APLICACIÓN DE TÉCNICAS DE AGRUPAMIENTO A CORREDORES DE RESISTENCIA PARA LA ESTIMACIÓN DEL UMBRAL DE LACTATO

Urtats Etxegarai[1], Eva Portillo[1], Jon Irazusta[2], Itziar Cabanes[1], Asier Zubizarreta[1]

{urtats.etxegarai, eva.portillo, jon.irazusta, itziar.cabanes, asier.zubizarreta}@ehu.eus
[1]Departamento de Ingeniería de Sistemas y Automática, Escuela de Ingeniería de Bilbao (UPV/EHU)
[2]Departamento de Fisiología, Facultad de Medicina y Enfermería (UPV/EHU)

Resumen

Hoy en día el running es, sin lugar a dudas, uno de los deportes más populares que se incorpora como hábito saludable en la vida cotidiana. Dicha práctica deportiva se ve además incentivada por la celebración de múltiples competiciones populares, lo que ha favorecido el cada vez mayor interés por mejorar el rendimiento deportivo así como la seguridad del y la deportista. En este sentido, en la última década se ha extendido enormemente la utilización de Relojes Inteligentes para monitorizar variables fisiológicas. Sin embargo, no todas las variables fisiológicas de interés pueden ser monitorizadas, destacando entre ellas el lactato, cuya medida requiere de muestras de sangre. La curva de lactato permite extraer el umbral de lactato, el cual es fundamental a la hora de planificar entrenamientos y conocer el estado del atleta. Por ello, en este trabajo se propone el diseño y desarrollo de un sensor virtual de lactato. Más concretamente, en este artículo se explora la aplicación de técnicas de agrupamiento de deportistas como vía hacia una estimación personalizada del umbral de lactato.

Palabras clave: umbral de lactato, clusterización, clasificación, estimación personalizada, running.

1 INTRODUCCIÓN

En los últimos años es patente el creciente auge de los deportes de resistencia, en especial de las carreras de media y larga distancia, así como el triatlón de resistencia. Actualmente, los y las participantes de dichos deportes utilizan Relojes Inteligentes o pulsómetros junto con aplicaciones software para el control del entrenamiento, las cuales permiten analizar a posteriori los datos de los recorridos, realizar comparativas, obtener valores estadísticos,… facilitando la realización de análisis completos y sistemáticos de los avances del deportista a lo largo de la temporada y, por tanto, dando apoyo a la hora de establecer directrices de entrenamiento.

Entre las directrices de mayor importancia en la planificación de la temporada se encuentra el "umbral de lactato" [13]. Hoy en día está demostrado que el umbral de lactato, es decir, la intensidad de ejercicio a la cual empieza a aumentar significativamente la concentración de lactato sanguíneo respecto a los valores de reposo, es más determinante para el rendimiento deportivo que el consumo máximo de oxígeno o la economía de carrera. Sin embargo, para poder medir la concentración de lactato es necesaria la extracción de sangre, personal cualificado y equipamiento específico, lo que supone un coste elevado y el desplazamiento a un centro especializado. En este sentido, la *Sensorización Virtual* aparece como una alternativa interesante posibilitando la estimación de *variables de interés* dificiles, costosas o imposibles de medir con técnicas de sensorización tradicionales, a partir de variables fácilmente medibles mediante dichas técnicas, y con las que se encuentran relacionadas las variables de interés.
Así, en el marco de este trabajo se propone el diseño y desarrollo de un Sensor Virtual de Lactato que permita estimar el umbral de manera autónoma e independiente, sin necesidad de adherir al cuerpo del deportista aparatos adicionales, y eliminando las muestras de sangre realizadas con el medidor de lactato. Concretamente, en este trabajo se explora la aplicación de dos enfoques, uno basado en la clasificación y otro en la *clusterización,* ambos aplicados a estadísticos y características representativas de la problemática con el fin de poder obtener estimaciones personalizadas del umbral de lactato.

2 ANTECEDENTES

A nivel científico, se están realizando importantes esfuerzos en cuanto a la medida no invasiva del lactato. Una de las líneas de investigación más tradicionales consiste en medir lactato en el sudor. Muestra de ello es un biosensor electroquímico para la medida no invasiva en tiempo real del lactato mediante la sudoración [11], cuya estrategia se ha extendido también a la inferencia de glucosa [3]. Sin embargo, como los propios autores reconocen, todavía se debe demostrar la correlación entre el lactato en

sangre y el lactato en el sudor. Es más, existe una corriente contraria a la existencia de dicha correlación. En la revisión de [5], en la que se repasan los trabajos en este ámbito desde 1934, los autores destacan que la inmensa mayoría de las publicaciones demuestran que el lactato sanguíneo y el lactato en el sudor son independientes ya que el lactato en el sudor es enteramente un derivado del metabolismo de las glándulas sudoríparas.

También es relevante la línea de investigación centrada en el diseño y desarrollo de dispositivos adheribles al cuerpo para la estimación del umbral de lactato [4]. Sin embargo, dichas soluciones suponen adherir al cuerpo dispositivos o elementos adicionales cuya colocación puede perjudicar la práctica deportiva, además de un coste elevado que no todos los deportistas aficionados se pueden permitir. Esto se hace aún más evidente en deportes como el triatlón en los que las transiciones entre disciplinas resultan determinantes.

Por otro lado, está ampliamente reconocido que la concentración de lactato está relacionada con múltiples características del atleta como son el ritmo cardiaco (HR), las pulsaciones en recuperación (HRR), la edad, la dieta y el estado de forma del atleta [12]. Sin embargo, las relaciones explícitas entre dichas características y el umbral de lactato son desconocidas y de naturaleza compleja [13]. Es por ello que las estrategias de modelado basado en la experiencia son probablemente muy adecuadas al no ser posible hoy en día establecer modelos analíticos. En el caso de la estimación del lactato, lo cierto es que las soluciones basadas en técnicas de Aprendizaje Autómatico han sido escasamente utilizadas. Entre los pocos trabajos encontrados, se propone un modelo basado en una red perceptron multicapa (MLP) para estimar el ritmo cardiaco HR en el punto OBLA [6]. El punto OBLA corresponde a una concentración en sangre fija de lactato de 4 mmol/l. Los autores consideraron varios parámetros derivados de HR como entradas de la red, y la precisión obtenida fue satisfactoria. Sin embargo, los mismos autores reconocen en su publicación que son necesarios más ejemplos de entrenamiento y testeo que consideren grupos más heterogéneos. Es más, hoy en día está ampliamente reconocido que el método basado en el punto OBLA es poco adecuado y está claramente superado por el método Dmax [14]. Precisamente en este método se ha basado nuestro trabajo previo [7-9]. En [8], se propone la estimación del umbral de lactato a partir de redes neuronales recurrentes y el método Dmax, cuyo diseño y desarrollo se ha basado en el diseño de una metodología experimental y consiguiente captura de una base de datos con información y experimentos realizados con 105 voluntarios y voluntarias. Dicha estimación ha sido mejorada en un trabajo reciente [9] a través de una propuesta metodológica que permite definir un estimador con altas capacidades de generalización dentro de la población objetivo. Sin embargo, toda solución basada en técnicas de Aprendizaje Automático se encuentra fuertemente condicionada por la base de datos utilizada, especialmente en términos de capacidad de generalización. En este sentido, resulta interesante abordar estrategias adaptativas que conduzcan hacia una *estimación personalizada*, de manera que se haga posible la estimación del umbral de lactato de un individuo nuevo a partir de los datos de individuos similares disponibles en la base de datos, favoreciéndose así soluciones escalables y versátiles.

3 METODOLOGÍA EXPERIMENTAL

La prueba de lactato se trata de una prueba de esfuerzo incremental realizada sobre un tapiz rodante. En este trabajo, se ha definido una ergometría en la que se comienza corriendo a una velocidad de 9,0 km/h, con un incremento entre *escalones* de 1,5 km/h hasta los 13,5 km/h, a partir del cual los incrementos entre escalones pasan a ser de 1 km/h (*protocolo 1,5-1*). Cada escalón supone 4 minutos corriendo y un minuto de reposo durante el cual se toma una muestra de sangre capilar para determinar la concentración de lactato sanguínea. La frecuencia cardiaca (pulsaciones por minuto) es capturada durante toda la prueba con un periodo de muestreo de un segundo. Para ello se ha utilizado el reloj Garmin Forerunner 910XT con cinta HRMTM, de manera que posteriormente los datos son descargados. En una primera fase se ha realizado un total de 105 pruebas correctas con el apoyo de una médica especializada en el ámbito deportivo. Los corredores han realizado la prueba tras ser debidamente informados y firmar la hoja de consentimiento. Asimismo, se han registrado otros datos de interés tales como: información acerca de su entrenamiento y competiciones, edad, peso, talla, temperatura y humedad, escala del esfuerzo percibido (Borg), etc.

4 HACIA UNA ESTIMACIÓN PERSONALIZADA

El *modelado personalizado* tiene por objetivo obtener un modelo específico sobre un individuo nuevo y desconocido a partir de los datos de los individuos similares disponibles en la base de datos. En comparación con los modelos *globales* entrenados con todos los individuos disponibles en la base de datos, los modelos personalizados tienen el potencial de poder proporcionar estimaciones más precisas sobre un individuo desconocido, así como sistemas de modelado escalables [17]. Así, la estrategia planteada en este trabajo consiste en construir un estimador para un atleta nuevo y desconocido a partir de datos de

individuos similares disponibles en la base de datos, tal y como se muestra en la Figura 1.
Tal y como se concluye en los trabajos previos [7-9], dicho modelo o estimador proporcionará una estimación del umbral de lactato según el método Dmax (ver Figura 1).

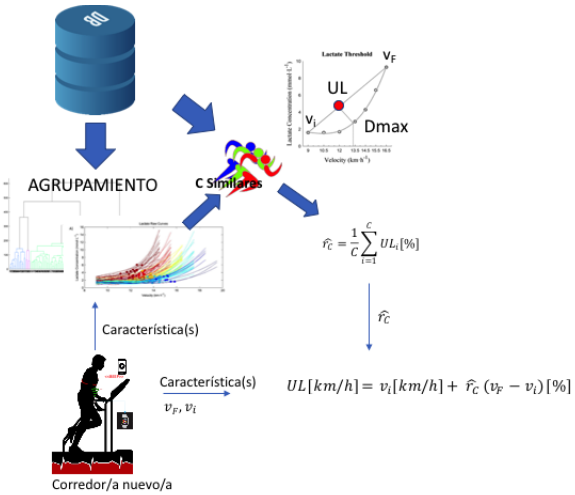

Figura 1: Estrategia basada en *modelado personalizado*

De cara a extraer los individuos similares que permitan realizar la estimación personalizada propuesta en este trabajo, se valoran dos alternativas: una basada en *clasificación* y otra en *clusterización*. Las técnicas de clasificación pertenecen al ámbito del aprendizaje supervisado, de manera que las *clases* o etiquetas son conocidas, en el caso de problemas complejos gracias al conocimiento experto. Por el contrario, en las técnicas de clusterización se aplica aprendizaje no supervisado, resultando una vía adecuada cuando no existe conocimiento previo sobre las clases o la distribución de los grupos o *clústeres* debido a la propia complejidad del problema, si bien es clave seleccionar de manera adecuada las características de los individuos a partir de las cuales poder generar los clústeres. En este último caso, el incremento exponencial de muchas bases de datos gracias al desarrollo de los últimos años de los conceptos *big data* y computación en la nube, han dado lugar a un mayor interés en la aplicación y mejora de las técnicas de clusterización [1].

4.1 ESTIMADOR DEL UMBRAL DE LACTATO

A la vista de los resultados de los trabajos previos [8-9], se define un estimador *parsimonioso* del umbral de lactato

$$UL[km/h] = v_i[km/h] + r\,(v_F - v_i)[\%] \quad (1)$$

Donde r es un factor de proporcionalidad, v_F es la velocidad del último escalón finalizado [7], y v_i es la velocidad inicial de la prueba de lactato.

El estimador expresa que el umbral de lactato (calculado en términos relativos respecto al eje de velocidad según [8]) se encuentra situado en un cierto porcentaje r de la diferencia entre la velocidad máxima e inicial. Para ello se ha estandarizando el eje de velocidad, tal y como se muestra más adelante en la Figura 3 (es decir, poniendo los valores del eje de abscisas de la curva de lactato en la misma escala). Tanto los valores v_F como v_i son valores conocidos y dependientes de la realización de la prueba de lactato del individuo. Por tanto, se debe estimar el factor r que sea representativo del grupo o clúster correspondiente, de manera que permita estimar los umbrales de lactato de los individuos pertenecientes al mismo. Para ello, por cada grupo o clúster, se toman las curvas de la base de datos correspondientes a ese clúster y se calcula, mediante el método Dmax, el umbral UL[%] real de cada individuo, de manera que el factor r del clúster se estima como la media de los umbrales relativos UL[%]:

$$\hat{r_C} = \frac{1}{C}\sum_{i=1}^{C} UL_i[\%] \quad (2)$$

Donde C es el número de individuos o atletas pertenecientes al clúster, y UL_i es el umbral de lactato relativo del individuo i perteneciente al grupo o clúster.

4.2 AGRUPACIÓN DE ATLETAS SIMILARES MEDIANTE CLASIFICACIÓN

En este caso la agrupación de atletas similares se realiza aplicando directamente el criterio de los expertos de dominio. Concretamente, la hipótesis es que aquéllos y aquéllas atletas que hayan alcanzado un mismo escalón presentan un estado de forma similar en términos de umbral de lactato. Por tanto, las clases se pueden establecer de manera sencilla en función del último escalón alcanzado en la prueba de lactato (ver Figura 2). Así, el estimador del umbral de lactato de un nuevo atleta se calculará de manera sencilla a partir de los datos disponibles en la base de datos de aquéllos individuos cuyo último escalón alcanzado coincida con el del nuevo.

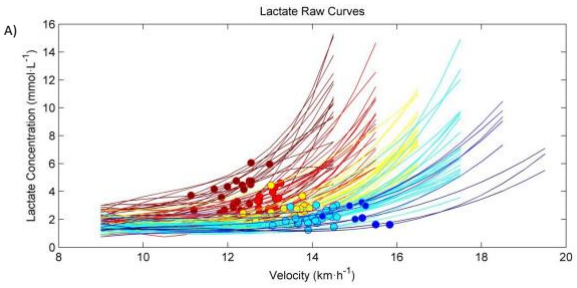

Figura 2: Clasificación en función del último escalón alcanzado durante la prueba de lactato

Una de las ventajas que presenta esta aproximación, es su sencillez para abordar la estimación del umbral de lactato ante los cambios de estado de forma de los corredores, ya que la estimación se podrá actualizar en función del último escalón alcanzado durante la prueba de lactato.

4.3 AGRUPACIÓN DE ATLETAS SIMILARES MEDIANTE CLUSTERIZACIÓN

A pesar de que la clasificación de los atletas en función del último escalón alcanzado sea una hipótesis lógica desde el punto de vista fisiológico, lo cierto es que estamos ante un problema de naturaleza compleja en el que influyen múltiples factores, lo que hace que se desconozca de manera explícita qué factores permiten definir el grado de similitud de individuos en términos de umbral de lactato. Es más, dichos factores (o su grado de influencia) podrían variar en función de los individuos y, por tanto, de los grupos. Una vía para poder abordar esta falta de conocimiento explícito es, precisamente, aplicar técnicas de clusterización con el fin de generar grupos o clústeres a partir de entradas o características relevantes de los individuos en términos de umbral de lactato.

Tal y como se especificará más adelante, entre los factores importantes a la hora de clusterizar, destaca en primer lugar la selección de las entradas o características a utilizar, ya que los grupos formados van a depender directamente de las mismas. Asimismo, se debe definir la medida de similitud a utilizar y el algoritmo de clusterización propiamente dicho [1]. Además, se debe tener en cuenta si los datos a considerar están formados por valores estáticos o por series temporales, ya que afectará a los métodos y a su aplicación.

Para la realización del análisis planteado en este trabajo se ha utilizado la toolbox Statistics and Machine Learning Toolbox™ de Matlab™ .

4.3.1 Características o entradas para clusterizar

A partir del conocimiento de los expertos de dominio, en este trabajo se consideran dos características o entradas de cara a analizar la viabilidad de esta propuesta: una característica estática, concretamente el *Peak Treadmill Speed* PTS, uno de los mejores indicadores del rendimiento de un deportista y que se define como el punto de intensidad de finalización del ejercicio [7]; y una característica dinámica o serie temporal, la frecuencia cardiaca (HR) durante la prueba de lactato, la cual se encuentra relacionada con el umbral de lactato [7-8]. Cabe destacar que el comportamiento de ambas características se encuentra relacionado con el estado de forma del corredor. Así, los cambios de estado de forma de un mismo corredor darán lugar a cambios en el comportamiento de PTS y HR y, por ende, a la identificación de individuos similares diferentes. Por tanto, este enfoque también permite abordar la estimación del umbral de lactato ante los cambios de estado de forma de los corredores.

4.3.2 Algoritmo de clusterización

Los principales grupos de algoritmos de clusterización válidos para características estáticas y series temporales son tres: jerárquicos, particionales y basados en densidad [1]. Para la realización de este trabajo se ha seleccionado el primer grupo, el cual ha demostrado un buen comportamiento con series temporales de extensión limitada. Los algoritmos jerárquicos buscan construir una jerarquía de grupos. Como resultado, el algoritmo ofrece un diagrama arborescente o *dendrograma* que representa de manera gráfica las relaciones entre los clústeres. Una de las ventajas de este algoritmo es su versatilidad en comparación con los otros tipos principales de algoritmos [10]. Existen dos tipos de estrategias: la aglomerativa o *Bottom-Up*, y la divisiva o *Top-Down*. En este trabajo se ha aplicado el enfoque aglomerativo mediante las funciones disponibles en Matlab. Asimismo, se ha utilizado el método del centroide para el cálculo de la distancia entre clústeres.

4.3.3 Medida de similitud

Todo algoritmo de clusterización requiere de una medida de similitud que permita cuantificar la mayor o menor distancia entre las características o entradas utilizadas para clusterizar. Si bien existen otras propuestas, las dos medidas de similitud por excelencia son la distancia euclídea y *Dynamic Time Warping* DTW. La selección de la medida tiene un protagonismo especial en el caso de series temporales o secuencias de datos ya que aspectos como el tamaño de la serie temporal y las características temporales de los fenómenos a clusterizar influyen directamente en la decisión. Mientras DTW es adecuada para clusterizar series temporales de distinto tamaño y fenómenos que no dependen del instante temporal en el que se producen, la distancia euclídea requiere de series temporales del mismo tamaño ya que realiza las mediciones de distancia por pares de puntos que se encuentran en el mismo instante. Si bien es cierto que la duración de la prueba de lactato depende del último escalón alcanzado por el atleta durante la prueba, dando lugar a curvas de distinta longitud, en este trabajo ha sido posible utilizar la distancia euclídea tras preprocesar las curvas de lactato. Concretamente, ha sido posible igualar la longitud de las curvas de lactato estandarizando el eje de velocidad (es decir, poniendo los valores del eje de abscisas de la curva de lactato en la misma escala). Se debe tener en cuenta que todas las curvas de lactato presentan una forma convexa con independencia de la velocidad máxima alcanzada, y el umbral de lactato no depende del valor absoluto de la intensidad del ejercicio, sino que se encuentra en la zona del punto de tangencia [8].

https://doi.org/10.17979/spudc.9788497497169.071

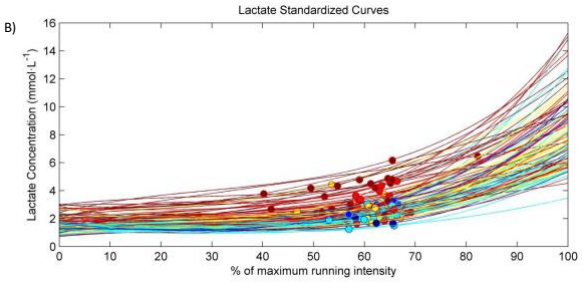

Figura 3: Curvas de lactato con eje de velocidades estandarizado

4.4 MÉTRICA DE EVALUACIÓN

En lo que concierne especialmente a las técnicas de clusterización, cabe señalar que las métricas habitualmente utilizadas para evaluar la calidad de los clústeres, como son Silhouettes, SSE… [16], son métricas no supervisadas al desconocerse la distribución real de los grupos. Dichas métricas pretenden reflejar dos aspectos: la cercanía entre los elementos pertenecientes a un mismo clúster (distancia intra-clúster) y la lejanía entre los elementos pertenecientes a distintos clústeres (distancia inter-clúster). Sin embargo, la falta de conocimiento previo sobre los grupos, desde el punto de vista de la problemática concreta a resolver, introduce un alto grado de incertidumbre sobre el clasificador no supervisado obtenido [15]. De hecho, han surgido alternativas mixtas de *clusterización supervisada* [2] y *aprendizaje parcialmente supervisado* [15], que tratan de aunar las ventajas de una y otra opción, a través de estrategias tales como el etiquetado parcial de la base de datos o la inclusión de conocimiento sobre el problema, de manera que este conocimiento permita indirectamente guiar el proceso de clusterización.

En este sentido, en el presente trabajo cabe señalar que ha sido posible aplicar una perspectiva de evaluación supervisada ya que, si bien es cierto que la distribución de los grupos o clústeres de individuos similares es desconocida, se cuenta con una base de datos experimental que, entre otros, contiene los umbrales de lactato reales y, por tanto, permite conocer si una estructura de grupos es representativa de la similitud de los individuos en mayor o menor medida.

Tabla 1: Márgenes de error aceptable para el umbral de lactato

Ritmo de carrera (min/h)	**Error máximo**	
	±(s/km)	**±(%)**
[3,5 - 4)	5	2,5
[4 - 4,5)	10	4,2
[4,5 - 5)	15	5,5
≥ 5	20	6,6

Por tanto, de cara a poder analizar los resultados de ambas perspectivas, se define una métrica de evaluación R(%). Para ello, se tienen en cuenta los márgenes de error aceptable según el ritmo de carrera del atleta en el umbral de lactato definidos en la Tabla 1, tal y como se justifica en [8].

Así, la métrica ofrece el porcentaje de aciertos respecto a los márgenes de error definidos según el caso (ver ecuación 3).

$$\hat{R}(\%) = \frac{100}{N} \sum_{i=1}^{N} B_i \qquad (3)$$

Donde N es el número total de individuos de la base de datos (105), y B se define de la siguiente manera

$$B \begin{cases} 1, \left|UL - \widehat{UL}\right| \leq Error\ máximo \\ 0, \left|UL - \widehat{UL}\right| > Error\ máximo \end{cases}$$

El cálculo del porcentaje de aciertos R(%) se realizará mediante *validación cruzada dejando uno fuera (Leave-one-out cross-validation)* siguiendo la estrategia ilustrada en la Figura 1: si X es el individuo *desconocido* (es decir, aquél para el cual se quiere realizar la estimación), por cada estructura de clústeres o grupos obtenida, se seleccionará el clúster al que pertenece X, de manera que se tomarán los datos necesarios para calcular el factor $\widehat{r_C}$ de la base de datos de todos los individuos que pertenecen a dicho clúster, exceptuando X. A partir del factor estimado $\widehat{r_C}$ y de los datos v_F y v_i pertenecientes a X, se calculará el umbral de lactato estimado $\widehat{UL}$, y se comparará con el UL real, de manera que será computado como caso de acierto si el error se encuentra dentro de los límites establecidos en la Tabla 1 (es decir, B_i=1). Este cálculo se realizará para los 105 individuos de la base de datos, obteniendo el porcentaje de aciertos correspondiente por cada estructura de clústeres o grupos considerada.

5 RESULTADOS

En esta sección se muestran y analizan los resultados obtenidos para las dos perspectivas planteadas: clasificación y clusterización. Tal y como se ha comentado, en el primer caso la base de datos ha sido clasificada según el último escalón alcanzado durante la prueba de lactato. Concretamente, se han definido seis etiquetas o clases entre los escalones 14,5 y 19,5 km/h. En el segundo caso, la base de datos ha sido clusterizada, por un lado, con la *serie temporal* HR, y por el otro con el valor PTS de los atletas de la base de datos.

Concretamente, en el análisis se han considerado entre 4 y 10 clústeres definidos a partir de los dendrogramas resultado del proceso de clusterización.

La Figura 4 muestra el porcentaje de aciertos R(%) por cada estructura de grupos o clústeres considerada (k[4-10]) en los casos de clusterización (HR, PTS), además del resultado obtenido para 6 escalones. Tal y como se puede observar, en términos globales el porcentaje de

aciertos se encuentra entre aproximadamente el 85,5% y el 91,5%. Entre las características comparadas, en términos generales, HR resulta ser la característica con menor capacidad de representación de individuos similares, mientras que la clasificación por escalones ha dado lugar al mayor porcentaje de aciertos. En el caso de la clusterización mediante PTS, alcanza porcentajes de aciertos cercanos al mejor caso para k[6-8].

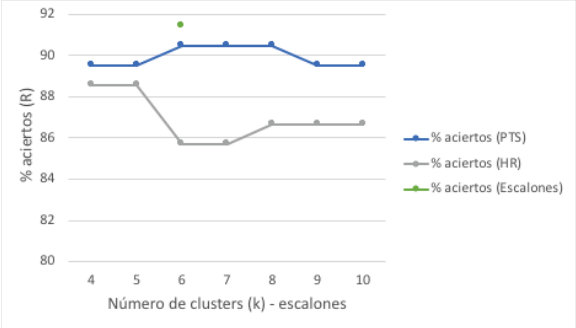

Figura 4: Porcentaje de aciertos R(%) por cada estructura de grupos o clústeres

Asimismo, se ha realizado un análisis del acierto o fallo por cada uno de los individuos de la base de datos y por cada estructura de grupos considerada (en la Figura 5 se ilustra con 23 de los 105 atletas).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1																								
2		Athleta																						
3	Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
4	HR k=4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
5	HR k=5	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
6	HR k=6	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	
7	Grandes	0	1	1	0	2	0	0	2	0	1	0	1	0	0	0	0	0	0	1	1	1	0	1
8	HR k=7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0
9	HR k=8	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
10	HR k=9	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	1	0
11	HR k=10	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0
12	Pequeños	1	1	1	1	1	0	0	1	1	0	2	0	0	0	0	0	2	1	2	2	0	2	0
13		1	3	3	1	5	0	0	5	1	2	2	2	0	0	0	0	2	1	4	4	2	2	2
14	PTS k=4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	PTS k=5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
16	PTS k=6	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
17	Grandes	0	0	1	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0
18	PTS k=7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	PTS k=8	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	PTS k=9	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
21	PTS k=10	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
22	Pequeños	2	0	0	0	0	1	2	1	0	0	0	1	0	0	1	1	0	1	0	0	0	0	0
23	Escalones	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Figura 5: Resultados por individuo

Así, se ha podido observar que para todos los casos de clusterización ha sido posible encontrar al menos un clúster a partir del cual hacer una estimación certera del umbral de lactato (a diferencia de la clasificación por escalones, en la que se considera una única estructura de grupos y, por tanto, la estimación ha sido fallida para el 8,5%).

Además, en los casos de clusterización, se han analizado los casos de fallo de estimación por individuo en función de la característica o entrada considerada (HR, PTS) y del tamaño del clúster. Para ello, se han dividido los clústeres en "grandes" G, para k[4, 6] y "pequeños" P, para k[7, 10]. En el análisis se debe tener en cuenta que el algoritmo de clusterización aplicado es jerárquico y, por tanto, los clústeres más grandes se forman acumulando clústeres de menor tamaño siguiendo el dendrograma correspondiente (ver ejemplo de la Figura 6).

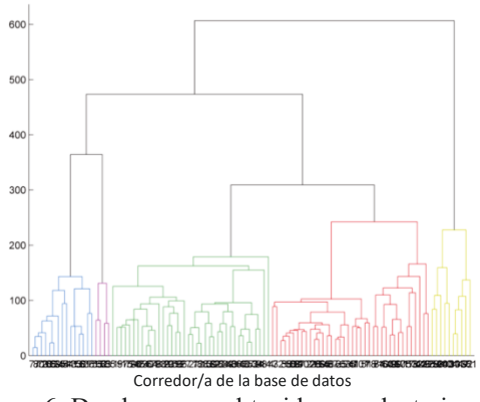

Figura 6: Dendrograma obtenido con clusterización HR

Los resultados muestran que el número de estimaciones fallidas ocurridas para alguna de las estructuras de clústeres k[4, 10] es algo superior para clusterización mediante HR que para clusterización mediante PTS (96 y 74 casos, respectivamente, de las 735 estimaciones realizadas por cada característica HR y PTS). Las 96 estimaciones fallidas se han dado para 56 individuos en clusterización mediante HR, mientras que los 74 fallos con PTS corresponden a 54 individuos. Sobre sendos totales de individuos con casos de fallo (56 en HR y 54 en PTS), en la Figura 7 se muestra la proporción de individuos con casos de fallo solo con clústeres grandes (G: k[4-6]), solo con clústeres pequeños (P: k[7-10]), y tanto con clústeres grandes como con pequeños (GyP: k[4-10]).

En el caso de HR, se han producido fallos de estimación tanto con clústeres de tamaño grande como pequeño para un mismo individuo en el 37,5% de los casos, mientras que para PTS el porcentaje ha sido inferior, 20,3%. Así, estos fallos pueden deberse a la poca o despreciable influencia de la característica considerada (HR o PTS) en el umbral de lactato de esos individuos concretos, lo que de nuevo sugiere la menor capacidad de representación de individuos similares de HR en comparación con PTS. En cuanto al porcentaje de individuos que han tenido estimaciones fallidas solo con clústeres de tamaño pequeño (y no de tamaño grande), se ha producido un 50% con PTS y un 39,3% con HR. Por tanto, en estos casos han sido necesarios tamaños de clúster grandes para realizar una estimación certera del umbral de lactato, lo que puede significar que, para esos individuos, el grado de influencia de la característica utilizada para clusterizar, si bien existe, no es muy elevada. Por último, en el 23,21% de los individuos con estimaciones fallidas con HR, y en el 29,6% de los individuos con estimaciones fallidas con PTS, la estimación para un mismo individuo ha fallado con clústeres grandes, pero ha sido certera con pequeños,

lo que permite pensar que en esos casos el grado de influencia en el umbral de lactato de la característica utilizada para clusterizar es muy significativa.

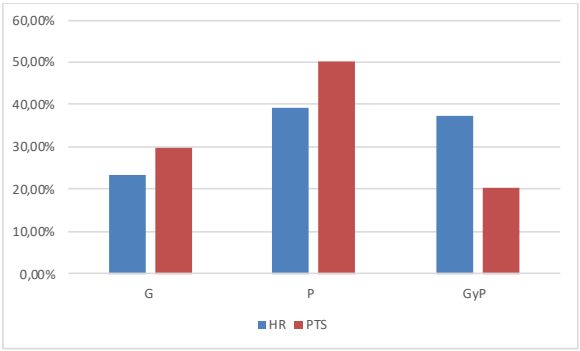

Figura 7: Proporción de individuos con estimaciones fallidas solo con clústeres grandes (G: k[4-6]), solo clústeres pequeños (P: k[7-10]), y con clústeres grandes y pequeños (GyP: k[4-10])

El resultado del análisis realizado permite pensar en establecer en un futuro una métrica de evaluación M_C de los clústeres generados que permita inferir la probabilidad de acierto del umbral del lactato para el individuo nuevo. Para ello, será fundamental la identificación de las características o *features* relevantes mediante técnicas de Ingeniería de Características.

6 CONCLUSIONES

En este trabajo se explora la aplicación de dos enfoques, uno basado en la clasificación y otro en la clusterización, para agrupar corredores de resistencia con el objetivo de poder generar grupos o clústeres que aglutinen individuos similares en términos de umbral de lactato, y así poder realizar estimaciones personalizadas de dicho umbral. Para la clasificación de los individuos, se ha tenido en cuenta directamente el último escalón alcanzado durante la prueba de lactato, mientras que, en el caso de la clusterización, el análisis se ha realizado con dos características relevantes en términos del umbral de lactato: HR y PTS. Si bien los resultados muestran, en términos globales, un mayor porcentaje de aciertos del umbral de lactato cuando se aplica la perspectiva de clasificación, un análisis individualizado de los resultados permite definir una línea futura de investigación en la que se defina una métrica de evaluación de los clústeres que permita estimar la probabilidad de acierto del umbral de lactato de un individuo nuevo mediante los datos de los individuos similares del clúster al que pertenece.

Agradecimientos

Este trabajo ha sido financiado por Grupo Campus S.L. [proyectos Universidad-Empresa LACTATUS 2017, LACTATUS 2016 y LACTATUS]; el Departamento de Desarrollo Económico y Competitividad del Gobierno Vasco [Gaitek 2015]; la Universidad del País Vasco UPV/EHU [proyectos GIU18/162, PPG17/56 y PPG/17/40]; y el Departamento de Educación del Gobierno Vasco [beca PRE 2015 1 0129], así como al Programa ERASMUS MUNDUS PANTHER (Phd agreement No. PN/TG1/AUT/PhD/06/2017). Agradecemos el apoyo de los miembros del Departamento de Fisiología y del Departamento de Educación Física y Deporte. Asimismo, se quiere agradecer la ayuda de Iñigo Meabe, Mikel Echezarreta, Jon Larruskain y Ainhoa Insunza.

English summary

APPLICATION OF SUPERVISED AND UNSUPERVISED CLASSIFICATION APPROACHES FOR THE ESTIMATION OF LACTATE THRESHOLD OF ENDURANCE RUNNERS

Abstract

Nowadays, running is one of the most popular sports and it is considered a healthy habit in everyday life. This sport practice is also encouraged by the celebration of multiple popular competitions, which are driving a growing interest in improving athletic performance as well as the safety of the runners. In this sense, in the last decade the use of Smart Watches to monitor physiological variables is becoming more and more common. However, not all the physiological variables of interest can be monitored, among them the lactate, whose measurement requires blood sampling. The lactate curve enables to extract the lactate threshold, which is essential for training-load prescription and assessment. In this work the design and development of a virtual lactate sensor is proposed. More specifically, this article analyzes the applicability of grouping techniques as an alternative towards a personalized estimation of the lactate threshold.

Keywords: lactate threshold, clustering, classification, personalized estimation, running.

Referencias

[1] Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y., (2015) "Time-series clustering: a decade Review", *Information Systems*, 53, pp. 16-38.

[2] Bair, E. (2013). "Semi-supervised clustering methods", *Wiley interdisciplinary reviews. Computational statistics,* 5, pp. 349-361.

[3] Bandodkar, A.J., Jia, W., Yardımcı, C., Wang, X., Ramirez, J., Wang, J., (2015) "Tattoo-Based Noninvasive Glucose Monitoring: A Proof-of-Concept Study" *Anal. Chem.*, 87 (1), pp. 394-398.

[4] Borges, N. R., Driller, M. W., (2016) "Wearable lactate threshold predicting device is valid and reliable in runners" *Journal of Strength and Conditioning Research*, 30 (8), pp. 2212-2218.

[5] Derbyshire, P.J., Barr, H., Davis, F., Higson, S.P.J., (2012) "Lactate in human sweat: a critical review of research to the present day" *Journal of Physiological Sciences*, 62, pp. 429-440.

[6] Erdogan, A., Cetin, C., Goksu, H., Guner, R., Baydar, M. L., (2009) "Noninvasive detection of the anaerobic threshold by a neural network model of the heart ratework rate relationship" *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology,* 223 (3), pp. 109-115.

[7] Etxegarai, U., Insunza, A., Larruskain, J., Santos-Concejero, J., Gil, S.M., Portillo, E., Irazusta, J., (2018) "Prediction of performance by heart rate-derived parameters in recreational runners" *Journal of Sports Sciences,* 36 (18), pp. 2129-2137.

[8] Etxegarai, U., Portillo, E., Irazusta, J., Arriandiaga, A., Cabanes, I., (2018) "Estimation of lactate threshold with machine learning techniques in recreational runners" *Applied Soft Computing*, 63, pp. 181-196.

[9] Etxegarai, U., Portillo, E., Irazusta, J., Koefoed, L. A., Kasabov, N., (2019) "A heuristic approach for lactate threshold estimation for training decision-making: An accessible and easy to use solution for recreational runners" *arXiv:1903.02318*.

[10] Jain, A.K., Murty, M.N., Flynn, P.J., (1999) "Data Clustering: A Review" *ACM Computing Surveys.*, 31 (3), pp. 264-323.

[11] Jia, W., Bandodkar, A.J., Valdés-Ramírez, G., Windmiller, J.R., Yang, Z., Ramírez, J., Chan, G., Wang, J., (2013) "Electrochemical Tattoo Biosensors for Real-Time Noninvasive Lactate Monitoring in Human Perspiration" *Anal. Chem.*, 85 (14), pp. 6553-6560.

[12] López Chicharro, J., Aznar Laín, S., Fernández Vaquero, A., López Mojares, L. M., Lucía Mulas, A., Pérez Ruiz, M., (2004) Transición aeróbica-anaeróbica: concepto, metodología de determinación y aplicaciones, 1st Edition, Master Line & Prodigio.

[13] Proshin, A. P., Solodyannikov, Y. V., (2013) "Mathematical Modeling of Lactate Metabolism with Applications to Sports" *Automation and Remote Control*, 74 (6), pp. 1004-1019.

[14] Santos-Concejero, J., Granados, C., Irazusta, J., Bidaurrazaga-Letona, I., Zabala-Lili, J., Gil, S. M., (2013) "Onset of blood lactate accumulation as a predictor of performance in top athletes" *RETOS. Nuevas tendencias en Educación Física, Deporte y Recreación*, 23, pp. 67-67.

[15] Schwenker, F., Trentin, E., (2014) "Pattern classification and clustering: A review of partially supervised learning approaches" *Pattern Recognition Letters*, 37, pp. 4-14.

[16] Tan, P.N., Steinbach, M., Kumar, V., (2005) Introduction to Data Mining, Addison-Wesley.

[17] Vijayalakshmi, P., Priya, N., (2016) "K-NN Classification in Integrated Method for Personalized Modelling in Bio-Medical Applications" I*nternational Journal on Advanced Computer Theory and Engineering*, 5 (2), pp. 24-27.