Original research / Artículo original / Pesquisa original - Tipo 1

Towards an automatic detection system of sports talents: An approach to Tae Kwon Do

Román Alcides Lara Cueva / ralara@espe.edu.ec

Alexis Darío Estévez Salazar / adestevez1@espe.edu.ec

Universidad de las Fuerzas Armadas, Sangolquí-Ecuador

ABSTRACT Tae Kwon Do is a Korean martial art included as an Olympic sport, where several tools have been developed from the engineering point of view,mainly focused on improving the capacity of the athletes. Nevertheless, there is a breach in the selection process of high performance athletes. For this reason, this research was focused on developing a system based on the information of the classification for the athletes in the Tae Kwon Do Ecuadorian Federation by using the wrapper and embedded modes and the Decision Tree and Support Vector Machines machine learning algorithms. These algorithms and modes were used to assess the different factors considered in this classification. The main contribution of this work is to provide a support system for the selection of these athletes.

KEYWORDS Tae Kwon Do; machine learning; wrapper; embedded; decision tree; support vector machine.

Hacia un sistema de detección automática de talento deportivo: una aplicación al Tae Kwon Do

RESUMEN El Tae Kwon Do es un arte marcial coreano reconocido como deporte olímpico, para el cual se han desarrollado diferentes herramientas desde la ingeniería, principalmente enfocadas en mejorar la capacidad de los competidores. Sin embargo, existe una brecha en el proceso de selección de atletas de alto rendimiento. Por ello, está investigación se enfocó en desarrollar un sistema basado en la información de la clasificación de los deportistas de la Federación Ecuatoriana de Tae Kwon Do, utilizando los métodos wrapper y embedded y los algoritmos Decision Tree y Support Vector Machine para la valoración de los diferentes factores considerados en dicha clasificación. La principal contribución de este trabajo es proporcionar un sistema de apoyo objetivo para la selección de dichos atletas.

PALABRAS CLAVE Tae Kwon Do; aprendizaje de máquina; *wrapper*; *embedded*; árbol de decisiones; máquinas de soporte vectorial.

Em direção a um sistema de detecção automática de talento esportivo: uma aplicação para o Taekwondo

RESUMO O Taekwondo é uma arte marcial coreana reconhecida como um esporte olímpico, para o qual foram desenvolvidas diferentes ferramentas a partir da engenharia, focadas principalmente em melhorar a habilidade dos atletas. No entanto, existe uma lacuna no processo seletivo para atletas de alto rendimento. Portanto, esta pesquisa se focou no desenvolvimento de um sistema baseado na informação da classificação dos atletas da Federação Equatoriana de Tae Kwon Do, utilizando os métodos wrapper e embedded e os algoritmos Decision Tree e Support Vector Machine para a avaliação dos diferentes fatores considerados na referida classificação. A principal contribuição deste trabalho é fornecer um sistema de apoio objetivo para a seleção dos atletas.

PALAVRAS-CHAVE Taekwondo; aprendizagem de máquina; *wrapper*; *embedded*; máquinas de suporte vetorial.

Received / Recepción: September 11, 2018 - Accepted / Aceptación: September 30, 2018 *doi:10.18046/syt.v16i47.3213*

I. Introduction

The use of Machine Learning [ML] theory has been extended to several emerging areas of study, such as data security and commerce, among others (Trejo & Miramá, 2018; Urcuqui & Navarro, 2016; Vergara, Martínez & Caicedo, 2017). In such a way, sport is effectively combined with ML Theory due to the large amount of data that can be extracted from a singular athlete or team. In this sense, ML is one of the most used theories for analysis in sports, which has focused on sport performance (Alderson, 2015), diagnosis of sport injuries (Zelic, Kononenko, Lavrac, & Vuga, 1997), and forecasting sport results (Valero, 2017).

Tae Kwon Do is a well known Korean martial art and Olympic combat sport which stands out for the variety and the impressive of its kicking techniques. In such a sense, several proposes have been developed to improve the competitors training by using a motion system with body and visual sensors and ML for analysis (Kwon & Gross, 2005). A hybrid approach sensing technique in conjunction with Hidden Markov Model [HMM] (Kwon, 2013) and a humanoid robot able to interact with athletes, in order to give instructions and improve training (Muscolo & Recchiuto, 2016). In the mean time, there exists works focused on the athlete to analyze the complex techniques in contact sports by using video frames and Deep Learning [DL] for predicting the action to be executed (Kong, Wei, & Huang, 2018). Furthermore, an approach to develop a dynamic evaluation of Tae Kwon Do by using the classification method of Genetic Algorithms with Support Vector Machine [GA-SVM] was proposed by Zhong, Hung, Yang, and Huang (2016) .

However, experts consider there is a gap in the process of athlete selection according to their expectations and reality, and the best of our knowledge, no principled studies have been conducted to recognize athletes in Tae Kwon Do and identify the main features of athletes with high competitive performance. For this reason, the aim of this paper is to develop a classification based system for determining the key features for identifying athletes towards a high performance in Tae Kwon Do; in order to fill out this paper, we apply feature selection and classification stages to data provided by the Federación Ecuatoriana de Tae Kwon Do [FETKD]. For the former stage, we propose to use wrapper and embedded methods, as long as for the next stage, supervised classification was considered, by using two well-known algorithms such as Decision Tree [DT] and Support Vector Machine [SVM]. This approach could allow us to make decisions –with reliable results as possible– about athletes suitability with major expectation of high performance.

The main contribution of this work is to provide a support system for the athlete selection based on exports opi-

I. Introducción

El uso de la teoría de aprendizaje de máquina [ML, *Machine Learning*] se ha extendido a diversas áreas de estudio, tales como la seguridad de datos y el comercio (Trejo & Miramá, 2018; Urcuqui & Navarro, 2016; Vergara, Martínez & Caicedo, 2017). Puntualmente, el deporte ha sido combinado efectivamente con la teoría de ML debido a la gran cantidad de datos que pueden extraerse de un deportista o equipo en particular. Bajo este contexto, ML es una de las teorías más utilizadas para el análisis en el deporte y se ha enfocado en el desempeño de los deportistas (Alderson,2015), en el diagnóstico de lesiones deportivas (Zelic, Kononenko, Lavrac, & Vuga, 1997) y en la predicción de resultados (Valero, 2017).

El Tae Kwon Do es una conocida disciplina coreana de artes marciales y un deporte olímpico que sobresale por lo impresionante de sus técnicas de patada. Por esto, se han desarrollado diversas propuestas para mejorar el entrenamiento de los competidores al utilizar un sistema de movimiento con sensores corporales y visuales, junto con ML para el análisis de datos (Kwon &Gross, 2005).Un ejemplo de esto es el enfoque híbrido de técnicas de teledetección en conjunto con el uso del modelo oculto de Márkov [HMM, *Hidden Markov Model*] (Kwon, 2013) y un robot humanoide capaz de interactuar con los deportistas para darles instrucciones y mejorar el entrenamiento (Muscolo & Recchiuto, 2016). Por otra parte, existen trabajos enfocados en el atleta para analizar las complejas técnicas en deportes de contacto que utilizan tramas de video y aprendizaje profundo [DL, Deep Learning] para predecir la acción a ejecutar (Kong, Wei, & Huang, 2018). Zhong, Hung, Yang, y Huang (2016), por su parte, propusieron un enfoque para desarrollar una evaluación dinámica del Tae Kwon Do utilizando los métodos de clasificación de los algoritmos genéticos y máquinas de vectores de soporte.

Sin embargo, expertos consideran que hay una falencia en el proceso de selección de atletas de acuerdo con sus expectativas y realidad y, por lo que los autores pudieron investigar, no hay estudios llevados a cabo para reconocer atletas en Tae Kwon Do e identificar sus características principales con un alto desempeño competitivo. Por esta razón, el objetivo de esta investigación fue desarrollar un sistema de clasificación para determinar las características clave para identificar atletas con potencial de alto rendimiento en la disciplina. Para esto, se llevó a cabo la selección de características y etapas de clasificación a datos suministrados por la Federación Ecuatoriana de Tae Kwon Do [FETKD]. Se propuso utilizar los métodos wrapper (envoltorio) y embebido, mientras que para la siguiente etapa, se consideró la clasificación supervisada, al utilizar dos algoritmos bien conocidos como árboles de decisión [DT, Decision Trees] y máquinas de vectores de soporte [SVM, *Support Vector Machines*]. Este enfoque permitió tomar decisiones con resultados fiables acerca de la idoneidad de los atletas con mayor expectativa de un alto desempeño.

La principal contribución de este trabajo es proveer un sistema de soporte para la selección de atletas basado en opiniones expertas, donde se identifique a los mejores candidatos y se

extraigan características clave que pueden ser utilizadas para entrenamiento específico orientado a mejorar las debilidades del atleta obtenidas por el sistema. Se plantea hacer el seguimiento de un atleta de alto rendimiento enfocado en los Juegos Olímpicos y otros eventos, nacionales e internacionales, donde los atletas del equipo nacional puedan participar desde una temprana edad.

El resto del documento está organizado de la siguiente manera: la sección II presenta los materiales y métodos empleados para el set de características y el preprocesamiento, además de la selección de características y clasificadores evaluados por su desempeño; la sección III presenta los resultados experimentales del trabajo y la sección IV discute las conclusiones obtenidas de esta investigación.

II. Materiales and métodos

En esta sección se definen los sets de características y preprocesamiento, la selección de características y clasificadores y las métricas para evaluar su desempeño, como se indica en la Figura 1. El set de diseño fue propuesto y extraído por expertos con base en sus experiencias en la selección de deportistas en la FETKD, además, en la etapa de selección de características se introdujo el uso de métodos wrapper-embebidos en unión con SVM y DT como algoritmos de clasificación. Los métodos embebidos y wrapper son actualmente usados para seleccionar el mejor set de características. Por esta razón se eligió un método de *wrapping* basado en eliminación de características recursivo [RFE, *Recursive Feature Elimination*], que cual puede lograr una mayor exactitud [A, *Accuracy*] en la clasificación, mientras que un método embebido utiliza selección de características y clasificadores en conjunto para características clave de aprendizaje, lo cual contribuye a mejorar A y evitar el sobreajuste (Liu, Wang, Zhao, Shen, & Konan, 2017; Blum & Langley, 1997; Langley, 1994). Se ha descartado el uso de filtros debido a que el uso de los métodos wrapper y embebido sobrepasa a los algoritmos de filtrado (Suto, Oniga, & Sitar, 2016). Para la clasificación se ha empleado un algoritmo bien conocido — DT— porque emula el razonamiento humano y presenta una estructura jerárquica simple para el entendimiento del usuario y la toma de decisiones (Kotsiantis,2013; Badr, Abdelkarim,Hanane, & Mohammed, n.d.). También, se ha utilizado SVM puesto que provee un alto A, lo que la hace una técnica poderosa de ML que demuestra ser un algoritmo robusto que generaliza bien a la vida real en aplicaciones ingenieriles de predicción (Parikh & Shah, 2016; Shi, Duan, Ma, & Weng, 2012; Zhang, 2012). Los algoritmos de ML deben ser evaluados respecto del desempeño, consecuentemente se han elegido métricas asociadas a dichos algoritmos (Lara, 2015). Para el desarrollo del experimento se utilizó MATLAB 2016 en un computador con un procesador a 2.4 GHz y 8 GB de RAM.

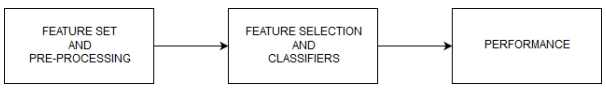

Figure 1. Block diagram for the proposed system
Diagrama de bloques para el sistema propuesto

nion whereby we can identify the best candidates. As well as the extraction of the key features, which can use for specific training oriented to improve the weak features provided by the support system. This work is according to lead an high performance athlete –firstly towards Olympic games and several events where national team could participate, locally or internationally–, from an early age.

The rest of this paper is organized as follows: in section II we define materials and methods for feature set and pre-processing, feature selection and classifiers, which are evaluated by their performance; in section III we show the experimental result and in section IV the conclusions and discussion obtained from this research.

II. Materials and Methods

In this section we detail feature set and pre-processing, feature selection and classifiers, and metrics to evaluate its performance, as depicted in Figure 1. About feature set design, this has been proposed and extracted by experts based on their experiences from athletes selection area at FETKD. Furthermore, in the feature selection stage, we have introduced the use of wrapper–embedded methods in conjunction with SVM and DT, respectively, as classification algorithms. Filter, embedded and wrapper methods are currently used to select the best set of features. For this reason, we have chosen a wrapper method based on Recursive Feature Elimination [RFE], which could achieve high classification Accuracy (A). Meanwhile, an embedded method uses feature selection and classifiers in conjunction for learning key features, which contribute to improve A and avoid over-fitting (Liu, Wang, Zhao, Shen, & Konan, 2017; Blum & Langley, 1997; Langley, 1994). We have dismissed the use of filters due to proposed wrapper and embedded methods outperforms filter algorithms (Suto, Oniga, & Sitar, 2016). Whereas the classification, we have used a well-known algorithm –DT–, because it closely resembles human reasoning and presents a simple hierarchical structure for the user understanding and decision making (Kotsiantis, 2013; Badr, Abdelkarim, Hanane, & Mohammed, 2014). Besides, we have used SVM due to it provides a higher A, and it is, for this reason, one of the most powerful techniques of ML which has been proven to be a robust algorithm that generalizes well into real life engineering applications for forecasting (Parikh & Shah, 2016; Shi, Duan, Ma, & Weng, 2012; Zhang, 2012). The ML algorithms must be assessed its performance, consequently we have chosen metrics associated to those algorithms (Lara, 2015). We have used Matlab®R2016a, a PC re(TM) i7-5500U with 2.4–2.39 GHz and 8GB of RAM for development of the experiment.

A. Feature Set and Pre-Processing Stage

An sporting talent is an athlete who possess main features required to get a higher probability of consolidation in a

sport. By this way, the traditional models of athletes selection are based on the ascription to a certain activity. It can be described two models to consider, which are the empirical or scientific model and the formative or development model (Brotons, 2005). The process to search and identify athletes potentially successful related to Tae Kwon Do is based on a mix of selection models previously mentioned. Experts have developed a recognition process based on well-defined features such as gender, category, weight and overweight, which are related to his/her physic somatotype. In addition, several tests have been developed to obtain some features like: physical abilities and techniques-tactics abilities, which are related to sports adaptation. For our understanding, gender feature determines if the candidate is male or female, while category places an athlete on their respective weight, age and gender established by the World Tae Kwon Do as described in Table 1.

Table 1. Relation between weight, age and gender to category
Relación peso, edad y género - categoría

Age (years)	*Male (weight)*	*Female (weigh)*	*Category*
Cadets (12-14)	*Even 33*	*Even 29*	*Fin*
	33-37	*29-33*	*Fly*
	37-41	*33-37*	*Bantam*
	41-45	*37-41*	*Feather*
	45-49	*41-44*	*Light*
	49-53	*44-47*	*Welter*
	53-57	*47-51*	*Light middle*
	57-61	*51-55*	*Middle*
	61-65	*55-59*	*Light heavy*
	Over 65	*Over 59*	*Heavy*
Junior (15-17)	*Even 45*	*Even 42*	*Fin*
	45-48	*42-44*	*Fly*
	48-51	*44-46*	*Bantam*
	51-55	*46-49*	*Feather*
	55-59	*49-52*	*Light*
	59-63	*52-55*	*Welter*
	63-68	*55-59*	*Light middle*
	68-73	*59-63*	*Middle*
	73-78	*63-68*	*Light heavy*
	Over 78	*Over 68*	*Heavy*
Senior (17+)	*Even 54*	*Even 46*	*Fin*
	54-58	*46-49*	*Fly*
	58-63	*49-53*	*Bantam*
	63-68	*53-57*	*Feather*
	68-74	*57-62*	*Light*
	74-80	*62-67*	*Welter*
	80-87	*67-73*	*Middle*
	Over 87	*Over 73*	*Heavy*

A. Set de características y etapa de preprocesamiento

Un talento deportivo es un atleta que posee las características principales requeridas para obtener una alta probabilidad de consolidación en un deporte. Por esto, los modelos tradicionales de selección de deportistas se basan en la adscripción hacia cierta actividad. Esta puede describir dos modelos a considerar: el empírico o científico y el formativo o de desarrollo (Brotons, 2005). El proceso de buscar e identificar atletas potencialmente exitosos para el Tae Kwon Do se basa en una mezcla de los modelos de selección previamente mencionados. Los expertos han desarrollado un proceso de reconocimiento basado en características bien definidas, como el género, la categoría, el peso y el sobrepeso, las cuales están relacionados con el somatotipo del sujeto de estudio. Adicionalmente, se han desarrollado varias pruebas para la obtención de ciertas características, como habilidades físicas y técnico-tácticas, las cuales se relacionan con la adaptación al deporte. En otras palabras, la característica del género determina si el candidato es hombre o mujer, mientras que la categoría ubica a un atleta dado su respectivo peso, edad y género establecido por la federación internacional en cabeza del deporte llamada World Tae Kwon Do, como se describe en la Tabla 1.

La relación entre estas características es entendible, puesto que el atleta pertenece a cierta categoría donde está limitado por valores máximos y mínimos de peso. El deportista puede ser clasificado en tres posibles casos: por debajo, dentro y fuera del límite. El sobrepeso tiene una relación positiva y negativa respecto de los hechos mencionados; las habilidades físicas y técnico-tácticas tienen una subdivisión relacionada con una etapa de entrenamiento u orientación, la que es necesario trabajar en el proceso deportivo. Las habilidades físicas están asociadas con las capacidades del deportista, como fuerza, velocidad, resistencia, flexibilidad y coordinación, mientras que las habilidades técnico-tácticas permiten la adopción de las etapas de entrenamiento y evaluación, como sea necesario por parte de los entrenadores, y condicionadas a las necesidades específicas del deporte. Se ha desarrollado el preprocesamiento en un contexto general para mejorar la discriminación de todas las características del set de datos (dataset) empleado al eliminar la tendencia lineal y etiquetar todas las características a utilizar. Después, la tendencia lineal ha sido suprimida al utilizar media cero y varianza igual a 1 ($\mu=0$, $\nu=1$), lo cual permite mejorar la visualización del set de características en el mismo rango. Por otra parte, el set de características se etiquetó de la siguiente manera: género (X1), categoría (X2), peso (X3), sobrepeso (X4), habilidades físicas (X5) y habilidades técnico-tácticas (X6).

B. Etapa de selección de características y clasificadores

La etapa de selección de características fue desarrollada para identificar los principales sets de características relevantes de los atletas, lo que permitió determinar las principales características a trabajar hacia un rendimiento de atleta competitivo. Un estudio de referencia se llevó a cabo para dos de los métodos de selección de características más utilizados, denominados embebido (embedded) y wrapper. El objetivo de estos

métodos es obtener matrices que provean la mayoría de la información discriminatoria para clasificar al atleta evitando la sobrecarga. Al utilizar el método embebido, se requiere utilizar un algoritmo que utiliza como criterio el de información mutua [MI, Mutual Information] entre la característica x y la salida y, como lo muestra la Ecuación 1.

$$I(x;y) = H(y) - H(y|x), \qquad (1)$$

donde la entropía marginal se define como H(y), mientras que la entropía condicional se asocia con H(y|x) entre la saliday y el set de características x a través de la generación iterativa de ejecuciones al dividir los datos y tomando ventaja de acuerdo con su importancia para la tarea de clasificación.

El algoritmo utilizado para esto fue DT, que es considerado como un algoritmo de aprendizaje supervisado no paramétrico y utilizado principalmente para problemas de regresión y tareas de clasificación. La meta de este algoritmo se orienta hacia un modelo que pueda predecir el valor de una variable al aprender reglas de decisión inferidas desde características de los datos. El parámetro libre del algoritmo DT es la profundidad de la frondosidad y debe ser ajustado con el fin de maximizar el rendimiento de la clasificación, evitando la sobrecarga en el set de entrenamiento (training set). El árbol es moldeado por un nodo raíz, nodos internos y nodos terminales. Además, se establece una regla en cada nodo, dando así confianza a producir la selección binaria y extenderla hasta el nodo final, que representa una clase.

Todas las posibles ramificaciones son dependientes de cada valor que el nodo pueda tomar. Por esto, el algoritmo genera decisiones secuenciales para predecir valores como características representativas de los datos e introduce un acercamiento basado en teoría de la información donde la elección de una característica está directamente relacionada con la entropía, la cual se describe como una medida de la incertidumbre del sistema que permite conocer la cantidad promedio necesaria de bits que deben ser adaptados a la salida del algoritmo. Este parámetro se representa por la Ecuación 2

$$\sum_{i \in C} -p_i \log_2 p_i, \qquad (2)$$

donde C describe un set de la clase a la que podría pertenecer, como ejemplo y pi es la probabilidad de que dicho ejemplo pertenezca a la clase i-ésima. Para el método wrapper se empleó el RFE, el cual tiene como base un método de eliminación en reversa, cuya operación se basa en remover características iterativas de los datos, buscando elegir las características que llevan al margen más largo de separación de clases al utilizar SVM como clasificador. En el caso descrito en este artículo, el clasificador elegido fue ν-SVM, habilitando la variación necesaria de un parámetro libre como ν, el cual controla el número de vectores de soporte. El algoritmo ν-SVM se define brevemente a continuación (el lector puede referirse a Schölkopf & Smola (2002) para más detalles). Al utilizar un set etiquetado de datos para entrenamiento (3):

The relationship between these features is understandable in a way that an athlete belongs to a category, which is limited by the higher and lower weight. The athlete could be located in three possible cases; under, into, and over the limit; the overweight has a positive and negative relation concerning the facts mentioned previously; physical and technical-tactics abilities has a subdivision related to training stage or orientation that is necessary to work into the sporting process. Physical abilities are associated to the athlete capacities as strength, speed, endurance, flexibility and coordination, while, technical-tactics abilities allow us to adopt the train stage and evaluate as necessary for coaches, conditioned by specific sport needs. We developed pre-processing in a general context to enhance the discrimination in all features on our data set, by eliminating the lineal trend and label all the features to be used. Over our case we removed lineal trend by using zero mean and variance equal to one (μ=0, ν=1), which allows improving visualization of our feature set in the same range; on the other hand, feature set is labeled such as gender (X1), category (X2), weight (X3), overweight (X4), physical abilities (X5) and technical-tactical abilities (X6).

B. Feature Selection Stage and Classifiers

Feature selection stage was developed to identify the principal sets of relevant features from athletes, which will enable us to determine the main features to work toward high competitive athlete performance. It is performed a benchmark study of two most used feature selection methods which are named "embedded" and "wrapper". The goal of this methods it to obtain matrices which provide most of the discriminative information to classify the athlete, while avoiding over-fitting. By using the embedded method, in this work is necessary to select an algorithm, which as the main criterion uses Mutual Information [MI] between feature x and the output y, as follows in Equation 1.

$$I(x;y) = H(y) - H(y|x), \qquad (1)$$

where, the marginal entropy is defined as H(y), while conditional entropy is associated with H(y|x) between output y and feature set x, through generating an iteratively builds by dividing the data taking advantage according to its importance for the classification task. The algorithm used is DT, which is considered a non-parametric supervised learning algorithm and is principally used for both, classification and regression problems. The goal of this algorithm is oriented towards a model which can predict the value of a variable by learning decision rules inferred from the data features. The free parameter of DT algorithm is the depth or leafiness and it has to be adjusted in order to maximize the classification performance, avoiding over-fitting to the

training set. The tree is shaped by a root node, internal nodes and terminal nodes; moreover, in each node a rule is established, which is the entrusted to produce the binary selection extend to the final node which represents a class. All the possible branches are dependents to each node values can take. In this way, the algorithm generates sequential decisions to predict values, as of representative features of the data. Introduces an approach based in information theory, where the choice of a feature it is directly related with entropy, which is described as a measure in a system uncertainty that allows us to know the necessary average amount of bits can be adapted to the output of the algorithm. This parameter is represented by Equation 2.

$$\sum_{i \in C} -p_i \log_2 p_i, \qquad (2)$$

where, C describes a set of the class which may belong to such an example and pi is the likelihood that given example belong to i-th class. For a wrapper method, we used RFE, which has a base on a backward elimination method. Their operation is based on iteratively removing features from data, seeking to choose the features which lead to the largest margin of class separation by using SVM as a classifier. In our case, the selected was ν-SVM, enabling the variation necessary of a free parameter ν which control the number of support vectors. The ν-SVM algorithm is defined in summary as follows (see Schölkopf & Smola (2002) for details. By using a labeled training data set (3):

$$\{x_i, y_i\}_{i=1}^{n} \qquad (3), \text{ where}$$

$$x_i \in \mathbb{R}^N \qquad (4) \text{ and}$$

$$y_i \in \{-1, +1\} \qquad (5)$$

and given a nonlinear mapping $\phi(\cdot)$ the ν-SVM methods solves (6).

$$\min_{w,\xi_i,b,\rho} \left\{ \frac{1}{2}\|w\|^2 + \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \right\} \qquad (6),$$

subject to (7) and (8)

$$y_i(\langle \phi(x_i), w \rangle + b) \geq \rho - \xi_i \qquad \forall i = 1, \ldots, n \qquad (7)$$

$$\rho \geq 0, \xi_i \geq 0 \qquad \forall i = 1, \ldots, n \qquad (8)$$

where, w and b define a linear classifier in the feature space and the positive slack variables enabling to deal with errors, it is associated with ξi . It should be taken that the appropriate choice of nonlinear mapping θ allows us guarantees that the transformed samples hold a major proba-

$$\{x_i, y_i\}_{i=1}^{n} \qquad (3), \text{ donde}$$

$$x_i \in \mathbb{R}^N \qquad (4), \text{ y}$$

$$y_i \in \{-1, +1\} \qquad (5)$$

y dado un mapeo no linear $\phi(\cdot)$, el método ν-SVM soluciona (6).

$$\min_{w,\xi_i,b,\rho} \left\{ \frac{1}{2}\|w\|^2 + \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \right\} \qquad (6), \text{ sujeto a (7) y (8)}$$

$$y_i(\langle \phi(x_i), w \rangle + b) \geq \rho - \xi_i \qquad \forall i = 1, \ldots, n \qquad (7)$$

$$\rho \geq 0, \xi_i \geq 0 \qquad \forall i = 1, \ldots, n \qquad (8)$$

donde w y b definen un clasificador lineal en el espacio característico y en las variables sueltas positivas, habilitando el tratar con errores; está asociado con ξi. Se debería considerar que la elección apropiada del mapeo no lineal θ permite garantizar que las muestras transformadas tienen una mayor probabilidad de ser separables en el espacio característico. Bajo este contexto, se identificó que las variables son controladas a través de coeficientes, lo cual provee un nuevo grado de libertad al margen. Por consiguiente, el tamaño del margen se incrementa linealmente con la variación del parámetro ρ y al ajustar ν en el rango [0;1], el algoritmo ν-SVM permite el intercambio entre el error de entrenamiento y el error de generalización, el cual se define como la frontera superior de la fracción del margen de errores y, a su vez, es la frontera inferior de la fracción de vectores de soporte. La solución óptima al problema primal (6) podría obtenerse utilizando la contraparte del problema dual al introducir (9).

$$w = \sum_{i=1}^{n} y_i \alpha_i \phi(x_i) \qquad (9)$$

mientras que la función de decisión para cualquier vector de texto x*es descrita por la Ecuación 10.

$$f(x_*) = sgn\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x_*) + b \right) \qquad (10)$$

Es posible describir restricciones en (6) como multiplicadores de Lagrange definidos por αi, donde son los vectores de soporte [SV, Support Vectors] los que entrenan las muestras xi con multiplicadores de Lagrange diferentes de cero $\alpha i \neq 0$ y el término de sesgo b calculado utilizando dichos multiplicadores ilimitados (11).

$$b = 1/k \sum_{i=1}^{k} (y_i - \langle \phi(x_i), w \rangle), \qquad (11)$$

donde k es el número de multiplicadores ilimitados de Lagrange ($0 < \alpha i < C$).

El algoritmo SVM presenta una particularidad alrededor de la función de decisión f(x), definida como función de un pequeño subconjunto de los ejemplos de entrenamiento descritos por los vectores de soporte. Estos son ejemplos cercanos del límite de decisión y caen en el margen junto con los ejemplos mal clasificados. La existencia de dichos vectores de soporte se encuentra en el origen de las propiedades computacionales de SVM y de su desempeño de clasificación competitivo. El lector puede remitirse a Guyon, Weston, Barnhill, and Vapnik (2002) para más detalles acerca del algoritmo SVM y su relación con los valores lineales y no lineales.

C. Desempeño

Esta etapa fue desarrollada para evaluar el desempeño de la clasificación. Al evaluar la determinación de los atletas etiquetados con el valor "1" que significa candidato apto y "-1" para un candidato no apto, se ejecuta la recopilación de información generada por las etiquetas reales. Los métodos establecidos para el desempeño de los clasificadores —exactitud (accuracy, A), precisión (P), sensibilidad (R), especificidad (S) y tasa de error balanceado [BER, Balanced Error Rate]— se describen con las Ecuaciones 12 a 16 respectivamente.

$$A(\%) = \frac{N_C}{N_T} \times 100, \qquad (12)$$

$$P(\%) = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100, \qquad (13)$$

$$R(\%) = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100, \qquad (14)$$

$$S(\%) = \frac{N_{TN}}{N_{TN} + N_{FP}} \times 100, \qquad (15)$$

$$BER = 1 - \frac{R + S}{2 \times 100}, \qquad (16)$$

donde:

N_C pertenece al número de patrones correctamente clasificados;

N_T hace referencia al número de patrones utilizados en la clasificación;

N_{TP} es el número de verdaderos positivos;

N_{FP} es el número de falsos positivos;

N_{TN} expresa el número de verdaderos positivos; y

N_{FN} el número de falsos negativos.

Estas medidas de desempeño fueron calculadas para cada validación empleada en todos los casos propuestos.

III. Resultados experimentales

Los resultados obtenidos a través de esta investigación permitieron realizar un análisis y enfoque hacia un atleta de alto rendimiento al seguir la identificación de características principales y a la clasificación del atleta. Fueron analizados datos de 76 atletas divididos en dos grupos, el primero corresponde al training set de los algoritmos e incluye a un total de 54 atletas, mientras que el test set tiene 22 atletas. Los últimos corresponden a lo más reciente obtenido en 2018.

bility for being linearly separable in the feature space. In this context, we can identify that the variable is controlled through coefficient, which provides a new degree of freedom to the margin. Furthermore, the size of the margin increasing linearly with the variation of the parameter ρ. Therefore, adjusting ν in the range [0;1] in the ν-SVM algorithm allows performing the trade-off between the training error and the generalization error, which is defined as an upper bound on the fraction of margin errors and is also a lower bound on the fraction of support vectors. The optimal solution of the primal problem (6) could be obtained by using its dual problem counterpart, introducing (9)

$$w = \sum_{i=1}^{n} y_i \alpha_i \phi(x_i) \qquad (9)$$

while decision function for any text vector x* is finally outlined by equation 10.

$$f(x_*) = sgn\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x_*) + b\right) \qquad (10)$$

It is possible to describe constraints in (6) as Lagrange multipliers defined by αi , being the Support Vectors [SV] those training samples xi with non-zero Lagrange multipliers αi ≠ 0; and the bias term b calculated by using the unbounded Lagrange multipliers as (11).

$$b = 1/k \sum_{i=1}^{k} (y_i - \langle \phi(x_i), w \rangle), \qquad (11)$$

where k is the number of unbounded Lagrange multipliers (0 < αi < C). SVM present a particularity around the decision function f(x), defined as a function of a small subset of the training examples described as the support vectors. Those are examples closes to the decision boundary and lie on the margin as well as those wrong-class examples. The existence of such support vectors is at the origin of the computational properties SVM and their competitive classification performance (see Guyon, Weston, Barnhill, and Vapnik (2002) for more details about the SVM algorithm related to linear and non-linear.

C. Performance

This stage was developed to evaluate the classification performance. By performing the determination of the athletes labeled with the value "1", equal to an suitable candidate, and "-1" for a not suitable candidate; an information collation is carried out by generated and real labels. The established measures for performance of classifiers –Accuracy (A), Precision (P), Sensitivity (R), Specificity (S) and Balanced Error Rate (BER)– are described with the equations 12 to 16, respectively.

$$A(\%) = \frac{N_C}{N_T} \times 100, \qquad (12)$$

$$P(\%) = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100, \quad (13)$$

$$R(\%) = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100, \quad (14)$$

$$S(\%) = \frac{N_{TN}}{N_{TN} + N_{FP}} \times 100, \quad (15)$$

$$BER = 1 - \frac{R + S}{2 \times 100}, \quad (16)$$

where:

N_C belongs to the number of patterns correctly classified;

N_T make reference to the number of the used patterns in the classification;

N_{TP} is the number of true positives;

N_{FP} is the number of false positives;

N_{TN} express the number of true negatives; and

N_{FN} the number of false negative.

We calculated these performance measures for each validation used in all the cases proposed.

III. Experimental Results

The results obtained throughout this research allows us to perform an analysis and approach toward a high-performance athlete, following the identification of main features and the athlete classification. The data to be analyzed corresponds to a total of 76 athletes, which was divided into two groups. The first group be owned by the training set of the algorithms, with a total of 54 athletes. While the test set has 22 athletes, these last are the most recent obtained in 2018.

In other words, our training set is equivalent to 71.052%, while the test set is 28.948%. This will allow us to carry out in a feasible way the feature selection and athletes detection. Enabling reliable results from the supervised algorithms and avoiding over-fitting.

By making use of a three-dimensional plane, in Figure 2a we can see the feature set surface provided, while Figure 2b presents to us the feature set surface after pre-processing. This stage works in the sense of removing lineal trend and place all the feature set on the same range, by using of μ=0 and ν=1.

The original output of athletes classification is depicted in Figure 3, which allows to compare with the output delivered by DT and SVM algorithms, described below, where, as mentioned, a value "1" is assigned to a suitable candidate, while "-1" corresponds to a non suitable candidate.

Dicho de otra manera, nuestro training set equivale a un 71.052%, mientras que el test set a 28.948%. Esto permite que la selección de características y la detección de atletas sea factible y que los resultados de los algoritmos supervisados sean confiables y se evite el ajuste excesivo.

Al hacer uso de un plano tridimensional, en la Figura 2a se puede observar el set de características superficiales provisto, mientras que en la Figura 2b se presenta el set de características superficiales después del preprocesamiento. Esta etapa es importante para remover tendencias lineales y ubicar todos los sets de características en el mismo rango al utilizar μ=0 y ν=1.

La salida original de la clasificación de los atletas se presenta en la Figura 3, donde permite la comparación con la salida entregada por los algoritmos DT y SVM. Aquí, un valor de "1" se asigna a un candidato apto y "-1" a uno no apto.

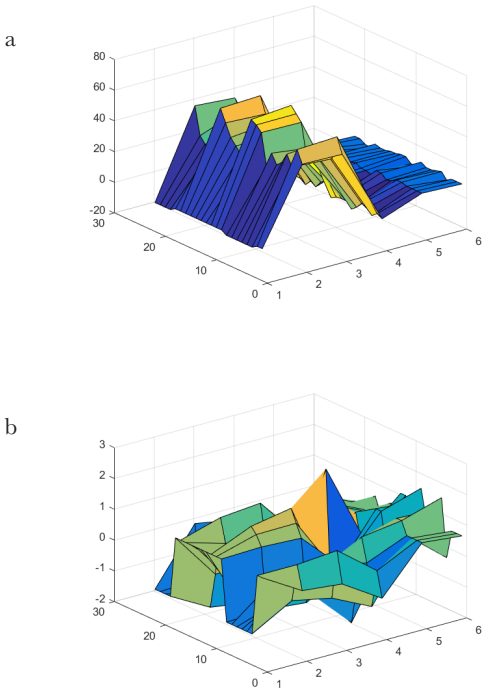

Figure 2. Surface representation of feature set: original (a) and re-processing (b) / Representación superficial del set de caractarísticas: original (a) y pre-procesado (b)

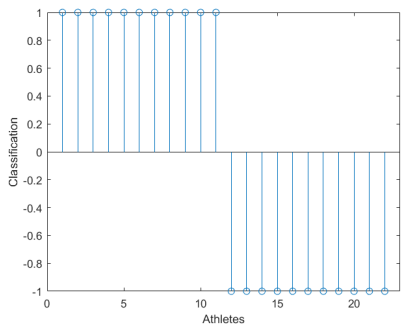

Figure 3. Original output of athletes classification
Salida original de la clasificación de los atletas

A. Resultados utilizando DT

El algoritmo DT obtuvo un modelo para la matriz de entrada al utilizar características previamente establecidas y la matriz de salida correspondiente a la clasificación de los atletas. Hizo posible inducir un árbol, como se muestra en la Figura 4, el cual eligió tres características clave. Empezando por el nodo superior con regla X5 ≥ 0.827944, seguido por X4 ≥ 0.366235 y finalmente X6 ≥ −0.0302377, lo cual hizo posible clasificar en cualquiera de las cuatro ramas.

Esta representación establece diferentes umbrales dependiendo de la amplitud, lo que hace posible identificar la elegibilidad de un atleta donde se determinó un nodo raíz X5. En caso de que el umbral fuera excedido, el atleta se consideró apto. Si el candidato es "no apto", se procede a tomar una nueva decisión.

El siguiente nodo para la toma de decisiones es X4, el cual no tiene que sobrepasar el valor de 0.36 para que el atleta no sea descartado. Esto permite una concatenación con la última característica X6, la cual puede ser mayor o igual a –0.03 para que el candidato sea apto. Al utilizar todos estos sets de características para una clasificación supervisada, se obtienen las siguientes medidas del desempeño: A%=86.3636, P%=90, R%=81.8182, S%=90.9091 y BER=0.1364. La salida de la clasificación de los atletas entregada por el algoritmo se muestra en la Figura 5. Utilizando las características principales propuestas por el algoritmo DT, se determinó que los parámetros de desempeño y clasificación son los mismos en el caso de emplear todas las características.

B. Resultados utilizando SVM

El algoritmo RFE obtuvo un modelo para la matriz de entrada al emplear las características establecidas previamente y la matriz de salida para la selección de características. Por esto, el framework de selección de características se hizo basado en el algoritmo SVM-RFE, los cuales identifican las características clave desde la principal hasta la menos importante, basándose en el peso de cada una. Para el caso de estudio, a través del empleo de este método fue posible determinar tres características clave (descritas en la Tabla 2), trasladadas a porcentaje como X4=70.820%, X6=22.289% y X2=6.891%.

Se decidió utilizar tres sets de características en los algoritmos de clasificación, los cuales fueron los sets entregados por DT, RFE, y todas las características. Además, se utilizaron dos

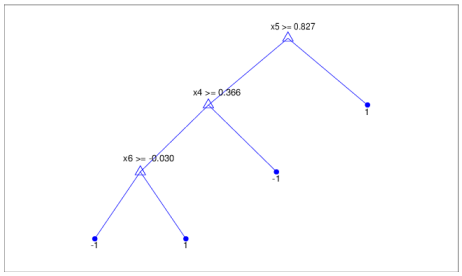

Figure 4. Tree representation considering main features, by using DT algorithm / Representación de árbol considerando las características principales al utilizar el algoritmo DT

A. Results Using DT

DT algorithm obtained a model for the input matrix by using the features established previously and the output matrix corresponding to athletes classification. It made possible to induce a tree, as depicted in Figure 4, which chosen three key features. Beginning from the top node with the rule X5 ≥ 0.827944, followed by X4 ≥ 0.366235 and finally X6 ≥ −0.0302377, which made possible to classify into any 1 of the 4 possible leafs.

This representation establishes different thresholds depending amplitude that would make it possible to identify the eligibility of an athlete. Where was determined a root node X5, in this way if threshold value it is exceeded the athlete is considered suitable, in the case of a candidate not succeeding this value proceed to take of a new decision. The next node for decision making is X4, will no have to go over the threshold value of 0.36, so that the athlete is not discarded, allowing a concatenation with the last feature X6, which could be greater than or equal to -0.03, so that the candidate be suitable. By using all the feature set for a supervised classification provides the following performance measures: A%=86.3636, P%=90, R%=81.8182, S%=90.9091 and BER=0.1364. The output of athletes classification delivered by the algorithm is depicted in Figure 5. Using main features proposed by DT algorithm, we determine that performance parameters and classification, which are the same in the case of use all features.

B. Results Using SVM

RFE algorithm obtained a model for the input matrix by using the features established previously and the output matrix for feature selection. In this sense, the feature selection framework was made based on SVM-RFE algorithms, which identify the key features from the main one toward less important feature, based on the weights of each one. For our case, through the use of this method, we determine three key features (described in Table 2), which are translated in percentage according to X4=70.820%, X6=22.289% and X2=6.891%.

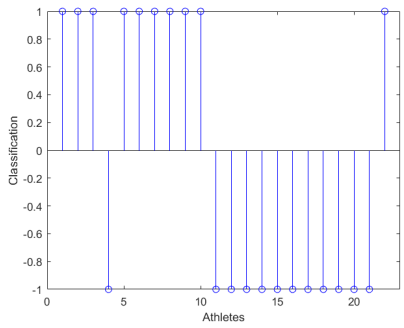

Figure 5. DT output of athletes classification
Salida del algoritmo DT para la clasificación de los atletas

Table 2. Main features delivered by using SVM-RFE algorithm
Características principales entregadas por el algoritmo SVM-RFE

Feature	*Weight*
X4	*21,460*
X6	*6,754*
X2	*2,088*

Table 3. Performance measures by using ν-SVM lineal kernel and different features sets / Medidas de desempeño al usar el kernel lineal ν-SVM lineal y distintos sets de características

v	*Feature set*	*A (%)*	*P (%)*	*R (%)*	*S (%)*	*BER*
0,29	*X4, X5, X6*	*86,363*	*90*	*81,818*	*90,909*	*0,136*
0,18	*X4, X6, X2*	*81,818*	*81,818*	*81,818*	*81,818*	*0,181*
0,24	*X1 ... X6*	*77,272*	*75*	*81,818*	*72,727*	*0,227*

Table 4. Performance measures by using ν-SVM RBF kernel and different features sets / Medidas de desempeño utilizando el kernel ν-SVMRBFy diferentes sets de características

v	*Feature set*	*A (%)*	*P (%)*	*R (%)*	*S (%)*	*BER*
0,29	*X4, X5, X6*	*86,363*	*90*	*81,818*	*90,909*	*0,136*
0,18	*X4, X6, X2*	*81,818*	*81,818*	*81,818*	*81,818*	*0,181*
0,24	*X1 ... X6*	*77,272*	*75*	*81,818*	*72,727*	*0,227*

We decided to use in classification algorithms three sets of features, which are the key sets provided by DT, RFE and all features. Furthermore, for classification we use two different kernel in ν-SVM algorithm, which are Lineal and Radial Basis Function [RBF] Kernel. The adjustment of parameter ν is carried out with a constant variation of 0.01 in the established range for the algorithm, those experimental results are described below. We detail performance measures obtained by using lineal kernel (see Table 3). Figure 6 depicts the output of athletes classification delivered by each algorithm proposed. Figure 6a represents the corresponding athletes classification provided by the ν-SVM lineal kernel algorithm, with DT feature set, where ν=0.29; Figure 6b shows the output of the algorithm with ν=0.18 by using RFE feature set; and Figure 6c define an amount of ν=0.24 for all features.

For the RBF kernel case, Table 4 depicts the measures of performance delivered by the algorithm and Figure 7 shows the athlete classification. The features employed in these algorithms are the same used in lineal kernel algorithms. Figure 7a depicts the corresponding athletes classification provided by the ν-SVM RBF kernel algorithm by using DT feature set, where 0.96 is the better value of ν. Figure 7b shows the output of the algorithm with ν=0.11 by using RFE feature set and Figure 6c define an amount of ν=0.23 for all features.

It can also seen in Figure 6a, 6b, 7a and 7b, by using the same sets previously mentioned, that there is a noticeable

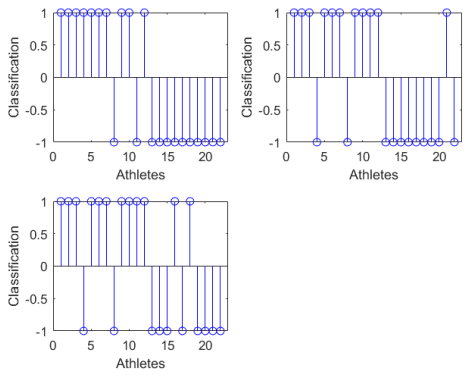

Figure 6. The output of athletes classification by using ν-SVM lineal kernel algorithm with: DT feature set (a); RFE feature set (b); and all features (c) / Salida de la clasificación de los atletas al utilizar el algoritmo de kernel lineal ν-SVMcon: set de características DT (a);set de características RFE (b);y todas las características (c)

diferentes kernels (núcleos) para clasificación en el algoritmo ν-SVM: kernel lineal y kernel de función base radial [RBF, Radial Basis Function]. La modificación en el parámetro ν se llevó a cabo con una variación constante de 0.01 dentro del rango establecido para el algoritmo, los detalles se describen a continuación. Además, se realizaron medidas detalladas de desempeño al utilizar un kernel lineal (ver Tabla 3). La Figura 6 presenta la salida de la clasificación de los atletas entregada por el algoritmo de kernel lineal ν-SVM con set de características DT y donde v=0.29. La Figura 6b muestra la salida del algoritmo con ν=0.18 al utilizar el set de características RFE y finalmente, en la Figura 6c define el valor en ν=0.24 para todas las características.

Para el caso de kernel RBF, en la Tabla 4 se presentan las medidas de desempeño entregadas por el algoritmo, mientras que en la Figura 7 se describe la clasificación de los atletas. Las características empleadas en estos algoritmos son las mismas que utilizan los algoritmos de kernel lineal. En la Figura 7a se presenta la correspondiente clasificación de atletas entregada por el algoritmo de kernel RFB ν-SVM al utilizar un set de características DT, donde 0.96 es el valor más alto de ν; en la Figura 7b se muestra la salida del algoritmo con ν=0.11 al

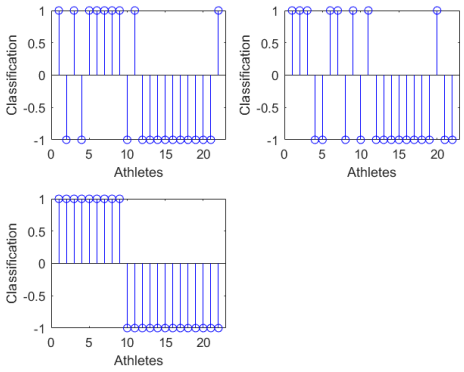

Figure 7. The output of athletes classification by using ν-SVM RBF kernel algorithm with: DT feature set (a); RFE feature set (b); and all features (c) / Salida de la clasificación de los atletas al utilizar el algoritmo de kernel ν-SVM RBF con: set de características DT (a);set de características RFE (b); y todas las características (c)

utilizar el set de características RFE; y en la Figura 7c se define el valor de ν=0.23 para todas las características.

Como se puede ver en las Figuras 6a, 6b, 7a y 7b, al utilizar los mismos sets mencionados existe una notable similitud entre la clasificación de los atletas con X4 y X6 como características comunes. La clasificación de los atletas puede representarse gráficamente con el uso de un plano tridimensional, habilitando el identificar la divisibilidad entregada por los algoritmos de acuerdo con el kernel utilizado, donde los ejes son asignados como (X4, X5, X6) a (X,Y, Z) y (X2, X4, X6) a (X,Y, Z) respectivamente. Para explicar mejor lo anterior, en la Figura 8 se presenta la clasificación de salida utilizando los kernels lineales y RBF con los sets TD y RFE. Aquí, el "+" corresponde a un atleta adecuado, mientras que el símbolo "°" indica un atleta no apto. Esta representación hace difícil identificar cuáles son las principales características y no es posible observar una gran diferencia entre los algoritmos. En las Figuras 8a y 8b se describe la correspondiente clasificación de deportistas generada por el algoritmo de kernel lineal ν-SVM al utilizar los sets de características de DT y RFE, respectivamente. En las Figuras 8c y 8d se muestra la salida del algoritmo en un plano tridimensional al utilizar los mismos sets.

similarity around athletes classification, providing X4 and X6 as common features. Nevertheless, athletes classification can be represented graphically with the use of a three-dimensional plane, enabling identify the separability delivered by the algorithms according to the used kernel, where the axes are assigned as (X4, X5, X6) to (X, Y, Z) and (X2, X4, X6) to (X, Y, Z) respectively. For our better knowledge, Figure 8 depicts the output classification by using lineal and RBF kernel, with DT and RFE sets, where (+) corresponds to a suitable athlete and (°) is a non suitable athlete. This representation makes hard to identify which are the main features and a great difference from one algorithm to another can't be observed. Figure 8a and 8b depict the corresponding athletes classification provided by the ν-SVM lineal kernel algorithm by using DT and RFE features sets, respectively. Figure 8c and 8d show the output of the algorithm in a three dimensional plane by using DT and RFE feature set.

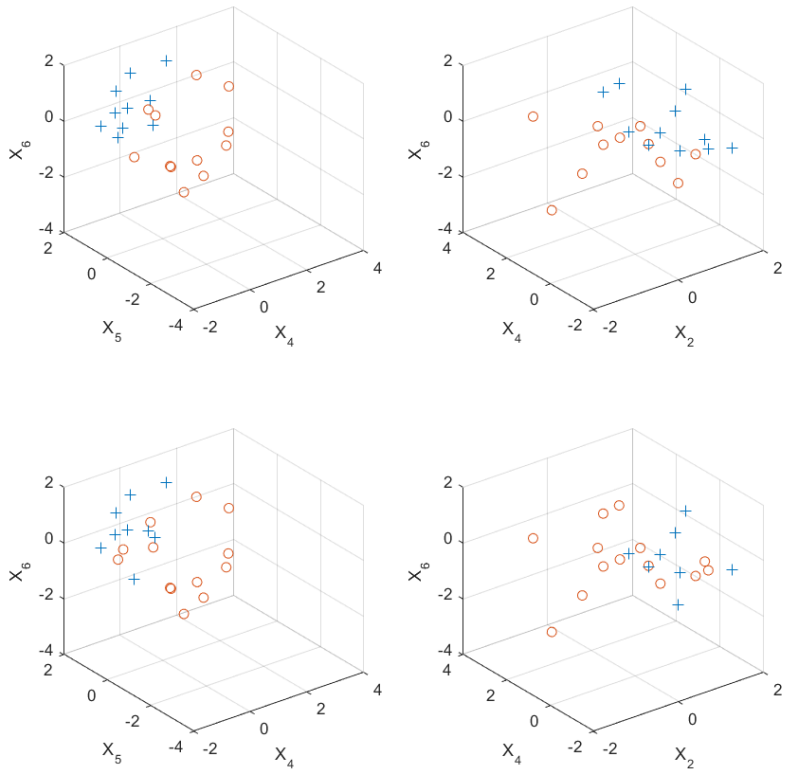

Figure 8. Figure 8. The output of athletes classification in a three-dimensional plane by using ν-SVM corresponds to: DT feature set with Lineal kernel (a); RFE feature set with Lineal kernel (b); DT feature set with RBF kernel (c); and RFE feature set with RBF kernel (d) / Salida de la clasificación de los atletas en un plano tridimensional utilizando ν-SVM en: un set de características DT con kernellineal (a);un set de características RFE con kernellineal (b);set de características DT con kernelRBF (c); y set de características RFE con kernelRFE (d)

IV. Discussion and Conclusions

In this paper we have proposed a detection of sports talents by using machine learning theory oriented toward Tae Kwon Do. First, a novel yet straightforward method has offered for feature selection and classifiers for an objective and impartial selection of athletes, which has been solved by using embedded and wrapper methods associated to DT and SVM algorithms. The analyzed scenario corresponds to the last years data corresponding to athletes from the Ecuadorian Tae Kwon Do National Team. Second, we supplied the managers with additional information about the most relevant features to be taken into account, for this purpose, the features of athletes were measured up, giving a clearer view of which is the most critical of them on a new systematic and easy-to-handle representation with the coaches. Features analysis has been allowed to detect the best candidates and identify features will be the essential item to work on it. The use of supervised algorithms makes this support system more than an athlete classification tool and is solidly based on the analysis of well-defined features. The application of these two cases of study (different theory of ML) highlights the practical convenience and usability of this approach. In our results, the analysis of feature selection showing a reduction to the half of these in both cases. It is possible identify two common features delivered by the algorithms, which are X6 and X4. For the DT algorithm, the performance measures are: A%=86.3636, P%=90, R%=81.8182, S%=90.9091 and BER=0.1364. Through the use of ν-SVM algorithm with lineal kernel, we identify the best case, which uses ν=0.29 and key feature set delivered by DT, providing the same performance measures of DT algorithm, however, it outputs a different classification. Furthermore, for the best case of ν-SVM algorithm by using RBF kernel, we obtain a value of ν=0.23 and using all features set; it provides the next performance measures: A%=90.9091, P%=100, R%=81.8182, S%=100 and BER=0.0909. Thus, the best algorithm for athletes classification in our case is associated to ν-SVM with RBF kernel, which outputs a high-performance measures. We are able to concluded that the proposed novel support system can be useful to determine the suitable athletes for the next competitions in this sport, while giving a robust and operative overview of features for athletes selection management purposes. Finally, though our formulation, it is suitable for Tae Kwon Do athletes, but it could also be useful in other combat or martial arts.

Acknowledgement

The authors gratefully acknowledge the contribution of the Universidad de las Fuerzas Armadas [ESPE] for the economical support for the development of this project under Research Grants 2013-PIT-014 and 2016-EXT-038.

IV. Discusión y conclusiones

En este documento se propuso un sistema de detección de talentos deportivos del Tae Kwon Do utilizando teoría de aprendizaje de máquina. Es un método novedoso, directo, que se presentó para la selección de características y clasificadores para una selección objetiva e imparcial de los deportistas, lo que fue solucionado al utilizar los métodos embebido y wrapper asociados con los algoritmos DT y SVM, respectivamente. El escenario analizado correspondió a los datos de los últimos años del equipo nacional ecuatoriano de Tae Kwon Do. Después, se presentó a los entrenadores información adicional acerca de las características más relevantes para tener en cuenta. Para esto, las características de los atletas se midieron, dando un claro panorama de cuáles son las más críticas en una representación sistemática y fácil de manejar para los entrenadores. El análisis de características permitió detectar a los mejores candidatos, la identificación de características fue esencial en este trabajo. El uso de algoritmos de aprendizaje supervisado hizo que el sistema de soporte fuera más que una simple herramienta de clasificación, puesto que está basada sólidamente en el análisis de características bien definidas. La aplicación de estos dos casos de estudio (diferente teoría de aprendizaje de máquina) resalta la conveniencia práctica y usabilidad de este enfoque. En los resultados encontrados, el análisis de selección de características mostró una reducción a la mitad en ambos casos. Es posible identificar dos características comunes derivadas de los algoritmos X6 y X4. Para el algoritmo DT, los valores de desempeño fueron: A%=86.3636, P%=90, R%=81.8182, S%=90.9091 y BER=0.1364. Con el uso del algoritmo ν-SVM con kernel lineal se pudo identificar el mejor caso, con un valor ν=0.29 y set de características clave entregado por DT, brindando el mismo rendimiento que el algoritmo DT, aunque con clasificación final diferente. Además, para el mejor caso del algoritmo ν-SVM utilizando un kernel RBF, se obtuvo un valor ν=0.23 y, utilizando todo el set de características, se obtuvieron los siguientes resultados: A%=90.9091, P%=100, R%=81.8182, S%=100 y BER=0.0909. Por ende, el mejor algoritmo para la clasificación de atletas en este caso de estudio es el ν-SVM con kernel RBF. Es posible concluir que el sistema propuesto puede ser utilizado para determinar los atletas aptos para futuras competiciones en este deporte, mientras proporciona un robusto y operativo resumen de características para propósitos de gestión de selección de deportistas. Finalmente, este enfoque, que es adecuado para deportistas de Tae Kwon Do, puede ser utilizado también en otras artes marciales.

Agradecimiento

Los autores agradecen la contribución de la Universidadde las Fuerzas Armadas [ESPE] por la ayuda económica para el desarrollo de este proyecto bajo las becas de investigación 2013-PIT-014y 2016-EXT-038.

References / Referencias

Alderson, J. (2015). A markerless motion capture technique for sport performance analysis and injury prevention: Toward a 'big data', machine learning future. *Journal of Science and Medicine in Sport, 19*(3), e79. doi: 10.1016/j.jsams.2015.12.192

Badr, H., Abdelkarim, M., Hanane, E., & Mohammed, E. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications, 2014*. doi: 10.14569/SpecialIssue.2014.040203

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1), 245 - 271. doi: https://doi.org/10.1016/S0004-3702(97)00063-5

Brotons, J. (2005). Propuesta de un modelo integral para el proceso de detección, selección y desarrollo de talentos deportivos a largo plazo. *Revista Digital, 10*(91). Retrieved from: http://www.efdeportes.com/efd91/selec.htm

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*(1-3), 389-422.

Kong, Y., Wei, Z., & Huang, S. (2018). Automatic analysis of complex athlete techniques in broadcast taekwondo video. *Multimedia Tools and Applications, 77*(11), 13643-13660. https://doi.org/10.1007/s11042-017-4979-0

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*(4), 261-283. https://doi.org/10.1007/s10462-011-9272-4.

Kwon, D. Y. (2013). A study on taekwondo training system using hybrid sensing technique. *Retos, 16*(12), 1439-1445. http://dx.doi.org/10.9717/kmms.2013.16.12.1439

Kwon, D. Y. & Gross, M. (2005). Combining body sensors and visual sensors for motion training. In: *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, (pp. 94–101). New York, NY: ACM. http://doi.acm.org/10.1145/1178477.1178490

Langley, P. (1994). Selection of relevant features in machine learning. In: *Proceedings of the AAAI Fall Symposium on Relevance* (pp. 140-144). AAAI.

Lara, R. (2015). *Real-time volcanic monitoring using wireless sensor networks* [doctoral dissertation). Universidad Rey Juan Carlos: Madrid, España.

Liu, C., Wang, W., Zhao, Q., Shen, X., & Konan, M. (2017). A new feature selection method based on a validity index of feature subset. *Pattern Recognition Letters, 92*(C), 1-8. doi: 10.1016/j.patrec.2017.03.018

Muscolo, G. G., & Recchiuto, C. T. (2016, September). T.P.T. a novel taekwondo personal trainer robot. *Robot Auton. Syst., 83*(C), 150-157. http://dx.doi.org/10.1016/j.robot.2016.05.009

Parikh, K. S., & Shah, T. P. (2016). Support vector machine – a large margin classifier to diagnose skin illnesses. *Procedia Technology, 23*, 369-375.

Scholkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT.

Shi, L., Duan, Q., Ma, X., & Weng, M. (2012). The research of support vector machine in agricultural data classification. In: D. Li & Y. Chen (Eds.), *Computer and Computing Technologies in Agriculture* (pp. 265-269). Berlin-Heidelberg, Germany: Springer.

Suto, J., Oniga, S., & Sitar, P. P. (2016, May). Comparison of wrapper and filter feature selection algorithms on human activity recognition. In: *2016 6th International Conference on Computers Communications and Control (ICCCC)*, (pp. 124-129). doi: 10.1109/ICCCC.2016.7496749

Trejo, O. & Miramá, V. (2018). Machine learning algorithms for inter-cell interference coordination. *Sistemas & Telemática, 16*(46), 37-57. doi:10.18046/syt.v16i46.3034

Urcuqui, C. & Navarro, A. (2016). Framework for malware analysis in Android. *Sistemas & Telemática, 14*(37), 45-56. https://doi.org/10.18046/syt.v14i37.2241

Valero, C. (2017). Aplicación de métodos de aprendizaje automático en el análisis y la predicción de resultados deportivos. *Retos, 34*, 377-382.

Vergara, J., Martínez, M. C., & Caicedo, O. (2017). A benchmarking of the efficiency of supervised ML algorithms in the NFV traffic classification. *Sistemas & Telemática, 15*(42), 47-67. doi:10.18046/syt.v15i42.2539

Zelic, I., Kononenko, I., Lavrac, N., & Vuga, V. (1997). Induction of decision trees and bayesian classification applied to diagnosis of sport injuries. *Journal of Medical Systems, 21*(6), 429-444. https://doi.org/10.1023/A:1022880431298

Zhang, Y. (2012). Support vector machine classification algorithm and its application. In C. Liu, L. Wang, & A. Yang (Eds.), *Information Computing and Applications*, (pp. 179-186). Berlin-Heidelberg, Germany: Springer.

Zhong, M., Hung, J., Yang, Y., & Huang, C. (2016). GA-SVM classifying method applied to dynamic evaluation of taekwondo. In: *2016 International Conference on Advanced Materials for Science and Engineering (ICAMSE)*, (pp. 534-537). doi: 10.1109/ICAMSE.2016.7840191

CURRICULUM VITAE

Román Alcides Lara Cueva *Ph.D Engineer in Electronics and Telecommunications from the Escuela Nacional Politécnica (Quito-Ecuador, 2001); Master in Wireless Systems and Related Technologies from the Politecnico di Torino (Italy, 2005); Master and PhD., in Telecommunication Networks for Developing Countries from the Universidad Rey Juan Carlos (Madrid-España, 2010/2015). He joined the Department of Electrical Engineering of the Universidad de las Fuerzas Armadas [ESPE] (Sangolquí-Ecuador) in 2002 and is a full professor since 2005. He has participated in more than ten research projects developed with public funds (five of them as main researcher). His main areas of interests are: digital signal processing, smart cities, wireless systems and automatic learning theory / Ph.D. en Ingeniería Electrónica y Telecomunicaciones de la Escuela Nacional Politécnica (Quito-Ecuador, 2001); Magíster en Sistemas Inalámbricos y Tecnologías Relacionadas del Politécnico di Torino (Italia, 2005); Magíster y Ph.D. en Redes de Telecomunicaciones para Países en Desarrollo de la Universidad Rey Juan Carlos (Madrid-España, 2010/2015). Se unió al Departamento de Ingeniería Eléctrica de la Universidad de las Fuerzas Armadas [ESPE] (Sangolquí, Ecuador) en 2002 y es profesor de tiempo completo de dicha institución desde 2005. Ha participado en más de diez proyectos de investigación desarrollados con fondos públicos (cinco de ellos como investigador principal). Sus áreas de interés son: procesamiento digital de señales, ciudades inteligentes, sistemas inalámbricos y teoría de aprendizaje automático..*

Alexis Darío Estévez Salazar *Candidate to Engineer in Electronics and Telecommunications at the Universidad de las Fuerzas Armadas [ESPE] (Sangolquí-Ecuador). In 2017 he joined to the Sistemas Inteligentes research group as assistant researcher. He completed the Cisco Certified Network Associate Fast Track courses and is candidate to CISCO certification. Actually is black belt –first dan– in Tae Kwon Do and coach of formative schools in this sport. His main areas of interest in research are machine learning and design of low cost technology related to sports / Candidato a Ingeniero en Electrónica y Telecomunicaciones en la Universidad de las Fuerzas Armadas [ESPE] (Sangolquí, Ecuador). En 2017 se unió al grupo de Sistemas Inteligentes como investigador asistente. Completó el curso de Cisco Certified Network Associate y es candidato a dicha certificación. Es cinturón negro en Tae Kwon Do y entrenador de escuelas formativas en este deporte. Sus áreas de interés son el aprendizaje de máquina y el diseño de tecnologías de bajo costo relacionadas ese deporte.*