

$$\dot{I}_{sc2}(t) = -k_{d} * I_{sc2}(t) + k_{d} * I_{sc1}(t) \quad (3)$$

$$\dot{I}_{p}(t) = -k_{e1} * I_{p}(t) + k_{d} * I_{sc2}(t) \quad (4)$$

donde $I_p(t)$ es insulina en plasma, y $I(t)$ dosis de insulina.

$$\dot{Q}_1(t) = -k_{a} * Q_1(t) + \omega(t) \quad (5)$$

$$\dot{Q}_2(t) = -k_{abs} * Q_2(t) + k_{a} * Q_1(t) \quad (6)$$

$$R_a(t) = (Q_2(t) * k_{abs} * f)/BW \quad (7)$$

donde $R_a(t)$ es la tasa de aparición de glucosa en plasma, $\omega(t)$ es la ingesta de carbohidratos and BW es el peso del paciente. Las constantes utilizadas en ambos filtros están ajustadas a valores medios de la población Tabla 1. El uso de estos filtros no resulta en degradación de la predicción cuando el control es aplicado finalmente a modelos estimados con datos reales. Los resultados de la predicción que se alcanzan con los modelos ARX se evalúan empleando métricas típicas para sistemas biológicos. Estas métricas son RMSE que corresponde a la raíz cuadrada del error medio entre la predicción y el valor real de glucemia y la Grilla de Error de Clarke que también son utilizadas en [1]. El método de Clarke define el nivel de riesgo para el paciente basado en que las magnitudes de error de predicción de los modelos pueden conducir a serios problemas de diagnóstico. En tal sentido considera 5 zonas de A hasta D, lo deseable es tener los porcentajes más altos en zona A o B como indicativo de la mejor calidad del modelo por su capacidad de predecir. Generalmente los horizontes de predicción van desde 30 a 120 minutos porque más allá de ese tiempo los resultados se degradan. Los resultados de la predicción con las métricas utilizadas en [1], RMSE (raíz cuadrada del error medio) y Grilla de Error de Clarke se observan en Tabla 2 comparadas con las predicciones con método de zero-order-hold (ZOH); este último método estima el valor de *y* en *y(k+HP) = y(k)*.

Tabla 1: Valores medios de la población de los parámetros utilizados en los filtros.

$k_{a\,BW}$= 0.01193 (1/min.)	k_d= 0.02 (1/min.)	f=0.9
k_{a1}=0.16 (1/min.)	k_{a2} = 0.0893 (1/min.)	

Tabla 2: Resultados promedios de la predicción (entre los 50 pacientes).

HP (min)	RMSE (mg/dl)	Zona Grilla de Clarke %		
		A	B	CDE
30 (arx-zoh)	11.44 16.42	94.31 91.62	2.25 3.34	3.44 5.04
60 (arx-zoh)	30.44 33.18	86.43 84.27	11.16 13.44	2.41 2.29
90 (arx-zoh)	47.27 51.34	55.13 52.34	39.79 37.25	5.08 10.41
120 (arx-zoh)	56.23 60.81	47.11 44.35	41.17 39.19	11.82 16.46

2.2.2 Modelos en funciones transferencia de primer orden con retardo

Una vez que se disponen de los modelos ARX validados, los mismos resultan útiles para estimar las respuestas glucémicas frente a excitaciones que se producen en modo no simultáneo. Esto es porque en la práctica los pacientes diabéticos se inyectan la dosis de insulina al mismo tiempo que ingieren los alimentos. Por tanto ahora los modelos ARX representan nuestros pacientes virtuales de los que trataremos de obtener modelos simplificados denominados funciones de transferencia. Para ello se excitan las dos variables de entrada consideradas (carbohidratos e insulina) mediante

funciones de tipo escalón de pequeña magnitud. Si la respuesta temporal de la glucemia es de tipo sigmoidal se pueden aplicar técnicas sencillas de identificación [13] para obtener funciones de transferencia en el dominio complejo de Laplace de primer orden con retardo (FOTD). Basados en el principio de superposición por tratarse de modelos linearizados se puede aproximar la respuesta glucémica sumando los efectos de ambas excitaciones.

$$G_{mi}(s) = \frac{K_{mi} e^{-\theta_{mi} s}}{1 + T_{mi} s} \tag{8}$$

$$G_{di}(s) = \frac{K_{di} e^{-\theta_{di} s}}{1 + T_{di} s} \tag{9}$$

donde las constantes *K* son las ganancias de los modelos, la relación entre la variación glucémica una vez alcanzado el estado estacionario respecto de la magnitud del escalón de la variable excitatriz. θ representa el retardo que se calcula como el tiempo en que se produce el escalón y el instante en que la glucemia comienza a variar debido a éste. Las constantes de tiempo *T* representan el tiempo desde la aparición del primer cambio en glicemia hasta que esta alcanza el 63,2% de su valor final. Los subíndices *mi* se refieren a la insulina y *di* a las ingestas.

Mediante las anti-transformadas de Laplace del producto de las funciones dadas por 8 y 9 por sus respectivas excitaciones permiten evaluar las respuestas temporales de glucemia frente a una dosis de insulina, y frente a una ingesta de carbohidratos respectivamente. En ecuaciones, $L^{-1}\{G_{ji}(s) * U(s)\} = y(t) = {}^{1}/_{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{st} G_{ji}(s) * U(s)\, ds$, donde *j= m,d* , y L^{-1} es el operador transformada inversa de Laplace. La respuesta del sistema *y(t)* resulta, si la entrada es un impulso unitario y por tanto *U(s)* una constante:

$$y(t) = K\left(1 - e^{-(t-\theta)/T}\right) \tag{10}$$

donde las constantes K, T y θ tienen los significados ya mencionados.

3 Breve revisión de PFC

PFC pertenece a la tercera generación de una familia de control predictivo basado en modelo. Consiste básicamente de cuatro elementos principales tales como un modelo dinámico del proceso, una trayectoria de referencia *yr(k)*, una metodología específica de compensación del error predicho y una estructura particular para la definición de la variable manipulada. El error futuro entre *yr(k)* y la salida predicha por el modelo se estima sobre el horizonte de coincidencia $[H_1;H_2]$ definido desde el tiempo muerto hasta el instante en que la respuesta alcanza el estado estacionario. La ventaja de este algoritmo es que se fijan determinadas condiciones sobre un número de puntos de coincidencia que como mínimo puede ser 1 disminuyendo la carga compu-

tacional para el cálculo de la acción correctiva por parte de la variable manipulada. La autocompensación permite estimar un error futuro para corregir las predicciones del modelo. El diseño de PFC, permite rechazar perturbaciones sobre la base de disponer de los modelos correspondientes, en ese caso el G_{di}, y las restricciones que presenta una bomba de insulina, por ejemplo, también se tienen en cuenta de una manera natural. Para una descripción más profunda, se recomienda ver [5]. El PFC fue testeado para pacientes diabéticos *in silico* en [13] dando como resultado que los pacientes estuvieran en su totalidad en las zonas A y B del CVGA que son las más favorables. En este trabajo también utilizamos PFC para aplicarlo a los 50 pacientes virtuales cuyos modelos ARX validados ya disponemos. El resultado de esta aplicación también se evalúa a través del CVGA y ese resultado es de utilidad para confrontarlo con la clasificación que se realiza empleando los árboles de decisión.

El PFC tiene tres entradas, la medición de glucemia que proviene del CGM, el set-point de glucosa que se fija en 100 mg/dl y el anuncio de ingestas dadas como pulsos de carbohidratos estimados. Los modelos internos utilizados son los estimados con la metodología de [14] y del tipo detallado en la sección 2. En [15] se aplica este tipo de control para dos pacientes con DMT1 representados por un modelo ARX cada uno, como los que se utilizan aquí, y observándose su buena capacidad para mantener a los pacientes dentro del rango saludable (70-180 mg/dl).

La representación gráfica del CVGA consta de un plano X-Y positivo donde tanto X como Y tienen unidades de glucemia en mg/dl que contemplan diferentes rangos para determinar zonas saludables y de riego para los pacientes. Por tanto, el diagrama se encuentra dividido en 9 zonas: A, B, B Superior, B Inferior, C Superior, C Inferior, D Superior, D Inferior y E, sobre las cuales se ubican los puntos que resultan de tomar el valor mínimo y máximo de glucemia obtenidos en este trabajo para 24 hs. de aplicación de PFC para cada paciente. El período puede contemplar diferentes magnitudes de tiempo.

4 Caracterización del paciente

A partir de la información que brindan los modelos simplificados (8) y (9) se calcula el *Índice de Severidad (Sev)* como

$$Severidad = \frac{K_{di}}{|K_{mi}|} \tag{11}$$

El valor promedio de Sev calculado para los 50 paciente fue 31.95 pmol·Kg/mg variando entre 11.82 pmol·Kg/mg y 102.4 pmol·Kg/mg. Mayores valores del índice pueden deberse a la mayor ganancia por carbohidratos para ganancias por insulina similares; o bajos valores de esta última para ganancias similares para carbohidratos. De los 13 pacientes con mayor índice de severidad aquí utilizados, 11 de ellos se corresponden a menor ganancia por insulina; mientras que los dos restantes se deben a las altas ganancias por carbohidratos.

Este índice es propuesto conjuntamente con el *Tiempo de Respuesta a Carbohidratos normalizado*, como posibles atributos predictores (entradas del modelo de árboles de decisión) de la zona del CVGA (Control Variability Grid Analysis) en que resulta-

rá el paciente luego de controlado. El *Tiempo de Respuesta a Carbohidratos normalizado (TRCHn)*, es el valor de T_{di} de cada paciente dividido el valor máximo obtenido de entre los 50 pacientes (ver **2.2.2**). El valor medio de esta nueva variable es 0.138 y varía entre 0.012 y 1. A menor tiempo de respuesta, y por consiguiente su normalización, más rápido resulta el paciente para alcanzar el valor máximo de glucosa ante una determinada ingesta.

Luego de obtenerse mediante la simulación del control predictivo funcional (PFC) la zona del CVGA en que resulta el paciente, se utiliza la misma como valor de "target" (salida del modelo de árboles de decisión) y los atributos como las variables que predicen tal salida; se constituye luego un conjunto de entrenamiento con todos los pacientes.

5 Árboles de decisión

Los árboles de decisión son una forma gráfica y analítica de representar todos los eventos que pueden surgir a partir de una decisión asumida en cierto momento. Luego permiten tomar la decisión probabilísticamente más acertada dadas ciertas posibles decisiones. Es un método utilizado para inferencia inductiva y para aproximación de funciones con valores de "target" discretos, donde dicha función aprendida es representada mediante el árbol (clasificación).

Cada nodo en el árbol (rectángulos en Fig. 1) especifica una prueba de algún atributo de la instancia que le precede, y cada rama descendiente del nodo corresponde a uno de los posibles valores para este atributo. Una instancia es clasificada comenzando en el nodo raíz del árbol testeando el atributo especificado por este nodo, luego moviéndose hacia abajo por la rama del árbol correspondiente al valor del atributo. El proceso es repetido para el "sub-árbol" a partir del nuevo nodo; la rama del árbol termina con el valor de "target" (valor de salida) de la clasificación realizada siguiendo tal rama.

En Fig. 1, los nodos se corresponden con los nombres de las variables utilizadas para clasificar, es decir Sev y TRCHn, mientras que las ramas saliendo de tales nodos representan diferentes valores de corte significantes del atributo del nodo del que parten, Sev > c, o TRCHn ≤ c, donde c es un número real que se adopta para cada caso particular de aplicación. Las ramas terminan en los valores de salida aquí utilizados, en este caso los "targets" corresponden a zona A y zona B que son las mejores posiciones para el CVGA y que demuestran que se ha logrado que el paciente esté en un rango saludable. Se debe remarcar que el mejor resultado es la zona A.

En general un árbol de decisión representa una disyunción de conjunciones de restricciones en los valores de los atributos de las instancias. Es decir las salidas pueden representarse de forma lógica como:

Si (primer valor Atributo 1) y (primer valor Atributo 2) o (segundo valor Atributo 1) y (tercer valor Atributo 2) Entonces: zona X. Donde *X* equivale a *A* o *B*.

Los algoritmos básicos utilizados fueron desarrollados por [16], [17]. Se define una propiedad denominada *ganancia de información* que mide cuan bien un atributo separa los ejemplos de entrenamiento de acuerdo a su valor de target. Para ello se

define primeramente la *entropía* de una colección de ejemplos *S* (*valor TRCHn, valor Sev, Zona X*) cuyos valores de salida pueden tomar solo dos valores, como una medida de la impureza de una colección arbitraria de ejemplos:

$$Entropia(S) = -p_+ \log p_+ - p_- \log p_- \tag{12}$$

donde p_+ es la proporción de ejemplos de una clase (de un valor de salida, denominada positiva en este caso), y p_- la de ejemplos de una clase "negativa". Luego la *ganancia en información*, es decir la efectividad de un atributo en clasificar los datos de entrenamiento, se define como la reducción esperada en entropía causada por particionar los ejemplos *S* de acuerdo al mencionado atributo *A* (*Sev > c, o TRCHn ≤ c, donde c es un número real*):

$$Ganancia(S, A) = Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{13}$$

donde, *Valores(A)* es el conjunto de todos los posibles valores para el atributo *A*, S_v es el subconjunto de *S* para el cual el atributo *A* tiene valor *v*. Se puede interpretar ahora tal ganancia como la información provista sobre el valor de la clase de salida, dado el valor de algún otro atributo *A*. Esta medida es utilizada por el algoritmo para seleccionar el mejor atributo a cada paso en el crecimiento del árbol (a mayor ganancia mejor es el atributo).

En el caso como el de este trabajo en el que los atributos (variables) utilizados son de valores continuos, el algoritmo crea valores discretos en los mismos a partir de generar intervalos dentro del rango continuo con valor de corte *c*, donde los nuevos atributos con valores discretos son *Ac*, verdadero si $A < c$, y falso en caso contrario. Este valor de *c* elegido de entre varios posibles valores de *c* es tal que produce el mayor valor de *ganancia en información* (estos valores de c se corresponden a los de las inecuaciones de las ramas de Fig. 1).

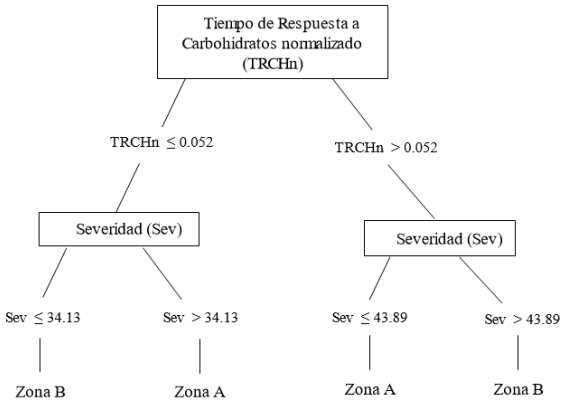

Figura 1. Árbol de decisión obtenido para el caso trabajado (50 pacientes reales).

6 . Resultados y discusiones

El CVGA resultante de aplicar PFC se muestra en Fig. 2, donde se aprecia que los porcentajes se distribuyen como 66% en zona A y 34% en zona B.

Tanto "zona A", como "zona B" constituyeron los valores de "target" a predecir mediante la aplicación de árboles de decisión.

En [6] se presenta un índice de controlabilidad para determinar cuan controlable o no es un paciente dando como resultado adicional una clasificación llevada a cabo de forma empírica. Separando mediante una recta dos grupos bien definidos entre el total de pacientes considerados: 30 con DMT1, 10 con DMT2, 1 adulto normal y 10 pacientes prediabéticos. Se utilizaron como predictores el peso del paciente y el *Índice de Severidad*. Todos los pacientes fueron simulados con el simulador Uva/Padova y se corresponden con datos de la patente del mismo [18]. Por último, se identificó una recta que es la que mejor separa en dicha gráfica los pacientes que terminaron en zona B y AB del CVGA (grupo 1), de aquellos que se ubicaron en zona A junto con los pacientes con DMT2, prediabético y normal (grupo 2).

El resultado final indica que la clasificación tiene un error 3.5% en clasificar mal a los del grupo 2, y 26.1% en identificar mal a los del grupo 1; con un error de 13.7% total. Si solo se tienen en cuenta los pacientes diabéticos tipo 1 el error final es de 16.7%, con un error de clasificar mal a los del grupo 2 de 0% y 21.7% en clasificar mal los del grupo 1.

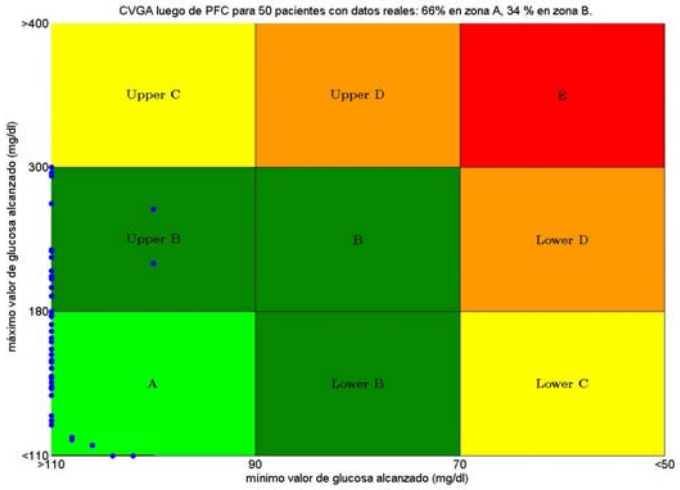

Figura 2. CVGA obtenido luego de aplicación de PFC.

Luego de aplicársele árboles de decisión (algoritmo C4.5) a los resultados obtenidos con pacientes reales utilizando los atributos vistos en el apartado anterior; y comprobarse que el error total era de 20% pero el error de clasificar mal los pertenecientes a la zona B era cercano al 50%, se optó por utilizar otros atributos. Se utilizaron entonces los mencionados en la sección **5**, por ser los directamente relacionados con

las dinámicas de la glucemia luego de ingestas y dosis de insulina, y por ende de los valores máximos y mínimos que esta puede alcanzar.

Se obtuvo un error total del 14% (comparable al obtenido en [6] pero con otras variables predictores), con un error de clasificar mal los del grupo 1 de 29.4% y 6.1% en clasificar mal los del grupo 2; donde grupo 1 son aquellos pacientes que acabaron en zona B luego de control y los del grupo 2 los que pertenecen a la zona A que es la más favorable. La Fig. 3 muestra la división de zonas en el semiplano positivo (*Sev-TRCHn*) resultante luego de la clasificación.

La necesidad de utilizar como variables el *Índice de Severidad* y el *Tiempo de Respuesta a Carbohidratos normalizados*, y no las variables utilizadas en [6], surgieron ante la idea de que no solo las ganancias del modelo (FOTD) iban a influir en si el paciente permanecía o no en el rango saludable (70-180 mg/dl) si no también la capacidad del sistema en asimilar la ingesta, dado por el tiempo de respuesta a la misma. Los valores más pequeños del *Índice de Severidad* indican o una ganancia pequeña de carbohidratos (picos de glucemia más pequeños) o bien una ganancia alta en insulina, y por lo tanto una mayor facilidad con menor insulina de contrarrestar el pico de glucosa en sangre producido por la ingesta. A su vez los valores más grandes en Sev indican una mayor ganancia de glucemia para la ingesta o una menor para la insulina. Esto último puede implicar que la insulina dada en un instante de tiempo (restringida por la bomba de insulina) no contrarreste adecuadamente el efecto de la ingesta.

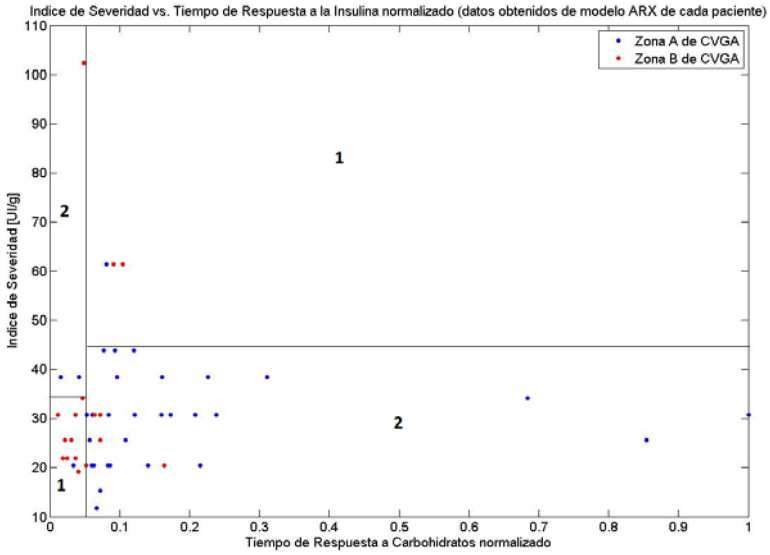

Figura 3. Resultados de la clasificación con árboles de decisión. Las zonas delimitadas caracterizadas con el número 1, son aquellos en las que el algoritmo interpretó que correspondían a zona B. El grupo 2 predice que el paciente estará en zona A.

A igual valores de Severidad, el tiempo de respuesta de carbohidratos resulta la variable decisiva. Un tiempo de respuesta más pequeño implica que la respuesta a la ingesta tenderá a alcanzar su valor máximo más rápidamente; por lo que existe una

menor posibilidad de que el suministro sucesivo de insulina logre contrarrestarlo, y una mayor posibilidad de que se llegue a un valor de glucemia mayor que para un tiempo de respuesta menor. Esto se observa en la Fig. 3, donde la línea vertical trazada (corte en la variable de la abscisa dado por el algoritmo), marca una clara diferencia entre puntos relacionados con zona B en un lado, y puntos relacionados con zona A en otro.

De los 4 pacientes con severidad más grande (de 60 a 105 pmol·kg/mg) tres de estos pertenecen a la zona B y están relacionados uno con una ganancia de carbohidratos muy alta y ganancia a insulina media, y los 2 restantes con ganancias a insulina media. A menor ganancia de insulina, más insulina se necesita para contrarrestar el efecto de la ingesta (insulina limitada además por las restricciones que impone la bomba de insulina). El total de los 9 pacientes cuya severidad varía entre 36 y 45 pmol·kg/mg se encuentra en zona A y ocho de ellos corresponden a ganancia a insulina pequeña y ganancia a carbohidratos entre pequeña y media (veáse Fig. 4). El resto queda prácticamente dividido por el tiempo de respuesta a carbohidratos normalizado.

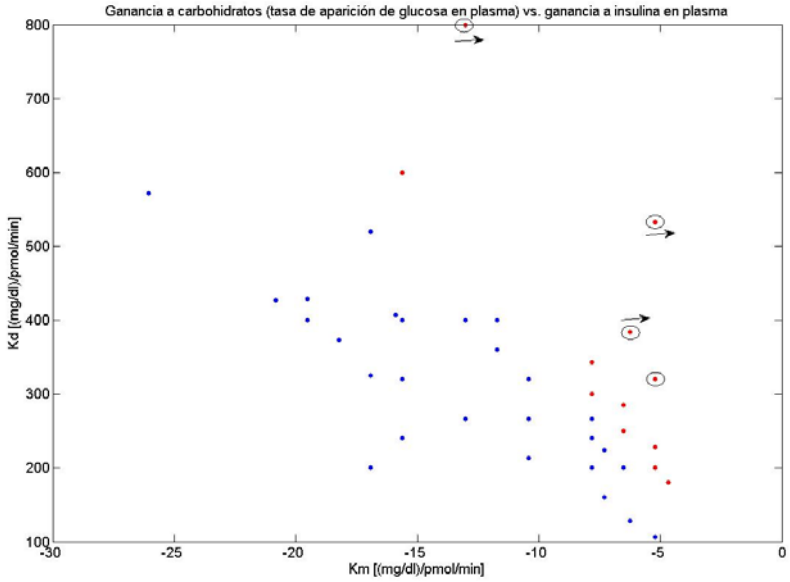

Figura 4. Ganancia a carbohidratos vs. Ganancia a insulina. Puntos rojos, aquellos cuya severidad es mayor o igual a 36 pmol·kg/mg; puntos azules lo contrario. Círculos negros en puntos rojos, severidad mayor o igual a 60 pmol·kg/mg. Flechas al lado de puntos rojos, severidad mayor o igual a 36 pmol·kg/mg y zona B en CVGA.

La técnica de árboles de decisión en cambio permite la obtención de zonas como la de la Fig. 3, las cuales no se corresponden con ninguna curva de separación si no con zonas rectangulares en el semiplano positivo de *Índice de Severidad* vs. *Tiempo de Respuesta a Carbohidratos normalizado*, en este caso. La cantidad de datos correspondientes a 50 pacientes no permitía en este caso contar con datos de validación y/o

de test, que permitirían apreciar el error alejado del sobre ajuste (*overfitting*), el cual puede observarse en la división de las dos zonas superiores las cuales poseen en total datos de 6 pacientes (12% del total) y por lo tanto no existe la seguridad de que la separación sea correcta si se cuentan con mayor cantidad de datos y distintos a los utilizados. Pero esta clasificación resulta contundente para los datos cuyo *Índice de Severidad* es menor a 45 pmol·Kg/mg.

Para ejemplificar esto último vale también añadir los resultados producto de aplicar el algoritmo C4.5 a datos de entrenamiento cuyas variables predictoras fueron en un caso ganancia a carbohidratos y tiempo de respuesta normalizado, y en otro ganancia a insulina y tiempo de respuesta normalizado. Los resultados arrojaron un error de 18% en ambos casos, pero solo utilizando información del tiempo de respuesta, resultando la variable restante irrelevante.

7 Conclusiones y Trabajos Futuros

Se estimaron modelos ARX para que su capacidad predictora permita considerarlos como pacientes virtuales a quienes se les aplica Control Predictivo Funcional para analizar el comportamiento de los pacientes bajo control. Una vez obtenidos los resultados del controlador expresados en el diagrama de CVGA se procedió a una posible clasificación de los mismos a partir de introducirse dos variables que pudieran influir en la zona del CVGA en que aparecería el paciente luego de controlado. A partir del Índice de Severidad y del Tiempo de Respuesta a Carbohidratos normalizado se aplicó la técnica de árboles de decisión obteniéndose resultados significativos. Se pudo determinar el rango de severidad y tiempo de respuesta que podía conducir al paciente a la mejor zona de control según el esquema del CVGA. Este primer resultado es promisorio en la búsqueda de variables que nos permitan predecir a partir de datos obtenidos rutinariamente por los pacientes, la característica de controlabilidad de los mismos antes de aplicar el algoritmo de control.

Se continuará trabajando en la búsqueda de otras variables que puedan brindar mejores resultados, así como también con datos de mayor cantidad de pacientes o con datos obtenidos de estudios diferentes. Se seguirán evaluando otros métodos de clasificación en busca de determinar si las variables que conforman la severidad pueden tener impacto por separado en la predicción de la controlabilidad de los pacientes. También se aplicarán metodologías que permitan la obtención de modelos a partir de los datos que puedan brindar mejores predicciones del comportamiento glucémico de los pacientes frente a otras perturbaciones como la realización de ejercicios físicos.

8 Agradecimientos

Los autores agradecen el financiamiento de CONICET, UTN-FRRo y UNR-FCEIyA para la realización de este trabajo.

REFERENCIAS

1. Griva, L., Basualdo, M.: Análisis del método Wiener para modelado del sistema endocrino de pacientes con Diabetes Mellitus Tipo 1. 45º JAIIO, 5º SII. Ciudad Autonoma de Buenos Aires, Argentina. (2016) 155-166
2. Cescon, M.: Modeling and Prediction in Diabetes Physiology. Ph. D. Tesis, Universidad de Lund, Lund, Suecia. (2013)
3. Stahl, F.: Diabetes Mellitus Glucose Prediction by Linear and Bayesian Ensembled Modeling, Universidad de Lund, Lund, Suecia. (2012)
4. Zecchin, C.: Online Glucose Prediction in Type 1 Diabetes by Neural Network Models. Ph. D. Tesis, Universidad de los Estudios de Padova, Padova, Italia. (2014)
5. Richalet, J., O'Donovan, D.: Predictive Functional Control – Principles and Industrial Applications. Springer. (2009)
6. Campetelli, G., Musulin, M. S., Basualdo, M.: A Novel Index to Evaluate the Blood Glucose Controllability of Type I Diabetic Patients. 18º IFAC, Congreso Mundial. Milan, Italia. (2011) 14223-14228.
7. Magni, L., Raimondo, D. M., Dalla Man, C., Breton, M., Patek, S., De Nicolao, G., Cobelli C., Kovatchev, B. P.: Evaluating the accuracy of closed-loop glucose regulation via control variability grid analysis. J. Diabetes Sci. Technol. (2008) 630-635
8. Kaur, G., Chhabra, A.. Improved J48 Classification Algorithm for the Prediction of Diabetes. International Journal of Computer Applications. (2014) 13-17
9. Huang, Y., McCullagh, P., Black, N., Harper, R.: Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med. (2007) 251–262.
10. Sigurdardottir, A. K., Jonsdottir, H., Benediktsson, R.: Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis. Patient Educ Couns. (2007) 21–31
11. Ljung, L.: System Identification: Theory for the User, 2nd Edition. Upper Saddle River, New Jersey: Prentice Hall. (1999)
12. Patek, S.D., Lv, D., Ortiz, E.A., Hughes-Karvetski, C., Kulkarni, S., Zhang, Q., Breton, M.: Prediction Methods for Blood Glucose Concentration: Design, Use and Evaluation. Empirical Representation of Blood Glucose Variability in a Compartmental Model. Springer International Publishing Switzerland. (2016) 133-157
13. Campetelli, G.: Desarrollo de Bio-modelos Computacionales para Asistir en la Toma de Decisiones Tendientes a Mejorar la Calidad de Vida de Pacientes Diabéticos. Ph. D. Tesis. Universidad de Rosario, Rosario, Argentina. (2014)
14. Cohen, G. H., Coon, G. A.: Theoretical consideration of retarded control. Trans. Amer. Soc. Mech. Eng.. (1953) 827-834.
15. Griva, L., Breton, M., Cherñavvsky, D., Basualdo, M.: Commissioning procedure for predictive control based on ARX models of Type 1 Diabetes Mellitus patients. 20º IFAC, Congreso Mundial. Toulusse, Francia. (2017) 11023-11028
16. Quinlan, J. R.: Learning efficient classification procedures and their application to chess end games. En R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), Machine learning: An artificial intelligence approach. San Matw, CA: Morgan Kaufmann. (1983)
17. Quinlan, J. R.: Induction of decision trees. Machine Learning. (1986) 81-106
18. Kovatchev, B. P., Breton, M., Cobelli, C., Dalla Man, C.: Method, system and computer simulation environment for testing of monitoring and control strategies in diabetes. Patente WO/2008/157781. (2008)