Análisis de datos de actividad física con aprendizaje supervisado

Luis Felipe Villaseñor - A01023976

ITESM

1. Introducción

Como parte de la actividad "Análisis de datos de actividad física con aprendizaje supervisado", utilizamos el acelerómetro que viene instalado en la mayoría de los celulares para obtener datos del sensor mientras se realizaban diversas actividades. Posteriormente utilizamos 5 modelos de clasificación para ver qué tan efectivos son para identificar la actividad física realizada.

2. Resumen de resultados

Los mismos clasificadores se probaron con dos conjuntos de datos, cada uno con 4 movimientos distintos. El primer conjunto de datos describe los siguientes movimientos:

1. Saltar la cuerda
2. Bailar como señora (cumbia)
3. Lagartijas
4. Sentadillas

Mientras que el segundo conjunto describe los siguientes movimientos:

1. Desplantes
2. Correr en círculos
3. Quedarse quieto (parado)

4. Rodar en el piso

Los clasificadores utilizados para intentar identificar los movimientos fueron Support Vector Machines (SVM), k-nearest neighbors (k-nn), un clasificador perceptrón multicapa, random forest y naive bayes.
La siguiente tabla resume las métricas de evaluación obtenidas para cada modelo utilizando un k-fold con 5 particiones:

Dataset 1									
Modelo	Acc.	Pre. 1	Pre. 2	Pre. 3	Pre. 4	Rec. 1	Rec. 2	Rec. 3	Rec. 4
SVM	0.963	0.956	0.945	1	0.951	0.957	0.940	0.992	0.961
K-NN	0.974	0.965	0.964	0.993	0.967	0.974	0.944	1	0.982
MLP	0.959	0.964	0.948	0.987	0.926	0.967	0.909	0.993	0.974
Random Forest	0.972	0.962	0.969	0.969	0.978	0.983	0.936	1	0.973
Naive Bayes	0.730	0.857	0.704	0.906	0.562	0.532	0.551	0.929	0.900
Dataset 2									
SVM	0.909	0.857	0.913	1	0.861	0.902	0.907	1	0.830
K-NN	0.899	0.808	0.904	0.969	0.936	0.908	0.923	1	0.762
MLP	0.913	0.899	0.938	0.993	0.824	0.877	0.893	1	0.888
Random Forest	0.930	0.881	0.911	1	0.936	0.899	0.953	1	0.882
Naive Bayes	0.772	0.567	0.892	1	0.781	0.942	0.830	1	0.297

La tabla resume la exactitud de cada modelo, así como la precisión y la exhaustividad por clase, para cada dataset. Como podemos observar, y a diferencia de lo que esperaba, los clasificadores tuvieron un mejor desempeño identificando los movimientos del dataset 1 que del dataset 2. Cabe destacar que todas las métricas de evaluación mencionadas en el documento se calcularon utilizando un k-fold de 5 particiones.
Para el dataset 1, los mejores clasificadores fueron K-NN y Random forest, mientras que el pero fue indiscutiblemente Naive Bayes. El movimiento que identificaron mejor fueron las lagartijas y los que más les costaron fueron bailar como señora y las sentadillas.
Para el dataset 2, también es claro identificar que el peor clasificador fue el Naive

Bayes. La actividad que identificaron mejor fue quedarse quieto parado (como era de esperarse) y la que más les costo identificar fueron los desplantes y rodar en el piso.

Posteriormente, aprovechando que uno de los clasificadores con mejor rendimiento fue el K-nn, aplicamos un método de obtención de hiperparámetros y de selección de características para ver qué tanto podía mejorar el modelo.

2.1. Obtención de hiperparámetros y selección de caracteristicas del dataset 1

Para la obtención de hiperparámetros utilizamos la función *GridSearchCV* de la librería *scikit learn*, con la cuál variamos el número de vecinos de 1 a 10, variamos la distancia utilizada de distancia de Minkowski de orden 1 a orden 3 y los pesos del modelo de uniforme a distancia. El método devolvió que los hiperparámetros óptimos son un vecino, distancia de Minkowski de orden 1 y el peso del modelo como uniforme. El modelo con dichos hiperparámetros obtuvo 0.98 de exactitud, 0.96, 0.98, 0.99 y 0.98 de precisión por clase y 0.99, 0.95, 1 y 0.99 de exhaustividad por clase

Posteriormente aplicamos el método de selección de caracteristicas al modelo con los hiperparámetros obtenidos. El método utilizado fue el método Sequential Feature Selector que es de tipo *Wrapper*. El modelo obtuvo mejores resultado utilizando entre 21 y 23 características, aunque la tendencia a este resultado se observa desde las 9. Por ejemplo, utilizando 23 de las 30 caracteristicas, el modelo obtuvo 0.99 de exactitud, precisión de 0.99, 0.99, 1 y 0.97 por clase y exhaustividad de 0.97, 0.98, 1 y 1 por clase.

2.2. Obtención de hiperparámetros y selección de caracteristicas del dataset 2

En el caso del conjunto de datos 2, seguimos exactamente el mismo procedimiento que se siguió para el conjunto de datos 1. El método de *Grid Search* encontró los mismos hiperparámetros óptimos, es decir, un vecino, distancia de Minkowski de orden 1 y el peso del modelo como uniforme. Con estos hiperparámetros se alcanzó una exactitud de 0.90, exhaustividad de 0.94, 0.92, 1, 0.72 por clase y precisión de 0.81, 0.88, 1 y 0.95 por clase.
En el caso de la selecció de caracteristicas con este conjunto de datos, el método *Sequential Feature Selector* encontró que el modelo obtiene mejores resultado utilizando entre 17 y 21 características. Utilizando 21 caracteristicas por ejemplo, el modelo obtuvo 0.93 de exactitud, 0.88, 0.90, 1 y 0.93 de precisión por clase y 0.93, 0.95, 1 y 0.83 de exhaustividad por clase.

3. Random Forest

De los 5 clasificadores utilizados, solamente random forest y naive bayes no los revisamos en clase. Considerando que random forest tuvo un mejor rendimiento en la clasificación de estos datos, explicaré este algoritmo.
Para poder entender el funcionamiento del algoritmo *Random Forest* debemos recordar el funcionamiento de los árboles de decisión ya que los árboles de decisión se juntan para formar el Bosque aleatorio [1].
El árbol de decisión es un modelo de clasificación en el cuál se dibuja un árbol al revés, es decir, con el nodo raíz hasta arriba. Cada nodo representa una condición que divide el árbol en sus diferentes ramas resultando en más nodos. Los nodos hasta abajo del árbol que ya no se dividen son la clasificación que se le asigna a los datos que caigan en ese nodo y se denominan como nodos hoja [2].
El árbol de decisión busca la característica y el umbral que resulta en grupos

cada vez más homogeneos utilizando métricas como el índice de Gini. Cabe destacar que se pueden agregar las condiciones necesarias para tener un árbol que clasifique correctamente todos los datos, pero hay que tener cuidado ya que se puede caer en *overfitting* que quiere decir que el modelo está demasiado ajustado a los datos de entrenamiento y pierde la capacidad de generalizar para otros datos [2].

Ya que recordamos el funcionamiento de los árboles de decisión, el *Random Forest* funciona el utilizar un conjunto de árboles de decisión, como el nombre sugiere. Cada árbol de decisión dentro del *Random Forest* predice una clase y aquella clase que tenga la mayoría de "votos.es la clase que predice el *Random Forest*.

La clave para el buen funcionamiento del *Random Forest* es la baja correlación entre los árboles de decisión, de lo contrario se obtendría el mismo resultado que usar solo un árbol de decisión [2]. Para obtener diferentes árboles de decisión, se toman diferentes muestras del conjunto de datos de entrenamiento para entrenar cada árbol [3]. De este modo el modelo se protege de los errores que puede tener un solo árbol de decisión confiando en que la mayoría de los árboles realizarán una clasificación correcta [2].

4. Consideraciones al usar datos de dispositivos móviles y *wearables*

Las mayores preocupaciones al momento de utilizar datos de dispositivos móviles y *wearables* como los utilizados en esta actividad incluyen que estos datos se vendan a terceros para que estos puedan conocer más acerca de tu estilo de vida. Esto con el objetivo de realizar analítica de datos de salud o para ofrecer publicidad más personalizada, por ejemplo, ofreciendo algún suplemento

o algún equipo para hacer ejercicio. De manera similar, otra preocupación es que los datos se vendan a aseguradoras para que sepan con mayor precisión tu estado de salud y poder cobrar mayores cargos si se detecta algún problema de salud o algún estilo de vida poco saludable [4]. Otras preocupaciones con el uso de estos datos son que estos dispositivos pueden ofrecer datos poco precisos en ocasiones [4], como con el contador de pasos o el oxímetro. Esto puede resultar en una mala interpretación y una toma de decisiones basada en datos erróneos. Finalmente, la que considero la mayor de las preocupaciones es la ciberseguridad. Considerando que se están trabajando con datos bastante personales que incluso pueden incluir tu ubicación, no queremos que cualquiera tenga acceso a información de este tipo [5].

Personalmente yo no tengo ningún problema con que se vendan este tipo de datos para ofrecer publicidad más personalizada. De este modo las empresas logran vender más sus productos y uno como consumidor se ve beneficiado por estos productos. No obstante, las demás preocupaciones mencionadas sí me parecen de mayor importancia.

Primeramente, el hecho de que se puedan vender a empresas aseguradoras ya me parece una violación a nuestra privacidad. Me haría sentir como si viviéramos completamente observados y controlados, además de que estos datos pueden ser poco precisos como menciona la siguiente preocupación.

Finalmente, la que me parece la mayor de las preocupaciones es la de la ciberseguridad. Verdaderamente se tratan de datos sensibles que en caso de caer en las manos equivocadas pueden perjudicarnos terriblemente.

5. Conclusión

Estoy completamente seguro de que el monitoreo de la actividad física puede ser benéfico para la salud. Si se usan estos datos de manera inteligente, podrían ayudar a entrenar de manera más eficiente, medir nuestros tiempos y en general mejorar nuestro rendimiento.

Además de las aplicaciones que monitorean la actividad física, ya existen aplicaciones que permiten medir nuestros niveles de estrés, monitorear el sueño y nuestros signos vitales, las cuales las considero como muy buenas ideas e incluso e llegado a utilizarlas. Estas aplicaciones aprovechan los diferentes sensores como el acelerómetro, el GPS, el giroscopio, la cámara (las aplicaciones que escanean los macronutrientes de los productos, por ejemplo), el micrófono (los detectores de COVID con la tos) o incluso un oxímetro de pulso para monitorear nuestra salud de muchas maneras.

Se deben tener en cuenta las diferentes consideraciones éticas que implica el trabajar con este tipo de datos, pero estoy seguro que pueden llegar a ser extremadamente benéficas para sus usuarios, hasta el punto en el que podrían llegar a detectar problemas de salud algún día e incluso podrían salvar vidas.

Referencias

1. T. Yiu. (2019) Understanding random forest. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2
2. P. Gupta. (2017) Decision trees in machine learning. [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
3. H. Deng. (2018) Why random forests outperform decision trees. [Online]. Available: https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5

4. U. of Chicago. (2020) Ethics of wearables: How health providers use health data insights from wellness technology. [Online]. Available: https://healthinformatics.uic.edu/blog/ethics-of-wearables/
5. S. Lee. (2017) The ethics of data collection: Smart phones and wearable technology. [Online]. Available: https://thisisglance.com/the-ethics-of-data-collection-smart-phones-and-wearable-technology/