Selección de características para atributos continuos en tareas de clasificación de actividad física

Features selection for continuous attributes in classification of physical activity tasks

Seleção de recursos para atributos contínuos em tarefas de classificação de atividade física

Enrique V. Carrera
Universidad de las Fuerzas Armadas ESPE, Ecuador
evcarrera@espe.edu.ec

Jefferson Stalin Rodríguez Páramo
Universidad de las Fuerzas Armadas ESPE, Ecuador
jsrodriguez2@espe.edu.ec

Resumen

Los dispositivos móviles contienen diversos sensores con capacidad para enviar datos que se utilizan en la toma de decisiones, un ejemplo es la clasificación de actividad física basada en el uso de acelerómetros y giroscopios. Las señales de los sensores se procesaron previamente aplicando diferentes técnicas que extrajeron un sinnúmero de atributos, los cuales sirvieron para el desarrollo de tareas de clasificación. La optimización de sistemas de clasificación requirió la disminución del número de características de entrada con la finalidad de sintetizar la dimensión de su conjunto y tiempo de aprendizaje. Este artículo empleó métricas de ganancia de información para atributos continuos, que redujeron la incertidumbre y extrajeron únicamente aquellas características más significativas a través de los datos procesados. El análisis de los resultados que se obtuvieron en la clasificación de actividad física usando redes neuronales, mostraron no solamente la disminución de características, sino también un error por debajo del 5 % y la reducción del tiempo de procesamiento en aproximadamente 55 %.

Palabras clave: aprendizaje de máquina, actividad física, selección de características, atributos continuos, ganancia de información.

Abstract

Mobile devices contain different sensors with the ability to send data that are used in decision-making, an example is the classification of physical activity based on the use of accelerometers and gyroscopes. The signals from the sensors were processed previously applying different techniques which extracted a countless number of attributes, which were used for the development of classification tasks. Optimizing systems of classification required the decrease of the number of input features with the purpose of synthesizing the dimension of its set and learning time. This article used metrics of information gain for continuous attributes, that reduced the uncertainty and extracted only those most significant characteristics through the processed data. The analysis of the results that were obtained in the classification of physical activity using neural networks, showed not only the reduction of features, but also an error below the 5% and the reduction of processing time by approximately 55%.

Key words: machine learning, physical activity, features selection, continuous attributes, information gain.

Resumo

Dispositivos móveis contêm vários sensores capazes de enviar os dados utilizados na tomada de decisões, um exemplo é a classificação de actividade física baseada na utilização de acelerómetros e giroscópios. Os sinais dos sensores são processados através da aplicação de diferentes técnicas extraídos inúmeros atributos, que serviram para o desenvolvimento de tarefas de classificação. A optimização do sistema de classificação necessária a redução do número de características de entrada, a fim de sintetizar a dimensão de tempo em conjunto e aprendizagem. Este artigo usou métricas de ganho informações para atributos contínuos, o que reduziu a incerteza e extraídas apenas as características mais significativas através dos dados processados. A análise dos resultados obtidos na classificação de actividade física utilizando redes neurais, não

só mostraram diminuição características, mas também um erro inferior a 5% e o tempo de processamento reduzido em cerca de 55%.

Palavras-chave: aprendizagem de máquina, atividade física, seleção de características, atributos contínuos, ganho de informação.

Fecha recepción: Enero 2016 **Fecha aceptación:** Junio 2016

Introducción

Los dispositivos móviles contienen diversos sensores que en la actualidad son utilizados en varios campos y con un sinnúmero de aplicaciones alrededor del mundo (Das, Green, Perez, y Murphy, 2010). Los dispositivos móviles "inteligentes" debido a su tamaño pequeño, capacidad para enviar y recibir datos y potencia de cálculo, permiten almacenar información que puede ser manipulada (Kwapisz, Weiss, y Moore, 2011).

Toda la información recopilada por estos dispositivos electrónicos brinda un aporte significativo para el desarrollo y monitoreo, aspectos referentes al cuidado de la salud, rehabilitaciones, diagnóstico de enfermedades, seguridad de las personas, entre otros (Mitchell, Monaghan, y O'Connor, 2013).

Las señales que son emitidas por los sensores no pueden ser clasificadas con algoritmos estándares, por lo que en primera instancia se deben transformar los datos de estado puro a información, cuyo procesamiento sea más sencillo en función del tiempo o la frecuencia (Weiss y Hirsh, 1998). De este modo, se consigue procesarlos y extraer un determinado número de características con base en diferentes métricas.

La gran cantidad de datos de entrada existentes provoca que el tiempo de procesamiento aumente (Han, Kamber, y Pei, 2011), lo que ocasiona que la optimización de sistemas de clasificación demande la reducción de estas. Para ello es necesario utilizar un algoritmo que permita la

selección de características, de tal manera que se sintetice la dimensión de su conjunto y el tiempo de aprendizaje (Yang y Wang, 2011).

La clasificación o selección automática de características es una de las tareas más comunes en donde las redes neuronales artificiales han demostrado su eficacia, ya que realizan un procesamiento automático de datos y están basadas en el sistema nervioso biológico (Isasi y Galván, 2004).

Es importante mencionar que las Redes Neuronales Artificiales, desde su aparición y por su desarrollo acelerado, han tenido un uso significativo como un tipo de tecnología para minería de datos, ello debido a que dicha tecnología tiene atributos para una modelación efectiva y eficiente de problemas complejos (Lu, Setiono, y Liu, 1996).

La presente investigación tiene como base el tratamiento de información de tipo continuo mediante métricas para ganancia de información, lo cual por medio de un algoritmo será discretizado. En consecuencia, será posible el proceso de selección de características para atributos continuos en tareas de clasificación de actividad física; es decir, se identificarán las características más relevantes para el proceso de clasificación, al reducir la incertidumbre y obtener únicamente aquellas más significativas.

Las características seleccionadas deben precisar la actividad física de una persona (caminar, subir o bajar gradas, sentarse, pararse, acostarse).

Cabe mencionar que el criterio de maximización de la ganancia de información produce un sesgo hacia los atributos que presentan gran cantidad de valores distintos, lo que resuelve este problema al usar la razón de ganancia como criterio de separación (Hong, 1997). Esta medida tiene en cuenta tanto la ganancia de información como las probabilidades de los distintos valores de los atributos; a su vez, dichas probabilidades son recogidas mediante la información de separación, que no es más que la entropía del conjunto de datos respecto a los valores de los atributos.

Los resultados de la clasificación de actividad física usando redes neuronales, como se describe más adelante, muestran que al utilizar: ganancia de información, puntos de quiebre para los 5 grupos de intervalos de selección y porcentaje de error en cada uno de ellos, se consiguió disminuir el conjunto de características (561) en 86 % (78), por lo que se percibe una optimización en cuanto al tiempo de procesamiento de datos.

La estructura de este artículo se detalla a continuación: en la sección 1 se muestra la descripción de los materiales y métodos adoptados, los cuales exponen el desarrollo experimental para la captura de datos, la descripción matemática del algoritmo propuesto y el proceso de entrenamiento de la red. En la sección 2 se exhiben los resultados experimentales y su análisis; y, finalmente en la sección 3 se exponen las conclusiones.

1. MATERIALES Y MÉTODOS

Conjunto de datos y sensores

En el mercado existe una amplia diversidad de dispositivos móviles para los que se han desarrollado distintos sistemas operativos como: iOS de Apple y Android de Google. En el presente trabajo se utilizó la base de datos "*Human Activity Recognition Using Smartphones Data Set"* (UCI HAR *Dataset*), del Repositorio de Aprendizaje de Máquina de la Universidad de California, misma que trabaja con un Smartphone (Samsung Galaxy S II) colocado en la cintura. A través de su acelerómetro y giroscopio embebidos, se obtiene la aceleración lineal y la velocidad angular en sus tres ejes XYZ. Los experimentos fueron grabados en video para etiquetar los datos de forma manual. El conjunto de datos se divide aleatoriamente en dos grupos. Ahí, 70 % (21 personas) de los voluntarios fue seleccionado para generar los datos de entrenamiento y 30 % (9 personas) proporcionó los datos de prueba.

Las características seleccionadas para esta base de datos provienen de las señales en bruto de los tres ejes del acelerómetro y giroscopio, las cuales en el dominio del tiempo se capturaron a una velocidad constante de 50 Hertz (Hz) y fueron muestreadas con ventanas deslizantes de ancho fijo de 2.56 segundos (s) y 50 % de solapamiento (128 lecturas/ventana). La señal de aceleración tiene dos componentes: gravitacional y movimiento del cuerpo; los cuales se separaron y

depuraron en aceleración del cuerpo y gravedad, recurriendo a un filtro pasa banda y otro de tercer orden pasa bajo Butterworth, ambos con una frecuencia de corte de 20 Hz para eliminar el ruido. La fuerza de gravedad posee únicamente componentes de baja frecuencia, por lo tanto, se utilizó un filtro pasa bajo Butterworth con frecuencia de corte de 0.3 Hz. A partir de cada ventana se obtuvo el vector de características mediante el cálculo de las variables de tiempo y dominio de la frecuencia.

La aceleración del cuerpo y la velocidad angular se derivaron en función del tiempo, para obtener las señales *Jerk*, y la magnitud de estas señales tridimensionales se calculó con el manejo de la norma euclidiana (distancia respecto al origen).

Se empleó una Transformada Rápida de Fourier (FFT) en algunas de las señales que fueron utilizadas para estimar variables del vector de características, las cuales proporcionaron una matriz de datos de 10.299 muestras y 561 características en dominio del tiempo y la frecuencia (Linchman, 2013).

Métricas

La base de datos del Repositorio de Aprendizaje de Máquina de la Universidad de California cuenta con 33 variables obtenidas de las señales en los tres ejes del acelerómetro y giroscopio, las cuales fueron procesadas con 17 métricas. Esto da un total de 561 características, derivadas de la multiplicación entre variables y métricas. A continuación, se observan las métricas y variables en sus tablas correspondientes.

La tabla 1 contiene algunas variables iguales pero en diferentes ejes, razón por la cual se contabiliza tres veces la variable para los tres ejes (X, Y, Z).

Tabla 1. Conjunto de Variables

#	Descripción
1,2,3	Aceleración del cuerpo en los tres ejes (XYZ), en función del tiempo.
4,5,6	Aceleración de la gravedad en los tres ejes (XYZ), en función del tiempo.
7,8,9	Derivada de la aceleración del cuerpo en los tres ejes (XYZ), en función del tiempo.
10,11,12	Velocidad angular del cuerpo en los tres ejes (XYZ), en función del tiempo.
13,14,15	Derivada de la velocidad angular del cuerpo en los tres ejes (XYZ), en función del tiempo.
16	Magnitud de la aceleración del cuerpo, en función del tiempo.
17	Magnitud de la aceleración de la gravedad, en función del tiempo.
18	Magnitud de la derivada de la aceleración del cuerpo, en función del tiempo.
19	Magnitud de la velocidad angular del cuerpo, en función del tiempo.
20	Magnitud de la derivada de la velocidad angular del cuerpo, en función del tiempo.
21,22,23	Aceleración del cuerpo en los tres ejes (XYZ), en dominio de la frecuencia.
24,25,26	Derivada de la aceleración del cuerpo en los tres ejes (XYZ), en dominio de la frecuencia.
27,28,29	Velocidad angular del cuerpo en los tres ejes (XYZ), en dominio de la frecuencia.
30	Magnitud de la aceleración del cuerpo, en dominio de la frecuencia.
31	Magnitud de la derivada de la aceleración del cuerpo, en dominio del tiempo.
32	Magnitud de la velocidad angular del cuerpo, en dominio de la frecuencia.
33	Magnitud de la derivada de la velocidad angular del cuerpo, en dominio de la frecuencia.

Tabla 2. Conjunto de Métricas

#	Métricas
1	Media
2	Desviación Estándar
3	Desviación Media Absoluta
4	Valor Máximo
5	Valor Mínimo
6	SMA
7	Energía
8	IQR
9	Entropía
10	Auto regresión
11	Correlación
12	Máximo Índice
13	Frecuencia Media
14	Skewness
15	Kurtosis
16	Energía de un intervalo de frecuencia
17	Ángulo entre vectores

Ganancia de Información

Ya se mencionó que el criterio de maximización de la ganancia de información está basado en la entropía de la teoría de la información, es decir, es una medida de incertidumbre de una variable aleatoria (Roobaert, Karakoulas, y Chawla, 2006).

Para determinar la ganancia de información del presente estudio se discretizaron los atributos, a partir de lo cual se calculó dicha ganancia para los 5 grupos de intervalos de selección; con ello se generó una lista ordenada y eliminaron aquellos atributos con los menores resultados.

Los 5 grupos utilizados fueron de: 4, 6, 8, 10 y 12 intervalos; la razón de emplear únicamente estos grupos obedece a que sus intervalos muestran una tendencia general, es decir, si se usa un mayor o menor número de intervalos (menos de 4 o más de 12), la media seguirá siendo la misma, lo cual no cambia los datos obtenidos.

En conclusión, se realizó la reducción esperada de la entropía de los datos al conocer el valor de atributos continuos en tareas de clasificación de actividad física.

Redes neuronales

Su principal función es el aprendizaje y su esquema determina el tipo de problemas que será capaz de resolver (Isasi y Galván, 2004). Por otro lado, los investigadores de inteligencia artificial y estadística se han interesado en las propiedades más abstractas de las redes neuronales, tales como su habilidad para desarrollar computación distribuida y tolerar el ruido en la entrada de la red (Cazorla, Colomina Pardo, y Viejo Hernando, 2011). Actualmente se entiende que otras clases de sistemas (incluyendo redes bayesianas) tienen estas propiedades, sin embargo, las redes neuronales son dignas de estudio pues permanecen como una de las formas más populares y efectivas al momento de construir sistemas de aprendizaje (Russell y Norving, 2004).

En el presente trabajo se precisa saber qué tipo de red neuronal ofrece una mayor eficiencia al momento de cumplir con los requerimientos propuestos. Se optó por las redes de alimentación

directa (*feedforward*), las cuales contienen una serie de capas: una con conexión de entrada a la red, una capa posterior que tiene una conexión de la capa anterior y una capa que produce la salida de la red. Estas redes con suficientes neuronas en su capa oculta pueden adaptarse a cualquier problema de asignación de entrada-salida finita. Dos tipos de redes de alimentación directa conocidas y utilizadas en la herramienta matemática MATLAB® son Fitnet y Patternnet.

Fitnet es una red de alimentación directa de dos capas con función de activación sigmoidea: una capa de neuronas oculta y otra de neuronas de salida lineal, que se acoplan a problemas de asignación multidimensionales con datos consistentes. La red será entrenada con un algoritmo de propagación hacia atrás "Levenberg-Marquardt", y en caso de que la memoria no sea suficiente se utilizará un algoritmo de propagación hacia atrás gradiente conjugado escalado (Monar, 2014).

Las redes de reconocimiento de patrones (Patternnet) es una red de alimentación directa de dos capas: una oculta y otra con neuronas de salida Softmax (patrón de red) con función de transferencia tipo sigmoidea. Esta red se entrena con el algoritmo de propagación hacia atrás gradiente conjugado escalado para clasificar las salidas de acuerdo a las entradas, y los datos de destino deben consistir en vectores de todos los valores de cero a excepción de un 1 en el elemento (*i)*, que es la clase a representar (Monar, 2014).

RESULTADOS Y ANÁLISIS

Con el fin de seleccionar la red neuronal que ofrece una mayor eficiencia respecto al tiempo de procesamiento y porcentaje de error, se realizaron tres pruebas de entrenamiento con 1, 70 y 561 características respectivamente, como se muestra en la tabla 3.

Tabla 3. Entrenamiento de redes neuronales.

Red Neuronal	Número de Características	Número de Muestras	Tiempo de Procesamiento (s)	Porcentaje de Error Train (%)	Porcentaje de Error Test (%)
Fitnet	1	7 352	6.296	15.821	15.845
	70	7 352	97.340	1.352	4.515
	561	7 352	325.761	0.4528	1.8709
Patternnet	1	7 352	9.647	29.491	29.583
	70	7 352	44.226	1.990	5.417
	561	7 352	112.262	0.701	2.322

En referencia a los resultados de los entrenamientos de cada red, se observa que la red Patternnet presenta un menor tiempo de procesamiento de datos y un porcentaje de error considerable en comparación a la red Fitnet, lo que permitirá conocer el estado de actividad física de una o varias personas simultáneamente y en tiempo real; dichas razones sustentan el uso de esta red.

Posteriormente se calculó la ganancia de información que tiene cada característica, dividiendo el conjunto de datos de entrenamiento en 5 grupos que contienen diferentes números de intervalos. Estos fueron estructurados de la siguiente manera: el primer grupo dividido en 4 intervalos, el segundo en 6 intervalos, el tercero en 8 intervalos, el cuarto en 10 intervalos y el quinto en 12 intervalos. Esto se realizó debido a que los datos presentan atributos continuos (Cao, Ma, Liu, y Guo, 2012).

Características del acelerómetro y giroscopio

Las características fueron ordenadas desde aquella que aporta mayor cantidad de información hasta llegar a la que menos aporta, como muestra la figura 1.

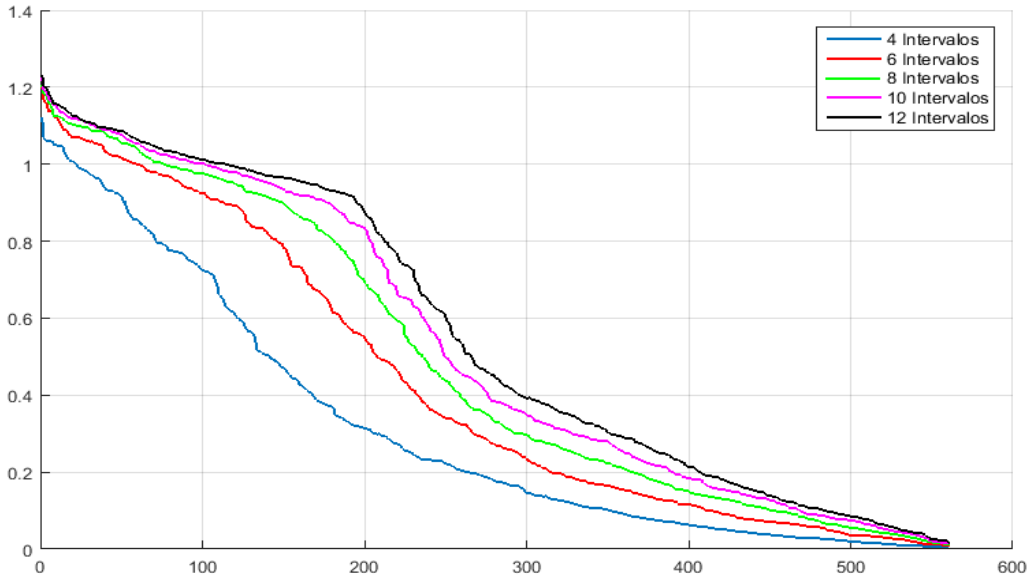

Figura 1.- Ganancia de información para los 5 grupos de intervalos (561 características).

Al calcular la derivada de ganancia de información se pueden apreciar los puntos de quiebre en cada grupo de intervalos, por ejemplo, alrededor de 200 características es clara la caída de la curva, a partir de ello se obtiene la media (promedio), que posteriormente servirá para establecer el número de características y el respectivo error que marcan tanto en el entrenamiento (*train*) como en la prueba (*test*), lo cual se aprecia en la tabla 4.

Tabla 4. Selección de puntos de quiebre.

	INTERVALOS						INTERVALOS					
	4	**6**	**8**	**10**	**12**	**Promedio**	**4**	**6**	**8**	**10**	**12**	**Promedio**
NÚMERO DE CARACTERÍSTICAS	280	276	278	280	278	278	133	131	139	129	133	133
	267	266	263	263	254	263	127	123	125	125	126	125
	261	259	259	254	253	257	103	102	104	101	109	104
	250	248	252	247	248	249	85	94	93	86	89	89
	234	234	232	231	229	232	85	83	85	81	78	82
	230	229	224	227	229	228	72	71	69	72	70	71
	211	214	213	215	213	213	61	63	60	64	61	62
	206	205	208	204	203	205	51	52	47	50	50	50
	198	201	195	192	191	195	34	43	40	38	39	39
	193	188	184	185	181	186	24	25	30	25	22	25
	178	179	170	179	171	175	16	14	13	16	14	15
	171	169	161	164	168	167	7	7	7	7	7	7
	157	150	150	152	151	152	5	4	4	4	3	4
	137	142	142	140	141	140	1	1	1	1	1	1

A través de la aplicación se determinaron todas las características idénticas de los atributos en los 5 grupos de intervalos, identificando de esa manera cuáles son las que contribuyen con más información relevante al proceso de clasificación de actividad física para el caso específico.

Es evidente que el porcentaje de error tanto para el conjunto de entrenamiento como para el de prueba, se incrementa gradualmente mientras se reducen las características. De esta forma se elige el número de datos mínimos, capaces de generar un error menor al 5 % para que la red neuronal aprenda y el clasificador pueda detectar los patrones que determinan la actividad física que se está realizando.

Tabla 5. Reducción de características del acelerómetro y giroscopio con características idénticas.

Número Características	Número Características Idénticas	Train (%)	Test (%)	Número Características	Número Características Idénticas	Train (%)	Test (%)
561	561	0.7743	2,9446	133	133	1.6391	3.6263
278	235	1.4187	3,1113	125	125	1.7344	3.6958
263	224	1.587	3,2117	104	104	1.6841	3.8639
257	221	1.4117	2,9988	89	89	6.6016	7.0473
249	218	1.5557	3,1409	82	82	6.9032	7.3374
232	213	1.4177	3,1252	71	71	7.0418	7.6495
228	211	1.5069	3,0083	62	62	7.204	7.6347
213	205	1.738	3,3246	50	50	7.3649	8.0591
205	201	1.547	3,0004	39	39	8.8933	8.7641
195	194	1.5908	3,2094	25	25	10.197	10.2485
186	185	1.3913	3,1513	15	15	12.2387	13.3877
175	174	1.4069	3,1168	7	7	13.216	14.8047
167	167	1.4222	3,0931	4	4	13.2696	14.865
152	152	1.5395	3,3428	1	1	15.2551	15.9221
140	140	1.6762	3.5242				

Como se evidencia en la tabla 5, conforme disminuye el número de características tiende a cotejarse con el número de características idénticas. De tal manera, si se observa para el número de características 167, se tiene la misma cifra de número de características idénticas, y conforme disminuyen dichas características el error tiende a incrementarse. Por lo tanto, al hacer el cálculo

combinado para el acelerómetro como para el giroscopio se obtiene una disminución conjunta de hasta 104 características, con un error del conjunto de prueba de 3.8639 %, que se convierte en el límite de disminución de características antes de superar el error de 5 %.

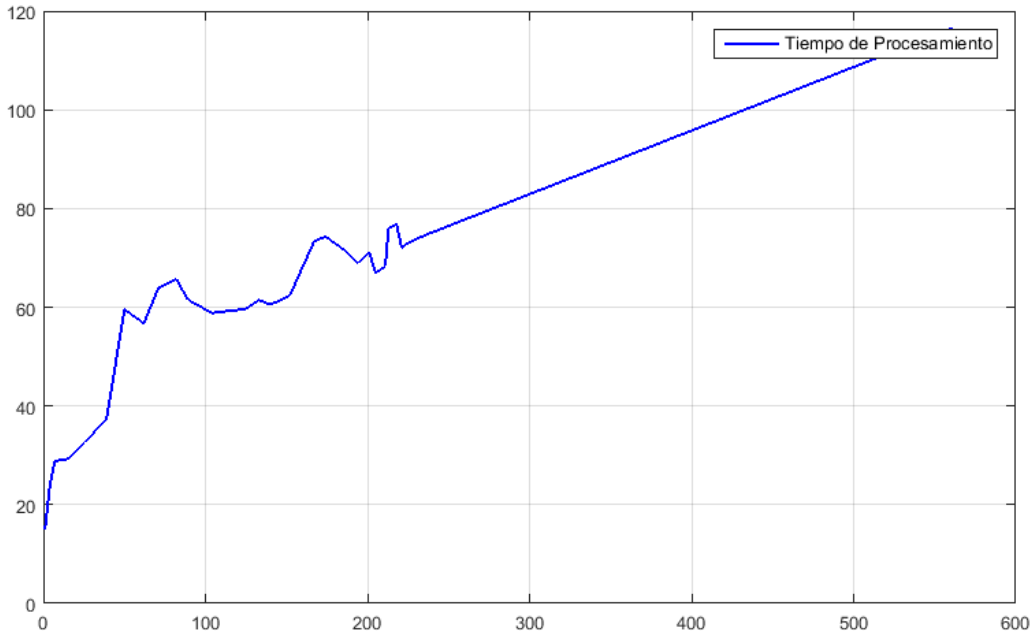

Figura 2.- Tiempo de procesamiento (todas las características).

Al disminuir el número de características como se aprecia en la figura 2, muestra una eficiente reducción del tiempo de procesamiento llegando a obtener un óptimo resultado de 58.8267 (s) para las 104 características.

Características del acelerómetro

La base de datos utilizada nos proporciona las 561 características, de las cuales 345 son del acelerómetro y 216 del giroscopio.

Una vez clasificados estos datos se realiza el mismo análisis de selección y reducción de atributos, tomando las características del acelerómetro y giroscopio por separado como se presenta en las siguientes tablas y figuras.

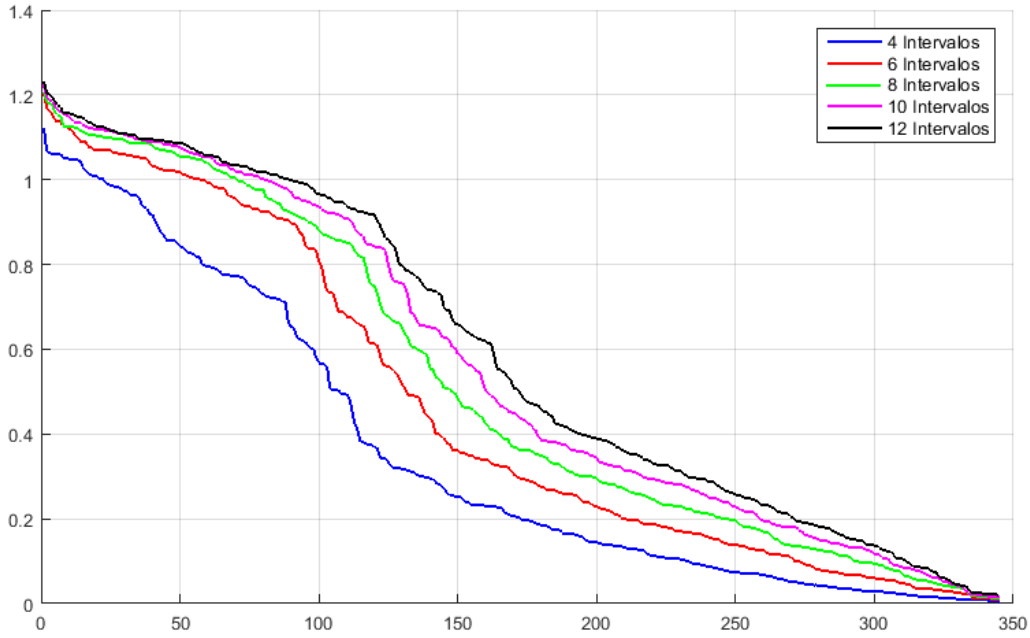

Figura 3.- Ganancia de información para los 5 grupos de intervalos (345 características del acelerómetro).

Por medio de la derivada de ganancia de información se pueden apreciar los puntos de quiebre en cada grupo de intervalos. En el caso del acelerómetro específicamente se denota una caída de la curva en aproximadamente 120 características, como muestra la figura 3. En la tabla 6 se establece la media para fijar el número de características con el cual se realizará el respectivo cálculo del error que marca tanto el conjunto de entrenamiento como el de prueba.

Tabla 6. Selección de puntos de quiebre (acelerómetro).

	INTERVALOS						INTERVALOS					
	4	**6**	**8**	**10**	**12**	**Promedio**	**4**	**6**	**8**	**10**	**12**	**Promedio**
NÚMERO DE CARACTERÍSTICAS	195	194	200	200	206	199	88	92	86	84	85	87
	187	181	182	178	184	182	78	82	80	78	74	78
	166	168	168	167	169	168	73	70	70	67	67	69
	152	161	159	159	163	159	64	61	61	62	64	62
	146	147	149	147	147	147	49	50	47	43	46	47
	144	141	144	143	144	143	38	38	40	33	34	37
	125	129	132	132	128	129	22	25	30	25	24	25
	121	122	121	124	122	122	14	16	13	14	16	15
	103	100	99	100	98	100	7	7	7	7	7	7
	91	94	95	90	96	93	1	1	1	1	1	1

A partir de los puntos de quiebre calculados, se aprecia en la tabla 7 que la disminución adecuada antes de sobrepasar 5 % de error, corresponde a 78 características, demostrando que en forma independiente se obtiene una mayor disminución de características en el caso particular del acelerómetro. Cabe recalcar que aplica el mismo cálculo y análisis para el caso del giroscopio.

Tabla 7. Reducción de características del acelerómetro con características idénticas.

Número Características	Número Características Idénticas	Train (%)	Test (%)	Número Características	Número Características Idénticas	Train (%)	Test (%)
345	345	0.7721	3.0793	87	87	1.3014	4.1031
199	164	1.0944	3.3315	78	78	2.9037	4.1573
182	155	1.1282	3.244	69	69	6.8814	8.3506
168	148	1.1226	3.5088	62	62	7.3665	8.47
159	145	1.1337	3.4383	47	47	8.1803	9.1972
147	140	1.279	3.3502	37	37	7.5788	9.337
143	138	1.1135	3.2756	25	25	9.6996	11.576
129	129	1.3649	3.6542	15	15	11.1481	12.1537
122	122	1.0565	3.8856	7	7	12.7406	13.3385
100	100	1.4052	4.0535	1	1	15.228	15.8485
93	93	1.1843	3.9107				

De igual manera se puede observar en la figura 4 que las 78 características idénticas se procesan en 55.9645 (s). Esto genera una reducción de tiempo significativo con respecto al procesamiento de datos de ambos sensores.

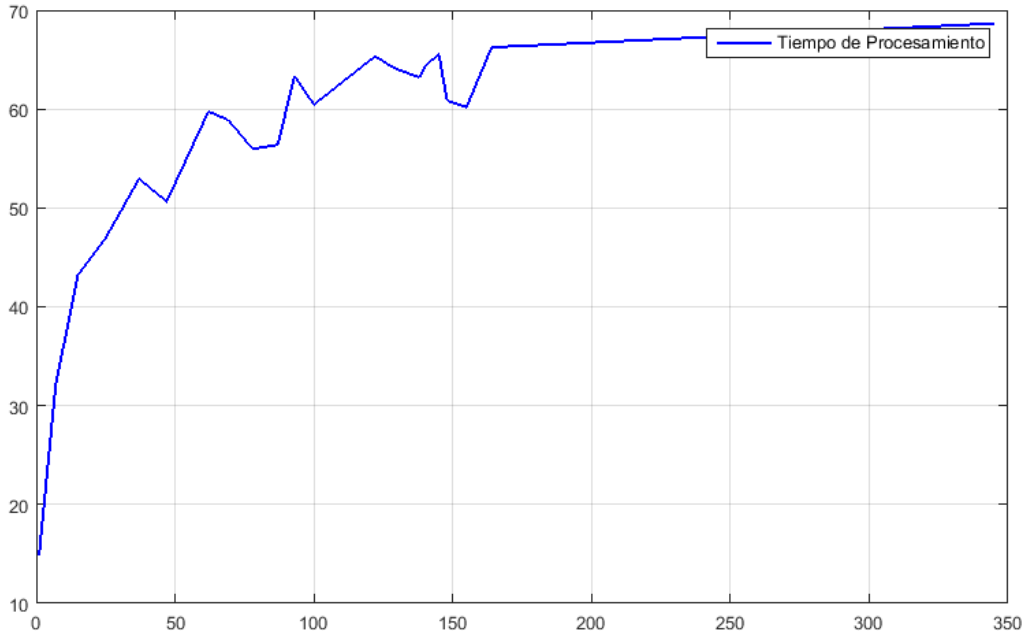

Figura 4.- Tiempo de procesamiento (características del acelerómetro).

Características del giroscopio

La figura 5 muestra una ganancia de información distinta para los 5 grupos de intervalos, por lo que al calcular su derivada los puntos de quiebre serán más evidentes. En este caso se observa una caída de la curva en torno a 80 características. Consecutivamente se realiza el análisis de selección y reducción de atributos únicamente con las características del giroscopio.

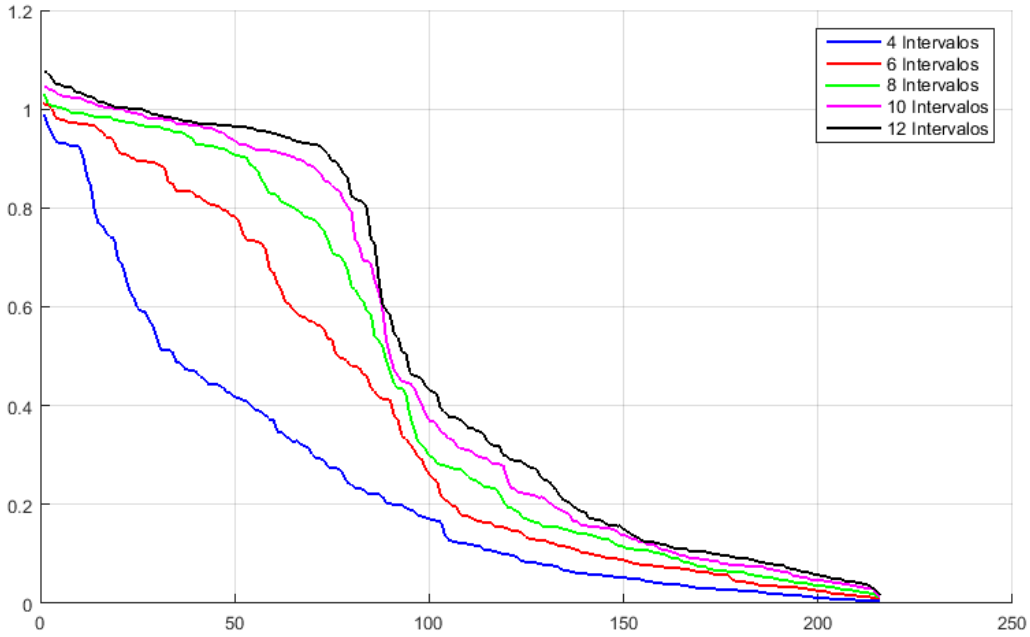

Figura 5.- Ganancia de información para los 5 grupos de intervalos (216 características del giroscopio).

La tabla 8 muestra una evidente disminución en la cantidad de puntos de quiebre comparados con los resultados anteriores, de igual manera se determina la media del número de características y posteriormente se calcula el porcentaje de error para el conjunto de entrenamiento y de prueba.

Tabla 8. Selección de puntos de quiebre (giroscopio).

	INTERVALOS					
	4	**6**	**8**	**10**	**12**	**Promedio**
NÚMERO DE CARACTERÍSTICAS	133	138	147	136	140	139
	122	124	128	122	128	125
	113	116	118	119	118	117
	103	102	101	102	102	102
	96	92	94	96	94	94
	88	84	85	88	86	86
	78	75	79	80	79	78
	60	58	61	53	55	57
	47	51	48	47	46	48
	34	32	39	34	35	35
	19	19	18	21	18	19
	13	16	12	14	14	14
	1	**3**	1	4	**3**	2

La menor cantidad de puntos de quiebre calculados, como se juzga en la tabla 9, muestra que la disminución adecuada antes de sobrepasar 5 % de error, corresponde a 78 características, lo que demuestra que en forma independiente las características del giroscopio también pueden ser utilizadas para el clasificador.

Tabla 9. Reducción de características del giroscopio con características idénticas.

Número Características	Número Características Idénticas	Train (%)	Test (%)
216	216	0.758	3.1706
139	116	1.257	4.1913
125	110	1.2988	4.2884
117	109	1.299	4.2975
102	102	1.4572	4.2413
94	94	2.025	3.7185
86	86	1.9113	4.0126
78	78	2.0773	3.7174
57	57	5.8204	6.3557
48	48	6.3161	6.6827
35	35	7.4344	7.5665
19	19	9.6658	11.5729
14	14	10.1099	12.0147
2	2	16.0832	17.1085

De manera similar, la gráfica del tiempo de procesamiento para las características del giroscopio expresa una disminución igual de eficiente a la del acelerómetro en sus valores, como se puede observar en la figura 6.

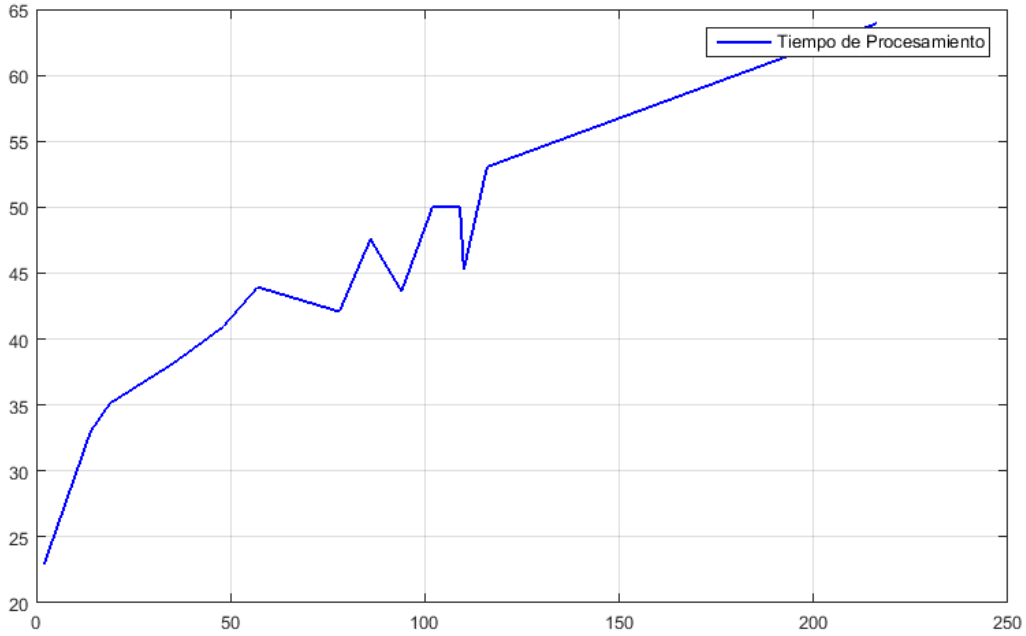

Figura 6.- Tiempo de procesamiento (características del giroscopio).

En el caso del acelerómetro, en cerca de 78 características idénticas se obtiene el rango de error permitido (3.7185 %) antes de exceder un error del 5 % y un tiempo de procesamiento eficiente de 42. 0811 segundos.

El análisis de las características para cada uno de los sensores comprueba que el proceso de selección es más eficiente si se trabaja de forma independiente. Aunque el dispositivo móvil cuente con los dos sensores (acelerómetro y giroscopio), es preferible que el proceso de clasificación de características se realice con uno de los dos, así se logrará optimizar al máximo el tiempo de procesamiento. Este resultado que no se consigue de manera eficiente si ambos sensores trabajan a la par.

Conclusión

Se utiliza 14 % del total de atributos continuos tomados como muestra, y se determina que se puede obtener una reducción del tiempo de procesamiento en aproximadamente 55 % y un error menor al 5 % en el proceso de selección de características sin afectar la clasificación de actividad física.

En el transcurso de la investigación se logró establecer mediante la reducción de características para atributos continuos, que se puede conocer la actividad física y/o estado de una persona (caminar, saltar, correr, etcétera) de forma más eficiente por medio del giroscopio o del acelerómetro en forma independiente.

El presente artículo sirve como fundamento para trabajos futuros con el planteamiento de otro tipo de métodos. Se puede citar como ejemplo a la mejora del algoritmo de ganancia de información a través de la introducción del grado de dependencia de atributos. Así, se conseguirá recoger los datos no solo de una persona, sino de varias simultáneamente y con un procesamiento de la información en tiempo real.

Bibliografía

Cao, D., Ma, N., Liu, Y., & Guo, J. (2012). A Feature Selection Algorithm for Continuous Attributes Based on the Information Entropy. *Journal of Computational Information Systems*, 1467-1475.

Cazorla, M., Colomina Pardo, O., y Viejo Hernando, D. (19 de mayo de 2011). Presentaciones de la asignatura Técnicas de Inteligencia Artificial (Curso 2010-2011). Obtenido de http://hdl.handle.net/10045/17323

Das, S., Green, L., Perez, B., & Murphy, M. (30 de julio de 2010). Detecting User Activities using the Accelerometer on Android Smartphones. *TRUST REU The Team for Research in Ubiquitous Secure Technology*, 29.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*, tercera edición, USA: Elsevier.

Hong, S. J. (1997). Use of contextual information for feature ranking and discretization. *IEEE transactions on knowledge and data engineering*, *9*(5), 718-730.

Isasi, P., y Galván, I. (2004). *Redes De Neuronas Artificiales. Un enfoque práctico*, primera edición, Madrid, España: Pearson.

Kwapisz, J., Weiss, G., & Moore, S. (2011). *Activity recognition using cell phones accelerometers. ACM SIGKDD Explorations Newsletter*, segunda edición, vol. 12, New York, USA.

Linchman, M. (04 de abril de 2013). UCI Machine Learning Repository. Recuperado el 10 de noviembre de 2015, de http://archive.ics.uci.edu/ml

Lu, H., Setiono, R., & Liu, H. (1996). Effective data mining using neural networks. *IEEE transactions on knowledge and data engineering*, *8*(6), 957-961.

Mitchell, E., Monaghan, D., & O'Connor, N. (19 de abril de 2013). Classification of Sporting Activities Using Smartphone Accelerometers. *Sensors*, *13*, 16.

Monar, W. L. (octubre de 2014). Repositorio Digital - Escuela Pilitécnica Nacional. Recuperado el 30 de julio de 2016, de http://bibdigital.epn.edu.ec/handle/15000/8711

Roobaert, D., Karakoulas, G., & Chawla, N. (2006). Information gain, correlation and support vector machines. *Feature Extraction. Springer Berlin Heidelberg, 207*, 463-470.

Russell, S., y Norving, P. (2004). *Inteligencia Artificial un enfoque Moderno,* segunda edición, Madrid, España: Pearson.

Weiss, G., & Hirsh, H. (27 de agosto de 1998). Learning to Predict Rare Events in Event Sequences. *Knowledge Discovery and Data Mining*.

Yang , B., & Wang, L. (2011). *The Construction Method of Knowledge Discovery Theory System Based on Cognitive.* (C. a. Circuits, Ed.) Wuhan: IEEE.