Análisis de datos de actividad física con aprendizaje supervisado

José de Jesús Gutiérrez Aldrete

3 de junio de 2021

Resumen

Las aplicaciones que monitorean la actividad física han traído innovación al mundo del deporte, en especial ahora que en la pandemia la gente opta por ejercitarse en casa. Algunas de ellas utilizan aprendizaje computacional para clasificar qué actividad se está realizando. En este proyecto se generaron 2 conjuntos de datos con 4 actividades físicas distintas cada uno utilizando información proveniente de sensores de un celular y después se entrenaron modelos clasificación para discriminar las actividades. Se probaron 5 modelos y se usaron sobre el K vecinos más cercanos técnicas de optimización de hiperparámetros y selección de características por el método de envoltura. Finalmente se obtuvieron dos modelos satisfactorios, un KNN con exactitud de 0.989 para el conjunto 1 y para el conjunto 2 un Random Forests con 0.93, ambos con valores de precisión y sensibilidad por clase encima de 0.88.

1. Introducción

La industria *fitness* ha cambiado para siempre. Incluso antes de la pandemia este sector ya estaba digitalizándose, y es que hoy es común ver a personas con relojes inteligentes o aplicaciones deportivas en sus dispositivos. El monitoreo de actividad física ya no solo es una afición, se ha vuelto una cultura [1].

Algunas de estas aplicaciones pueden determinar si acaso estás corriendo, trotando, descansando y demás. Esto es posible con ayuda de algoritmos de clasificación y los datos generados por el dispositivo móvil. Sin embargo, estos datos son de índole personal, y su manejo debe efectuarse con especial consideración.

Si bien las aplicaciones de salud derivadas del monitoreo de dispositivos móviles son diversas, y que también hay muchas opciones de variables a considerar, el objetivo de este proyecto es utilizar señales de 8 actividades divididas en dos conjuntos de cuatro para discriminarlas utilizando técnicas de aprendizaje computacional en función de datos derivados del acelerómetro, giroscopio y campo magnético de un celular.

El resto del documento se divide de la siguiente manera: en la sección 2 se definen las preguntas de investigación, en la 3 la metodología usada y en la 4 los resultados. Después se describe el algoritmo de clasificación de bosques aleatorios en la sección 5, y se diserta sobre la privacidad y manejo de datos en la sección 6. Finalmente en la sección 7 se abordan las conclusiones.

2. Pregunta de investigación

- ¿Es posible utilizar las señales de movimiento para discriminar entre actividades de un usuario que porta un teléfono móvil?
- ¿Qué consideraciones se deben tomar en cuenta a la hora de recolectar datos de un dispositivo móvil en cuanto a privacidad y manejo de los mismos?

3. Metodología

3.1. Extracción de datos

Con el uso de la aplicación móvil Sensor Stream IMU+GPS y un pograma de Python se registraron las señales de algunos sensores de un teléfono celular mientras se realizaron 8 diferentes actividades físicas.

Estas actividades físicas y sus respectivas señales están divididas en dos conjuntos de datos. El primero contiene "saltar la cuerda", "bailar como señora", "lagartijas" y "sentadillas". El segundo contiene "desplantes", "correr en círculos", "quedarse parado y quieto" y "rodar en el piso".

3.2. Modelado

Por cada conjunto de datos, se utilizaron los modelos de clasificación K vecinos más cercanos (KNN), máquinas de vectores de soporte (SVM), red neuronal perceptrón multicapa (MLP), bosques aleatorio (random forests) y bayes ingenuo (naive bayes), todos de sklearn [2].

Cada modelo fue evaluado con validación cruzada de 5 pliegues y usando las métricas de exactitud, precisión por clase y sensibilidad por clase.

3.3. Hiperparámetros y selección de características

En ambos análisis se tomó al knn para optimizar sus hiperparámetros utilizando búsqueda de cuadrícula (GridSearchCV de la librería sklearn).

Posteriormente, utilizando KNN los hiperparámetros dados por la búsqueda de cuadrícula, se utilizó un método de envoltura (SequentialFeatureSelector de sklearn) para seleccionar características. Lo anterior con el propósito de mejorar el desempeño del modelo.

4. Resultados

En el primer archivo de datos el modelo con el mejor desempeño fue el KNN con 3 vecinos, obteniendo una exactitud de 0.97367 y valores de sensibilidad y precisión por clase por encima de 0.94. No obstante, después de optimizar sus hiperparámetros se llegó a que la mejor exactitud se obtiene utilizando un solo vecino y la distancia de Manhattan, alcanzando el valor de 0.98177. Además, tras la selección de características con el método de envoltura, se tiene que con solo 23 atributos se logra elevar ligeramente ese puntaje a 0.98989, pero terminando con la precisión y sensibilidad de cada clase por encima de 0.97. Incluso las lagartijas se reconocen con un 100 % de precisión y un 100 % de sensibilidad. En resumen, se puede considerar a este modelo como uno que logra discriminar bastante bien las clases.

Por el otro lado, respecto al segundo archivo, el modelo con mejor desempeño fue el random forests con una exactitud de 0.93, y valores de precisión y sensibilidad para cada clase mayores a 0.88. Aún así, se optó por optimizar los hiperparámetros del KNN, estos siendo una vez más un solo vecino y la distancia de Manhattan, consiguiendo una exactitud de 0.9014. Usando el mismo modelo con dichos hiperparámetros en la selección de características, se obtuvo que con las mejores 21 se llega a una exactitud de 0.9284. Junto con ello, la precisión y sensibilidad de "quedarse parado y quieto" son del 100 %, aunque la sensibilidad de "rodar en el piso" es de 0.83. Lo anterior da indicios de que aún sin buscar optimizar los hiperparámetros, el random forests es levemente más certero. Puede que con estos datos el desempeño del modelo no se haya igualado al anterior, pero igual se puede decir que es un buen modelo para discriminar las actividades físicas dadas.

5. Clasificador de bosques aleatorios

El random forests es un algoritmo de aprendizaje de máquina que pertenece a la familia de los métodos de ensamble. Los métodos de ensamble combinan las predicciones de modelos base para poder mejorar la capacidad de generalización de un nuevo modelo [2].

Dentro de la familia de los métodos de ensamble, están los métodos de empaquetado o bagging. Lo que sucede en el bagging es que cada modelo se entrena con distintas porciones de los datos, de forma que al combinar sus resultados los errores individuales son compensados y la predicción es más robusta [3]. El random forests toma el bagging combinándolo con el promedio de un grupo de árboles no correlacionados [4].

Como el sesgo de un árbol individual es el mismo que el de un grupo de árboles, lo que le queda al random forests por mejorar es la reducción de la varianza del bagging, aminorando la correlación entre los árboles. Esto se logra en el proceso de crecimiento del árbol con la selección aleatoria de las variables de entrada [4].

Este algoritmo se construyen de la siguiente manera [5]:

1. Si N es el número de observaciones en el conjunto de entrenamiento, una muestra Z* de esos N casos se toma aleatoriamente con reemplazo. Este proceso de muestreo se llama bootstrap. Con dicho bootstrap se entrena el arbol T_b.
2. Si existen p varibles de entrada, se elige $m < p$ tal que para cada nodo, m variables se seleccionan aleatoriamente de p. Sobre m se aplica el criterio para ramificar el árbol. Este valor se mantiene constante durante la generación de todo el bosque, y usualmente es $\lfloor\sqrt{p}\rfloor$ [4].
3. Si bien el desarrollo de Breiman [6] extiende el árbol lo más posible y sin poda, en sklearn esto puede ser regulado con tamaño mínimo del nodo hoja n_min o la profundidad máxima.
4. Se juntan las predicciones de los B árboles y las nuevas instancias se predicen en base a la clase que obtuvo la mayoría de votos.
5. Nuevas instancias se predicen a partir de la mayoría de votos dada por las predicciones de los B árboles.

Lo que se resume en este pseudocódigo [4]:

Algoritmo 1 Clasificador random forests

for $b = 1$ to B **do**

1. Obtener una muestra de bootstrap $\mathbf{Z}^*$ de tamaño N del conjunto de entrenamiento.
2. Modelar un árbol T_b con los datos del bootstrap repitiendo recursivamente los siguientes pasos para cada nodo terminal, hasta que el tamaño mínimo del nodo n_{min} es alcanzado.
   - I Seleccionar m variables aleatoriamente de las p que tienen los datos.
   - II Buscar el mejor punto de división para las variables m.
   - III Partir el nodo en dos nodos hijos.

La salida es el ensamble de los árboles $\{T_b\}_1^B$

Para hacer una predicción en un nuevo punto x:

Sea $\hat{C}_b(x)$ la clase predicha por el b-ésimo árbol. Entonces $\hat{C}_{rf}^B(x)$ = la mayoría de votos $\{\hat{C}_b(x)\}_1^B$

Durante el bootstraping aproximadamente un tercio de las observaciones no se toman. A ellas se les llama "Fuera de la bolsa"(OOB, por sus siglas en inglés) y sirven para evaluar el error de la

clasificación y estimar la importancia de cada variable. Una vez que el error OOB se estabiliza, el entrenamiento del algoritmo puede terminarse [6].

6. Privacidad y manejo de datos

Cuando se habla de datos provenientes de dispositivos móviles, se tiene que reparar en consideraciones especiales. Por ejemplo, a diferencia de una computadora, un celular usualmente viaja en los bolsillos de los usuarios, lo que habilita el rastreo de su localización. Un individuo puede razonablemente ser identificado a través de sus patrones de ubicación, y para muchos podría ser riesgoso. Hay visitas a lugares que uno quiere mantener en privado, y si se enterara la familia, el trabajo o el gobierno, las consecuencias podrían ser graves. Ejemplos de lugares sensibles son las clínicas de aborto, espacios de reunión LGBT+ o mezquitas [7].

Opino que esta época ha hecho más fácil la adquisición de datos, pero al mismo tiempo ha vuelto más complejo mantener su seguridad, ya que los sistemas son vulnerables a ciberataques. Encriptar los datos podría ser una buena medida de mitigación de amenazas, pero la protección de la información y el correcto manejo de la misma requiere de la sinergia de herramientas, protocolos y buenas prácticas de seguridad [8].

Así, la recolección de datos de dispositivos móviles debe cumplir con varios objetivos de seguridad, cuyo núcleo está representado por la "triada CIA" [8], compuesta por:

- **C**onfidencialidad: garantizar que los datos están protegidos contra divulgación no autorizada.
- **I**ntegridad: que los datos no sean modificados.
- **A**vailability (Disponibilidad): que siempre que los usuarios autorizados lo necesiten, puedan acceder a los datos.

Además, para proteger los derechos civiles y la privacidad de datos, se está desarrollado un marco de trabajo sobre el manejo correcto de información. Los principios clave se muestran en la tabla 1, donde la persona de la que se están extrayendo los datos es llamada *cliente*, y quien los recolecta es llamado *compañía*.

Propiedad de los datos	Las compañías solo custodian los datos, pero los verdaderos dueños son los clientes.
Conciencia	Los clientes deben estar informados de qué datos se están extrayendo y para qué se están usando.
Consentimiento	Los clientes deben ser capaces de decidir si compartir sus datos con terceras partes.
Acceso	Las compañías deben limitar el acceso de los datos de acuerdo a reglas y los deseos del cliente.
Seguridad e integridad	Las compañías deben asegurarse de que los datos sean válidos y a salvo del mal uso.

Cuadro 1: Principios de la recolección de datos [9]

Yo considero que los datos son la nueva moneda de cambio en el mundo, y por lo mismo su protección y buen uso son imprescindibles. Las compañías deberían apuntar a que parte de su propuesta de valor esté la privacidad de la información de sus clientes porque es algo que cada vez nos está preocupando más.

Debemos exigirle a los recolectores de datos que los protejan y los manejen con cuidado, pero también pienso que es absurdo pedir que dejen de obtener nuestros datos porque "están

monetizando con ello". Creo las compañías ha dejado el paradigma de "venta agresiva"muy atrás, y ahora su objetivo es poder ofrecer valor al cliente. Esto se da a través de la personalización de productos y servicios, posible por los datos. A cambio de nuestros datos estamos teniendo experiencias más prácticas, cómodas y positivas, y es razonable que ello tenga un precio.

7. Conclusiones

En conclusión, para los dos conjuntos de datos con 4 actividades físicas diferentes, se obtuvieron modelos de clasificación que discriminan satisfactoriamente las clases. Para ambos conjuntos resultó el K vecinos más cercanos con un vecino y distancia de Manhattan como el más prometedor. No obstante, se podría ahondar en la optimización de hiperparámetros y selección de características con otros modelos, como el random forests, para determinar si es posible brindar alguna mejoría en la generalización del clasificador.

Considero que el monitoreo de la actividad física propia es muy benéfico para la salud, principalmente porque tiene una función motivadora. Si se tiene un objetivo de salud, como bajar de peso o correr cierta cantidad de kilómetros, ser consciente de tu progreso te inspira a superarte, y eso se logra a través del monitoreo.

Otro tipo de aplicaciones basadas en el monitoreo con dispositivos móviles que podrían ser benéficas para la salud serían aquellas que van almacenando un historial médico. Si bien no es recomendable su uso para diagnosticarse, son útiles para recibir alertas sobre indicadores anormales. Si no se puede asistir al médico o si es muy difícil, otras aplicaciones envían directamente los datos del dispositivo al personal de salud para que pueda dar seguimiento de tu estado. Por ejemplo, a mí me aterra la idea de que el día de mi muerte nadie se entere que morí, así que tener un dispositivo que mida mi frecuencia cardiaca y mande alerta a mi médico me ayudaría a vivir más tranquilo.

Además del campo magnético, la aceleración o el giro, dispositivos más especializados podrían medir presión arterial, saturación de oxígeno o como ya mencioné, frecuencia cardiaca. También existen electrocardiogramas y electroencefalogramas portables [10, 11].

Como se puede ver, la inteligencia artificial y la salud están convergiendo en soluciones que pueden ayudar a mejorar la calidad de vida de las personas. Los datos podrían ser la respuesta de un futuro con mejor atención médica, pero no hay que olvidar que su recolección y manejo conllevan responsabilidades éticas.

Referencias

[1] M. Naldaiz, "¿Son las aplicaciones de fitness el futuro de la industria del fitness?: Virtuagym." https://business.virtuagym.com/es/blog/son-las-aplicaciones-de-fitness-el-futuro-de-la-industria-del-fitness/, Sep 2020.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] J. Martínez Heras, "Random forest (bosque aleatorio): combinando árboles." https://www.iartificial.net/random-forest-bosque-aleatorio/, Sep 2020.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Stanford, California: Springer, 2009. https://web.stanford.edu/ hastie/ElemStatLearn/.

[5] J. Orellana Alvear, "Arboles de decision y random forest." https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html, Nov 2018.

[6] L. Breiman and A. Cutler, "Random forests Leo Breiman and Adele Cutler." https://www.stat.berkeley.edu/ breiman/RandomForests/cc_home.htmworkings.

[7] S. A. Thompson and C. Warzel, "Twelve million phones, one dataset, zero privacy." https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html, Dec 2019.

[8] "Guidelines on the use of electronic data collection technologies in population and housing censuses." https://unstats.un.org/unsd/demographic/standmeth/handbooks/guideline-edct-census-v1.pdf, 2018.

[9] E. Naef, P. Muelbert, S. Raza, R. Frederick, J. Kendall and N. Gupta, "Using mobile data for development," 2014. https://www.betterevaluation.org/sites/default/files/Using-Mobile-Data-for-Development.pdf.

[10] Bitbrain, "Diadem." https://www.bitbrain.com/es/productos-neurotecnologia/dry-eeg/diadem, Mayo 2020.

[11] M. F. for Medical Education and Research, "Electrocardiograma (ecg)." https://www.mayoclinic.org/es-es/tests-procedures/ekg/about/pac-20384983.